Simple o3: Enhancing Visual Grounding in Multimodal Reasoning via Adaptive Tool Composition

Anonymous ACL submission

Abstract

The emerging "thinking with images" paradigm enhances model performance by incorporating visual information into the Chain-of-Thought (CoT). However, its efficacy is hampered by persistent visual grounding errors. We systematically identify and categorize three core grounding failure modes: Long-context Grounding Error, Regional Grounding Error, and Fine-Grained Grounding Error. To address these challenges, we propose three targeted compensation strategies: visual reuse, focus, and magnification. To validate their effectiveness, we construct the TWI-Tools-146K dataset and develop Simple-o3, a tool-augmented framework that interleaves image and text tokens. This design sustains visual attention, enforces regional focus, and enhances fine-grained perception by adaptively composing our strategies within a slow-thinking reasoning process. Our approach achieves performance competitive with more complex Reinforcement Learning (RL)-based models, establishing a new strong baseline for multimodal reasoning. Notably, this work goes beyond mere performance metrics by offering an in-depth analysis of how visual grounding failures impede reasoning, how each mechanism rectifies its targeted issue, and why such interventions are effective—thereby contributing to a more principled and reliable reasoning process.

1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated remarkable proficiency across a wide range of vision-language tasks (Bai et al., 2025a; Chen et al., 2024b; Liu et al., 2024a; Li et al., 2024a; Team et al., 2025), from image captioning to visual question answering (VQA). The emerging "thinking with images" paradigm, exemplified by models like o3, further emulates human-like reasoning by integrating visual information into textual Chain-of-Thought (CoT), thereby enabling the co-evolution of visual perception and cognitive processing. Prevailing tool-driven visual exploration often follows a pipeline of cropping regions of interest and then magnifying them (Su et al., 2025b; Zheng et al., 2025; Zhang et al., 2025c), substantially enhancing grounding capabilities for visual reasoning.

However, we contend that significant visual grounding challenges persist in "thinking with images" paradigm. Specifically, the monolithic cropping-then-resizing inference is prone to cumulative errors, where initial grounding inaccuracies directly propagate to reasoning failures. Further, the lengthy sequences generated through slow-thinking processes can introduce more grounding issues, necessitating dedicated visual grounding enhancement mechanisms. Despite these critical challenges, prior work also lacks a systematic analysis and targeted improvement of grounding within complex reasoning chains, leaving a key bottleneck unaddressed.

Systematically, we identify three core error categories of visual grounding, including Long-Context Grounding Error, Regional Grounding Error, and Fine-Grained Grounding Error. Regarding Long-Context Grounding Error, models suffer from progressive visual information decay during extended reasoning, which shifts decision-making from visual evidence toward textual priors, ultimately creating semantic inconsistencies with the input image. Regional Grounding Error occurs when models exhibit spatial misalignment in localization, primarily due to inadequate pixel-level alignment between visual concepts and their linguistic descriptions. When images lack critical details due to small size or blur, models experience perceptual degradation at the foundational level, leading to Fine-Grained Grounding Error.

To address these grounding challenges, we propose targeted compensation strategies: *reuse*,

$focus_area$, and $zoom_in$. $reuse$ strategically re-injecting the original input image into the reasoning chain. $focus_area$ crops the image using bounding boxes (bbox). $zoom_in$ applies interpolation-based upscaling to the entire image. By integrating these tools adaptively within a slow-thinking reasoning process, we preserve visual fidelity throughout the extended multimodal chain.

To validate these tools and enhance visual grounding, we build the TWI-Tools-146K dataset. Through Supervised Fine-Tuning (SFT), we develop Simple o3, a tool-augmented interleaved vision-language reasoning framework as shown in Figure 1. Experiments highlight distinct mechanistic roles: the $reuse$ operation maintains persistent visual tokens, ensuring continuous attention to original content across long reasoning chains; $focus_area$ restricts attention to bounding-box regions, eliminating cross-entity interference and achieving pixel-level precision; $zoom_in$ increases pixel density, enriching structural details and discriminative feature extraction. Moreover, composing these tools yields synergistic effects, substantially boosting overall visual grounding performance.

Our main contributions are fourfold: (1) We expose how deficient grounding results in flawed inference processes, leading to the formal definition and taxonomy of three distinct categories of visual grounding errors that emerge during extended multimodal reasoning processes. (2) We propose $reuse$, $focus_area$, and $zoom_in$ to directly compensate for visual grounding errors, providing the first in-depth mechanistic analysis of how and why each tool specifically enhances visual grounding capabilities. (3) We introduce a scalable data generation pipeline to construct the TWI-Tools-146K dataset. By SFT, we develop Simple o3, competitive with more complex reinforcement learning (RL)-based models. (4) Experimental results demonstrate that Simple o3 achieves significant improvements on various multimodal reasoning and perception benchmarks, validating the effectiveness of our approach.

2 Related Work

2.1 Multi Large Language Models

By integrating visual encoders with powerful LLMs, MLLMs gain visual perception capabilities (Alayrac et al., 2022; Awadalla et al., 2023), employing either MLP-based or attention-based connector modules to bridge the visual and language modalities (Liu et al., 2024b; Li et al., 2023a; Zhu et al., 2023; Bai et al., 2023; Chen et al., 2023). Since most visual encoders are constrained by fixed input resolutions and capturing only partial visual features, subsequent research has prioritized enhancing visual representations. Two predominant approaches have emerged: some directly extend the visual encoders' capabilities (Yuan et al., 2024; Ge et al., 2024; Fan et al., 2024), and others adopt dynamic resolution strategies that process high-resolution images by decomposing them into multiple sub-images while simultaneously analyzing downsampled global views through low-resolution encoders (Ye et al., 2023a; Li et al., 2024c,a; Chen et al., 2024b). While these innovations have substantially improved visual perception, the persistent limitations in reasoning capacity continue to constrain overall model performance.

2.2 Long CoT in MLLMs

Early MLLMs predominantly adopted a two-stage training paradigm consisting of vision-language pretraining followed by instruction fine-tuning (Dai et al., 2023; Liu et al., 2023; Ye et al., 2023b; Zeng et al., 2024). While this approach improved instruction-following capabilities, it inherently separates perception from reasoning. Subsequent research has sought to address this limitation by integrating various forms of textual auxiliary inputs, including detailed reasoning traces, spatial coordinates, attribute relations, and comprehensive image descriptions, to provide evidential grounding for model outputs (Shao et al., 2024; Xu et al., 2024). Recently, the success of RL reasoning has rapidly extended to multimodal domains. While recent research has increasingly focused on reward signal design (Huang et al., 2025; Liu et al., 2025; Zhang et al., 2025b; Zhou et al., 2025), OpenAI's o3 model pioneered the "thinking with images" paradigm and has gradually become pivotal for enhancing multimodal reasoning capabilities (Zhang et al., 2025a; Wang et al., 2025a; Bai et al., 2025b; Ni et al., 2025; Chung et al., 2025). However, prevailing tool-driven approaches like DeepEyes (Zheng et al., 2025) and Chain-of-Focus (Zhang et al., 2025c) often rely on a rigid pipeline of image cropping and then resizing (Zhu et al., 2025; Su et al., 2025a). This lack of adaptability fails to meet the demands of diverse scenarios, where different tasks may require more flexible visual processing strategies.

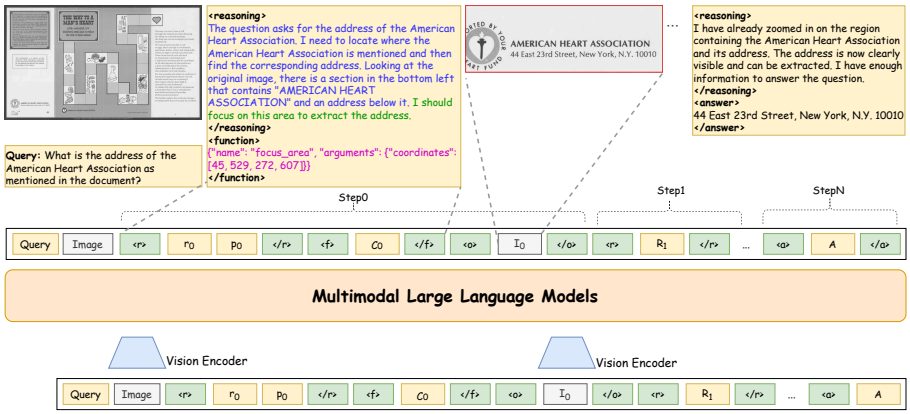


Figure 1: Overview of Simple o3. At Step 0, the blue text represents the atomic reasoning step r_0, while the green text represents the visual operation plan p_0. These two components constitute the reasoning content R_0. The pink text represents the tool command C_0 in JSON format, including $reuse$, $focus_area$ and $zoom_in$.

3 Method

In this section, we propose visual grounding enhancement tools in Section 3.1. Then, we describe the generation process of the TWI-Tools-146K dataset in Section 3.2 and Section 3.3, as well as the training and multi-step inference procedures of Simple o3 in Section 3.4.

3.1 Tool Integration and Invocation

To mitigate visual grounding errors, we introduce three dedicated tools: $reuse$ directly concatenates the original image into the generation sequence; $focus_area$ crops regions of interest based on a bbox; $zoom_in$ upscales the original image via interpolation. As illustrated in Figure 1, the model dynamically produces JSON-formatted tool commands C_t during reasoning, specifying function names and arguments. The executor parses these functions, performs operations on the original image, and concatenates the returned image into the input sequence. Specifically, during long-context text generation, if the model naturally loses visual information, it triggers the $reuse$ operation, prompting the executor to reintroduce the original visual tokens. When focused attention on specific entities or regions is required, the model calls $focus_area$ with a bbox parameter, and the executor returns images of key areas. In cases where image resolution is too low, the model employs $zoom_in$ with a magnification factor, and the executor provides a magnified image to alleviate recognition difficulties caused by discretization errors. By adaptively integrating these visual compensation mechanisms throughout the reasoning process, the system achieves significantly enhanced visual grounding performance.

3.2 Scalable Data Generation Pipeline

We propose the data generation pipeline consisting of a reasoning path generator and a two-step verification module, as shown in Figure 2. In the generation phase, the MLLM serves as an orchestrator that dynamically analyzes task requirements and coordinates predefined visual manipulations: $focus_area$, $zoom_in$, and $reuse$. We employ *gemini-2.5-Flash-Preview-05-20* in its non-thinking mode to optimize the trade-off between generation efficiency and computational overhead. The reasoning path generation follows an iterative cycle:

$$S = \{s_0, s_1, ..., s_t, A\} \tag{1}$$

$$s_t = (R_t, C_t, I_t) \sim P(\cdot \mid Q, I, H_{t-1}; \theta_{\text{MLLM}})(t \geq 0) \tag{2}$$

$$H_{t-1} = \begin{cases} \emptyset & \text{if } t = 0, \\ \{(R_i, C_i, I_i)\}_{i=0}^{t-1} & \text{if } t \geq 1. \end{cases} \tag{3}$$

Here, S represents the iterative reasoning chain comprising the serialized atomic reasoning step s, and final answer A. s_t denotes the output at reasoning step t, including the reasoning content R_t, the tool command C_t, and the returned image I_t, formed as the triple, by sampling from the MLLM while considering the historical reasoning path H_{t-1} that aggregates all previous steps' information as the reasoning process unfolds.

Our pipeline utilizes structured tags as special tokens to delineate distinct reasoning components. The reasoning content $R_t = (r_t, p_t)$, comprising atomic reasoning steps r_t and visual

planning p_t, is encapsulated within <*reasoning*> and </*reasoning*> tags. Tool command $C_t = (T_n, Param)$, specifying both operation names and input parameters, are marked by <*function*> and </*function*> tags. The system dynamically generates new visual states through image manipulations, with the resulting transformed images being encoded and embedded between <*observation*> and </*observation*> delimiters. This structured tagging scheme preserves a complete and parseable record of multimodal reasoning chain progression.

The verification process begins by assessing the geometric validity of tool interactions using *gemini-2.5-flash-lite-preview-06-17*. For $focus_area$, we retain samples where the bbox either precisely matches or fully contains the target entity. $zoom_in$ and $reuse$ operations are validated based on semantic consistency between the tool command C_t and the visual planning description p_t. Finally, the answer is verified by comparing system outputs against ground truth using *Qwen3-turbo*. For each question-image-answer triplet, the MLLM constructs a multi-step reasoning chain, potentially invoking visual tools. The system performs a maximum of two generation attempts per sample to enhance validity. All detailed prompt templates are listed in Appendix A.4.

3.3 TWI-Tools-146K Data Curation

We carefully curate the dataset from diverse tasks to maximize the data distribution, thereby enhancing the model's generalization capability. Our TWI-Tools-146K training dataset is sourced from following datasets: MATHV360K (Shi et al., 2024), LLaVA-CoT-100K (Xu et al., 2024), and ArxivQA (Li et al., 2024b). We exclude subtasks including math problems from MATHV360K, as geometry typically focuses more on the extension of textual tokens. Through proportional sampling across remaining sub-datasets, we initially generate 150K candidate entries. After rigorous verification, we retain approximately 100K high-quality samples, covering perceptual VQA, knowledge-based reasoning, chart analysis, and logical reasoning, representing a comprehensive collection of challenging real-world tasks. To establish robust visual understanding capabilities, we primarily select COCO QA pairs from LLaVA-CoT-100K, generating 27K samples featuring diverse real-world imagery. To further enhance abstract figure interpretation and reasoning, we extract 19K proportionally balanced samples from ArxivQA, covering all subcategories.

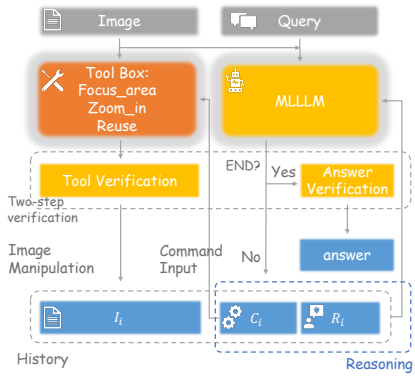


Figure 2: Overview of scalable data generation pipeline. This iteratively generates data through a cycle of MLLM-driven reasoning, function-based image manipulation, and semantic verification, persisting until a final answer is produced and validated.

3.4 Training and Inference for Simple o3

We train on multimodal reasoning sequences $S = (z_0, \ldots, z_t)$ with a modality-aware mask M_t that governs a masked cross-entropy loss. This objective optimizes textual generation by zeroing out gradients for visual tokens, thus ensuring proper gradient flow while maintaining cross-modal interactions. The specific formula derivation is described in Appendix A.1. We perform full parameter fine-tuning on Qwen2.5-VL-7B (Bai et al., 2025a) using TWI-Tools-146K data, with a mixture of partial general reasoning data. Notably, during training, $focus_area$ is followed by the full image with a red bbox as input, while during inference, the image is cropped using the corresponding bbox to inject focused visual tokens. After SFT, the model demonstrates proficiency in processing structured formats, enabling tool usage and multi-step interleaved vision-language reasoning. This capability facilitates dynamic alternation between text and image token generation until reaching an answer or hitting the round limit. The inference workflow is depicted in Appendix A.2. Some reasoning examples of Simple o3 are provided in Appendix A.5.

4 Analysis

To address the lack of long-context, regional, and fine-grained grounding capabilities, this section elaborates on three corresponding compensation mechanisms: $reuse$, $focus_area$, and $zoom_in$. We first analyze how the deficit in each ground-

ing capability impairs the model's reasoning, and then provide an in-depth explanation of how each mechanism specifically enhances its corresponding grounding ability and uncover the insights behind the observed performance gains. The effectiveness of all mechanisms is systematically validated through experiments. The experimental setup is described in Appendix A.3.

4.1 Simple Image Reuse Can Help Long-Context Grounding

Generally, visual features are injected only once at the sequence onset. During long-text generation, this initial visual input becomes progressively diluted by newly generated text tokens, causing the model to rely on textual context for fabrication rather than remaining visually grounded, a phenomenon termed visual forgetting. This shifts the model's decisions from visual evidence towards its internal language model priors, becoming a primary source of hallucinations in long-context reasoning.

Visual reuse computationally mirrors the human "speak-and-look" process—periodically glancing at the reference—to provide visual re-anchoring opportunities. This enables the model to recalibrate its understanding against the evolving context, thereby curbing error propagation in autoregressive generation.

To validate the $reuse$ operation, we modified Simple o3's system prompt to disable $focus_area$ and $zoom_in$, to strategically reinject the image during long reasoning chains. Table 1 demonstrates that the $reuse$ operation delivers significant performance improvements, achieving gains of 31.2 points on MME (Fu et al., 2024) and 4.8% on VStarBench (Wu and Xie, 2024), with at least 1-point improvements on other benchmarks. This indicates that $reuse$ enhances the model's perception and reasoning capabilities through the introduction of additional visual tokens. By establishing dynamic visual checkpoints, our approach systematically counteracts long-context information decay while ensuring persistent output alignment with the visual source.

4.2 Visual Focus Can Help Regional Grounding

Unlike the temporal "forgetting" effect in long-context scenarios, Regional Grounding Error is a form of spatial misalignment, characterized by models' limited precision in localization. We identify two predominant failure modes: In multi-entity scenarios, the model's inability to accurately localize specific instances causes attention dispersion across multiple entities, leading to confusion in semantic and spatial relationships (e.g., misidentifying subjects in "A is holding B" relations). This problem becomes particularly critical when handling entity-specific queries. In single-entity scenarios, the same deficiency causes misidentification of core attributes (e.g., shape, material, text content), as the model fails to concentrate attention on the entity's discriminative regions. While visual reuse primarily addresses temporal "attention drift", visual focus tackles spatial ambiguity. Even in globally clear images, key entities drown in visual clutter. Enhanced regional grounding mitigates information overload and cognitive bias through physical spatial focusing.

$focus_area$ compels the model to shift from text-prior shortcuts toward evidence-based visual reasoning, thereby curbing associative hallucinations. Furthermore, it atomizes multi-entity descriptions into sequential single-entity processing steps, ensuring precise attribute decoupling and preventing misassignment. Critically, the bbox serves as a verifiable intermediate representation that provides a geometric foundation for spatial reasoning while enhancing determinism and verifiability.

During training, we draw bbox on images based on the coordinates returned by $focus_area$. This stems from the fact that some coordinates only provide approximate focal regions, analogous to human visual attention mechanisms, which may either incompletely cover target instances or excessively include irrelevant areas. Directly cropping images based on samples with incomplete target coverage would lead to critical visual information loss, thereby compromising the reasoning chain integrity.

To better leverage regional grounding during inference, we evaluate three operational modes for $focus_area$ in Simple o3: after the model generates a bbox within a tool command, we either (1) return the full image with the bbox overlaid, (2) provide the image cropped to the specified region, or (3) simply reuse the original image. As shown in Table 2, experimental results demonstrate that the image cropping delivers optimal performance. Specifically, Simple o3 can accurately cover target entities or regions when executing $focus_area$ operations, without producing incomplete crops that trigger "I can't see" or "I can't find" responses.

Model	MME(R)	CharXiv	HR-4K	V*	HalB	InfoVQA
Qwen2.5-VL-7B	652.5	37.6	68.8	77.5	42.1	80.0
+ w/ *reuse*	683.7	39.0	70.5	82.3	43.3	81.6

Table 1: Simple o3 achieves significant performance gains by simply repeatedly injecting original visual information, through *reuse* mechanism.

Method	MME(R)	CharXiv	HR-4K	V*	HalB	InfoVQA
Crop	702.1	**41.8**	**76.2**	90.4	**44.4**	**82.0**
Draw	698.2	41.2	75.4	**90.6**	43.8	**82.0**
Reuse	**704.3**	40.7	74.2	89.3	43.8	81.8

Table 2: Performance comparison of *focus_area* operation variants in Simple o3, given bbox: image cropping, bbox drawing, and original image reusing.

This proves that repeatedly injecting key regional visual information effectively enhances model reasoning capabilities.

4.3 Raw Image Magnification Can Help Fine-Grained Grounding

Fine-grained visual grounding is fundamentally constrained by input signal quality. When input images lose critical details due to inherently small size or optical blur, the visual encoder encounters insufficient high-frequency signals and fails to reconstruct discriminative features from sparse or degraded pixels. This effectively renders the model "visually myopic"—unable to resolve essential fine-grained details, becoming constrained by ambiguous low-level perceptual input.

While conventional image enlargement shifts the model's processing from high-level cognitive focusing to low-level perceptual enhancement, we theoretically distinguish two magnification approaches: super-resolution reconstructs high-frequency details to achieve a qualitative "from-none-to-some" leap, whereas our interpolation-based method performs signal resampling to densify pixel distribution. *zoom_in* provides such high-resolution visual signals, and presents existing low-frequency information more continuously, effectively alleviating recognition difficulties caused by discretization artifacts.

To evaluate the impact of input quality, we systematically controlled the base model's maximum input resolution. As shown in Table 3, lower resolutions lead to over-compression and significant detail loss, severely impairing low-level perception, and performance improves monotonically with increasing resolution as high-frequency information is better preserved. These results confirm that high-quality visual input serves as a fundamental prerequisite for effective fine-grained grounding.

Res	MME(R)	CharXiv	HR-4K	V*	HalB	InfoVQA
Low*	651.1	33.0	45.8	42.9	42.3	54.3
Med*	652.5	37.6	61.2	66.5	42.1	80.0
High*	652.5	37.7	68.8	77.5	42.4	81.7
Low	699.2	32.4	56.9	61.3	41.5	62.3
Med	**702.1**	**41.8**	69.3	78.0	44.4	82.0
High	**702.1**	40.1	**76.2**	**90.4**	**46.7**	**82.6**

Table 3: Performance comparison of maximum input resolution variants of Qwen2.5-VL and Simple o3: low resolution ($256 \times 28 \times 28$), medium resolution ($2048 \times 28 \times 28$), and high resolution ($16384 \times 28 \times 28$). * denotes Qwen2.5-VL.

We then examine the effectiveness boundary of *zoom_in* in Simple o3, which is constrained by the maximum input resolution. At low limits, the zoomed-in region remains compressed, making *zoom_in* akin to *reuse*. However, with sufficiently high input resolution limits, the complete enlarged images are fed into the model without being compressed, thereby injecting high-quality, dense visual evidence into the sequence and fully unleashing the fine-grained grounding potential of *zoom_in*.

5 Experiments

5.1 Main Experiments

By delving into the performances in Table 4, we conclude that Simple o3 achieves significant improvements on multimodal reasoning benchmarks. It outperforms the base model by nearly 50 points on the MME reasoning subset and surpasses GPT-4o (Hurst et al., 2024) by 27 points, underscoring the efficacy of "thinking with images" where increasing the number of attended visual tokens enhances reasoning. Further improvements on CharXiv (Wang et al., 2024) and VisuLogic (Xu et al., 2025) confirm our method's ability to handle complex visual reasoning through robust regional grounding and multi-step atomic inference, progressively leading to accurate answers.

As shown in Table 5, Simple o3 consistently outperforms Qwen2.5-VL-7B across general VQA benchmarks. On ScienceQA (Saikh et al., 2022) and RealWorldQA, the gains reflect its stronger capacity for integrating multidisciplinary knowledge and generalizing to real-world scenarios. Furthermore, significant improvements on MMStar (Chen et al., 2024a) and MMVet (Yu et al., 2024) validate its advanced multimodal integration and complex

Model	Reasoning			Perception			Hallucination	
	MME(R)	CharXiv	VisLog	HR-4K	V*	COCOC	POPE	HalB
GPT-4o*	674.6	29.9	25.1	46.8	45.0	23.1	84.6	44.2
GPT4o-mini*	564.3	32.7	25.8	48.0	50.8	17.4	83.3	39.3
Ovis1.6-7B*	547.5	36.5	-	64.1	71.2	14.4	87.8	44.1
LLaVA-OV-9B*	415.4	22.4	22.5	64.7	72.8	9.9	88.3	30.3
Qwen2.5-VL-7B*	652.5	37.6	26.0	68.8	77.5	15.0	86.0	42.1
Simple o3*	702.1	41.8	29.3	76.2	90.4	18.1	87.6	44.4
Δ (vs Base)	+49.6	+4.2	+3.3	+7.4	+12.9	+3.1	+1.6	+2.3

Table 4: Comparison results on multimodal reasoning benchmarks: reasoning tasks in MME, CharXiv(val), and VisuLogic, perception benchmarks: HR-Bench 4K, VStarBench, and COCO Caption(val), and hallucination benchmarks: POPE and HallusionBench. We use Rouge-L scores for COCO Caption. For HallusionBench, we report the mean accuracy of unique questions and all figures. * denotes the result is reproduced by ourselves. For HR-Bench 4K, VStarBench, and POPE, the highest input resolution is set to $16384 \times 28 \times 28$, and for other benchmarks, it is set to $2048 \times 28 \times 28$.

Model	General VQA				Chart	
	Sci-QA	RW-QA	MM*	MMVet	DocVQA	InfoVQA
GPT-4o*	88.4	67.2	61.5	68.6	55.5	38.9
GPT4o-mini*	85.6	66.7	54.5	69.8	78.0	57.9
Ovis1.6-9B*	93.3	71.0	62.6	61.9	88.9	73.4
LLaVA-OV-7B*	95.3	69.7	61.9	52.7	87.0	66.4
Qwen2.5-VL-7B*	88.7	68.4	63.9	68.2	94.7	80.0
Simple o3*	90.0	69.5	66.5	69.7	94.8	82.0
Δ (vs Base)	+1.3	+1.1	+2.6	+1.5	+0.1	+2.0

Table 5: Comparison results on general VQA benchmarks: ScienceQA, RealWorldQA, MMStar, and MMVet, and chart comprehension benchmarks: DocVQA(val) and InfoVQA(val). The highest input resolution is set to $2048 \times 28 \times 28$.

Model	VisuLogic Spatial	MMVet rec_spat	LogicVista spatial
Qwen2.5-VL-7B*	22.9	66.7	20.5
Simple o3*	26.4	83.3	30.8
Δ (vs Base)	+3.5	+16.6	+10.3

Table 6: The performance of Simple o3 on spatially relevant benchmarks.

task handling through refined capability alignment. While DocVQA (Mathew et al., 2021) performance nears saturation, Simple o3 maintains a lead via fine-tuning, and its progress on InfoVQA (Mathew et al., 2022) demonstrates enhanced chart information extraction abilities.

For multimodal perception, Simple o3 demonstrates consistent improvements across granularity levels. On coarse-grained COCO Caption (Chen et al., 2015), it enhances holistic image understanding through increased image reuse. More notably, it achieves significant gains of 7.4% and 12.9% on HR-Bench 4K (Wang et al., 2025b) and VStarBench, highlighting its fine-grained perception capability in high-resolution scenarios where target objects occupy minimal pixels. As shown in Table 6, Simple o3 also excels in spatial reasoning tasks, particularly on subtasks of MMVet and LogicVista (Xiao et al., 2024), achieving improvements of 16.6% and 10.3%, indicating that our interleaved vision-language reasoning approach strengthens relational understanding of entities and regions. By accurately locating objects and aligning visual features with positional cues, the model achieves unified cross-modal comprehension.

Model	MME(R)	CharXiv	HR-4K	V*	HalB	InfoVQA
Qwen2.5-VL-7B	652.5	37.6	68.8	77.5	42.1	80.0
w/ Raw CoT	660.2	38.0	68.6	77.6	42.5	80.6
Simple o3	**702.1**	**41.8**	**76.2**	**90.4**	**44.4**	**82.0**

Table 7: Performance comparison of dataset variants: raw CoT dataset and TWI-Tools-146K (Simple o3).

Model	HR-Bench 4K	POPE	V*Bench
Qwen2.5-VL-7B*	68.8	86.0	77.5
DeepEyes	75.1	87.7	90.1
Chain-of-Focus	-	88.4	88.0
Simple o3*	76.2	87.6	90.4
Δ (vs Best)	+1.1	-0.8	+0.3

Table 8: Comparative performance of Simple o3 with RL-based interleaved vision-language reasoning methods: DeepEyes and Chain-of-Focus.

To validate the effectiveness of tool integration, we compare our proposed TWI-Tools-146K dataset with the original CoT data using identical queries, in Table 7. Our experiments reveal that direct SFT with the original 146K samples yields limited improvements over the base model, even showing slight performance degradation on HR-Bench 4K. This suggests that text-based CoT outputs alone have constrained utility for enhancing model capabilities. In contrast, our TWI-Tools-146K dataset demonstrates significant gains across all evaluation datasets. This improvement substantiates that our three proposed tools collectively enhance grounding capabilities, where iterative interactions be-

tween textual reasoning and critical visual information progressively strengthen the model's reasoning capacity.

We also compare Simple o3 with two RL-based approaches, DeepEyes and Chain-of-Focus, which both employ the cropping-then-resizing strategy in Table 8. Our method demonstrates superior performance on both HR-Bench 4K and VStarBench, further validating the effectiveness of the "thinking with images" in visual reasoning. Although Simple o3 does not achieve the best performance compared to other methods in POPE (Li et al., 2023b), it still improves on two hallucination benchmarks, especially by 2.3% in HallusionBench (Guan et al., 2024).

5.2 Ablation Study

We progressively activate visual grounding tools via system prompt customization in Simple o3, to quantify their cumulative impact, as detailed in Table 9. While all components contribute positively, the $zoom_in$ operation yields only marginal gains. In contrast, $focus_area$ proves critical, especially in fine-grained perception tasks such as HR-Bench 4K and VStarBench, where it brings improvements of 4.9% and 6.9%, respectively. Its strength lies in precise visual localization that filters irrelevant visual information while emphasizing essential regions, leading to consistent performance gains. These results indicate that the synergistic integration of these tools collectively elevates the model's reasoning upper bound.

Further, our analysis extends to the role of heterogeneous training data in shaping model capabilities, as summarized in Table 10. First, we establish a baseline by training the model on the general CoT reasoning dataset, as mentioned in Section 3.4, which yields a slight but consistent performance improvement. To strategically enhance specific capabilities, we then progressively integrate three specialized subsets from the TWI-Tools-146K dataset: namely, MATHV360K-Tools, LLaVA-CoT-100K-Tools, and ArxivQA-Tools. Our integration begins with MATHV360K-Tools, which broadens the data distribution by introducing complex visual QA, charts, and logical reasoning problems, leading to a performance leap. We attribute this to its large scale of complex VQA and chart data, which strengthens the model's representation of logical inter-object relations, thereby refining its strategic planning for visual tool operations. However, the initial training mix showed a relative dearth of perceptual samples, resulting in modest gains on fine-grained perception benchmarks. To remediate this, we introduced perceptual data from LLaVA-CoT-100K-Tools, an intervention that elevated the model's fine-grained perception capability without trading off reasoning performance. Furthermore, the logic-centric ArxivQA-Tools, despite its narrow distribution, directly bolsters atomic reasoning capabilities and substantially fortifies the model's logical reasoning faculties with its challenging abstract diagrams from Arxiv papers. Ultimately, the synergistic integration of these multi-source, heterogeneous datasets amplifies the model's higher-order iterative reasoning capabilities.

Model	MME(R)	CharXiv	HR-4K	V*	HalB	InfoVQA
Qwen2.5-VL-7B	652.5	37.6	68.8	77.5	42.1	80.0
+ w/ $reuse$	683.7	39.0	70.5	82.3	43.3	81.6
++ w/ $zoom_in$	689.6	39.6	71.3	83.5	43.7	81.8
+++ w/ $focus_area$	**702.1**	**41.8**	**76.2**	**90.4**	**44.4**	**82.0**

Table 9: Performance comparison of tool implementations variants in Simple o3: Only $reuse$, plus $zoom_in$, and with $focus_area$.

Model	MME(R)	CharXiv	HR-4K	V*	HalB	InfoVQA
Qwen2.5-VL-7B	652.5	37.6	68.8	77.5	42.1	80.0
w/ general	656.8	38.1	69.2	77.6	42.3	80.3
+ w/ MATHV360K*	683.4	39.4	71.3	81.4	42.8	81.2
++ w/ LLaVA-CoT*	688.6	39.7	74.9	88.2	44.1	81.4
+++ w/ ArxivQA*	**702.1**	**41.8**	**76.2**	**90.4**	**44.4**	**82.0**

Table 10: Impact of dataset on model performance: Only MATHV360K-Tools, plus LLaVA-CoT-100K-Tools, and with ArxivQA-Tools. * denotes subsets of TWI-Tools-146K.

6 Conclusion

In this work, we systematically identify and address three fundamental grounding errors that impede reliable multimodal reasoning: Long-Context, Regional, and Fine-Grained Grounding Errors. With the introduction of meticulously curated TWI-Tools-146K dataset and Simple o3 model, we demonstrate that structured, tool-augmented interleaving of image and text tokens can effectively sustain visual attention, enforce regional precision, and enhance fine-grained perception, through $reuse$, $focus_area$, and $zoom_in$ mechanisms, achieving competitive performance against more complex RL-based interleaved models. Crucially, our in-depth analysis reveals how each mechanism specifically augments its target capability and why such interventions are effective, moving beyond mere performance gains to shape a more reliable, transparent, and grounded reasoning process.

7 Limitations

Although this work employs the $reuse$, $focus_area$, and $zoom_in$ mechanisms to enhance the long-context grounding, regional grounding, and fine-grained grounding capabilities of MLLMs respectively, and achieves comparable performance to current RL-based baselines, it still relies solely on SFT, leaving room for improvement in the model's performance ceiling. Furthermore, the dataset we constructed intentionally excludes data such as mathematics, which emphasizes logical reasoning over visual information. While this aims to strengthen grounding capabilities across the three dimensions and achieve significant gains in perceptual reasoning benchmarks, the model may underperform on mathematical problems.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tianyi Bai, Zengjie Hu, Fupeng Sun, Jiantao Qiu, Yizhen Jiang, Guangxin He, Bohan Zeng, Conghui He, Binhang Yuan, and Wentao Zhang. 2025b. Multi-step visual reasoning with visual tokens scaling and verification. *arXiv preprint arXiv:2506.07235*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. 2025. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. *arXiv preprint arXiv:2505.18842*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, and 1 others. 2024. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. 2024. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. 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.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

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

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024c. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. 2025. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *arXiv preprint arXiv:2505.19702*.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *CoRR*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025a. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, and 1 others. 2025b. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.

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

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025a. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025b. 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.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13084–13094.

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

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, and 1 others. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong, and Ruihua Song. 2024. What matters in training a gpt4-style language model with multimodal inputs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7930–7957.

Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, and 1 others. 2025a. Cmm-cot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. *arXiv preprint arXiv:2503.05255*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025b. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and 1 others. 2025c. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and 1 others. 2025. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*.

A Appendix

A.1 Training Incorporating Image Masking

Our data generation pipeline produces interleaved vision-language sequences $\mathbf{Z} = (Q, I, S)$, where the iterative reasoning chain S can be further decomposed into individual elements z_t from different modalities, originating from the text space $\mathcal{T}$ or the visual space $\mathcal{V}$. To adapt the multi-turn "user-assistant-tool" dialogues obtained during data generation into a suitable training format, we consolidate their reasoning paths into single-turn "user-assistant" sequences. Prior to training, we also resolve potential coordinate system discrepancies by converting all spatial coordinates to the native representation format of the target model. For effective learning from multimodal reasoning sequences $S = (z_0, \ldots, z_t)$, we employ a selective loss computation mechanism governed by modality-aware masking M_t, which ensures proper gradient flow while maintaining cross-modal interactions.

$$M_t = \begin{cases} 1 & \text{if } z_t \in \mathcal{T} \quad \text{(compute loss)} \\ 0 & \text{if } z_t \in \mathcal{V} \quad \text{(ignore loss)} \end{cases} \qquad (t \geq 0) \tag{4}$$

The training objective employs a masked cross-entropy loss that optimizes textual generation, zeroing out gradients for visual tokens:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=0}^{T} M_t \cdot \log P(z_t \mid Q, I, H_{t-1}; \theta_{\text{MLLM}}), \tag{5}$$

where the history H_{t-1} is defined as:

$$H_{t-1} = \begin{cases} \emptyset & \text{if } t = 0, \\ \{z_i\}_{i=0}^{t-1} & \text{if } t \geq 1. \end{cases} \tag{6}$$

In our implementation, visual tokens are demarcated by special tokens like $< |\textit{image_pad}| >$ and serve only as contextual inputs. The binary mask ensures the model attends to visual data for understanding while restricting gradient updates to textual outputs.

A.2 Multi-step Inference with Tool Interaction for Simple o3

At each reasoning step t, the model outputs the reasoning content between *<reasoning>* and *</reasoning>* tags. Upon detecting the *</function>* token, the system automatically pairs it with the preceding *<function>* tag and extracts the JSON-formatted tool command C_t. This command is then passed to a predefined function for visual processing, and the resulting image I_t is embedded as visual tokens between *<observation>* and *</observation>* tags before being appended to the model's history. Finally, the model performs summary reasoning and outputs the answer between *<answer>* and *</answer>* tags, completing the inference cycle.

Algorithm 1 MLLM Response with Multi-Turn Inference

Require: Input query Q, Input image I, MLLM π_θ, tool box $\mathcal{T}$, maximum rounds M.
Ensure: Final response y.

Initialize multimodal sequence $y \leftarrow \emptyset$
Initialize round count $r \leftarrow 0$
while $r < M$ **do**
    Generate token $y_t \sim \pi_\theta(\cdot \mid Q, I, y)$ ▷ Append y_t to multimodal sequence y
    $y \leftarrow y + y_t$
    if $\langle$function$\rangle$ $\langle$/function$\rangle$ detected in y_t **then** ▷ Extract tool command C_t
        $C_t \leftarrow \text{Parse}(y_t, \langle\text{function}\rangle, \langle/\text{function}\rangle)$ ▷ Retrieve tool instruction
        $I_t = \mathcal{T}(I, C_t)$ ▷ Image manipulation using the tool box
        $y \leftarrow y + \langle\text{observation}\rangle\, I_t \langle/\text{observation}\rangle$
        Increment round $r \leftarrow r + 1$
    end if
    if $\langle$answer$\rangle$ $\langle$/answer$\rangle$ detected in y **then** ▷ Terminate generation
        return final generated response y
    end if
end while
return final generated response y

A.3 Setups

A.3.1 Implementation Details

We perform full parameter fine-tuning on Qwen2.5-VL-7B (Bai et al., 2025a) using TWI-Tools-146K data, with a mixture of partial general reasoning data. The training was completed in approximately 54 hours on 8 NVIDIA V100 GPUs. The model processes sequences with a maximum length of 8,192 tokens. Input images are resized to a resolution range between a minimum of $4 \times 28 \times 28$ pixels and a maximum of $1024 \times 28 \times 28$ pixels. We employ a cosine learning rate scheduler with an

initial rate of 1.0e-5 and 10% warmup ratio, training for 1 epoch with the batch size of 8. The vision encoder and multimodal projector remain frozen during training.

A.3.2 Baselines and Benchmarks

We compare Simple o3 with proprietary models like GPT-4o (Hurst et al., 2024) and 4o-mini, as well as competitive open-source models, LLaVA-OneVision-7B (Li et al., 2024a) and Ovis1.6-Gemma2-9B (Lu et al., 2024). We additionally evaluate it against other models adopting the "thinking with images" paradigm, including DeepEyes (Zheng et al., 2025) and Chain-of-Focus (Zhang et al., 2025c). To validate the effectiveness of Simple o3, we evaluate it on multimodal reasoning benchmarks, including the reasoning subtasks in MME (Fu et al., 2024) and reasoning questions in CharXiv (Wang et al., 2024), VisuLogic (Xu et al., 2025), as well as fine-grained perception benchmarks such as HR-Bench 4K (Wang et al., 2025b), VStarBench (Wu and Xie, 2024), and COCO Caption (Chen et al., 2015). Additionally, we assess its performance on general VQA tasks like RealWorldQA, ScienceQA (Saikh et al., 2022), MMStar (Chen et al., 2024a), and MMVet (Yu et al., 2024). Beyond these comprehensive evaluations, we also test the model on hallucination benchmarks, including POPE (Li et al., 2023b) and HallusionBench (Guan et al., 2024), and chart comprehension tasks such as DocVQA (Mathew et al., 2021) and InfoVQA (Mathew et al., 2022).

A.4 Detailed Prompt Templates for Data Generation Pipeline

Prompt Template for Data Generation

You need to perform image-text interleaved multi-step reasoning according to the "observation-reasoning-operation" paradigm.
1. Observe the current image, then perform atomic logical reasoning for the current round and determine the next visual operation (whether to focus on the key region, zoom in further, or reuse the original image) within <reasoning> and </reasoning> tags, and return the visual operation parameters.
2. Then, you receive the processed image after the visual operation.
3. Repeat Step 1-2 until sufficient information is obtained.
4. Make a concluding inference within <reasoning> and </reasoning> tags.
5. Provide the concise and accurate final answer or value within <answer> and </answer> tags.

- Given the ground truth, you must hide the prior knowledge of the rationale when outputting the complete reasoning path that leads to the ground truth.
- DO NOT MENTION ANY PARAMETERS within the <reasoning> and </reasoning> tags.
- Ensure the returned function name is semantically aligned with the determined visual operation.
- If you focus on the key region, you should ensure full coverage of all entities targeted for visual operations.
- All image operations are based on the original image.
- Ensure that each round of logical reasoning and visual operation contributes to reaching the ground truth.

The absolute image size is ({w},{h}).
Question: {query}
Answer: {gt}

Figure 3: Prompt template for the reasoning path generator.

```
tools = [
    {
        "type": "function",
        "function": {
            "name": "focus_area",
            "description": "returns the bbox of the focused area in the original image using normalized
                            coordinates [ymin, xmin, ymax, xmax] in range of [0, 1000]",
            "parameters": {
                "type": "object",
                "properties": {
                    "coordinates": {
                        "type": "array",
                        "description": "[ymin, xmin, ymax, xmax]",
                        "items": {
                            "type": "number",
                            "minimum": 0,
                            "maximum": 1000,
                        },
                        "minItems": 4,
                        "maxItems": 4,
                    },
                },
                "required": ["coordinates"],
            },
        },
    },
    {
        "type": "function",
        "function": {
            "name": "zoom_in",
            "description": "Magnifies image area using interpolation while maintaining original data",
            "parameters": {
                "type": "object",
                "properties": {
                    "scale": {
                        "type": "number",
                        "minimum": 1,
                    },
                },
                "required": ["scale"],
            },
        },
    },
    {
        "type": "function",
        "function": {
            "name": "reuse",
            "description": "Reuses the image without any modifications to enhance visual memory",
        },
    },
]
```

Figure 4: Tool definitions in OpenAI's dialogue format.

Tool Verification Prompt Template (focus_area)

You are an expert evaluator who is responsible for checking whether the drawn red bounding box matches or encompass the region of focus in the mentioned visual operation.

- Returns "TRUE" if the red bounding box matches or encompasses the target area/entity.
- Returns “FALSE” if the red bounding box deviates too much from the target area/entity, or is not semantically aligned with the visual operation planning.

NOW EVALUATE THE FOLLOWING:
{res_text}

The normalized coordinate (range [0, 1000]) of drawn bbox: [{ymin}, {xmin}, {ymax}, {xmax}]
ANSWER WITH "TRUE" OR "FALSE", DO NOT INCLUDE OTHER WORDS!

Figure 5: Tool verification prompt when executing $focus_area$. The model determines whether the input image and the returned coordinates are semantically aligned with the visual operation planning.

Tool Verification Prompt Template (reuse & zoom_in)

You are an expert evaluator who is responsible for checking whether the output JSON argument is semantically consistent with the visual operation planning in the reasoning content? The reasoning process is composed of two components: the atomic reasoning step followed by visual operation planning.

- Returns “TRUE” if the output JSON argument is semantically consistent with the visual operation planning
- Returns “FALSE” if the output JSON argument is not semantically consistent with the visual operation planning

NOW EVALUATE THE FOLLOWING:
Reasoning Content: {res_text}
JSON Arguments: {arguments}

ANSWER WITH "TRUE" OR "FALSE", DO NOT INCLUDE OTHER WORDS!

Figure 6: Tool verification prompt when executing $reuse$ and $zoom_in$. The model determines whether the returned function is semantically aligned with the visual operation planning.

Answer Verification Prompt Template

You are an expert evaluator tasked with determining if a SYSTEM'S ANSWER matches the GROUND TRUTH. Please carefully analyze the ground truth and the system's answer based on the question.

Apply the following evaluation criteria, and respond TRUE if:
- Focus on semantic equivalence rather than exact wording
- Allow for paraphrases and different phrasings that convey the same meaning
- Ignore minor grammatical differences or punctuation variations
- For categorical answers, match must be precise unless synonyms are clearly equivalent
- For list-type answers, order doesn't matter unless specified

Input:
Question: [Input Question]
GROUND TRUTH: [The correct reference answer]
SYSTEM ANSWER: [The answer to evaluate]

Output - respond with exactly one word - either "TRUE" or "FALSE":
["TRUE"/"FALSE"]

Now evaluate the following:

Question: {question}
GROUND TRUTH: {gt}
SYSTEM ANSWER: {ans}

ANSWER WITH "TRUE" OR "FALSE", DO NOT INCLUDE OTHER WORDS!

Figure 7: Answer verification prompt. The model judges whether the the system output is accurate based on the question and ground truth.

A.5 System Prompt and Inference Cases for Simple o3

System Prompt for Simple o3

You are proficient in interaction with the user.

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": {"type": "function", "function": {"name": "focus_area", "description": "returns the bbox of the focused area in the original image using absolute coordinates [xmin, ymin, xmax, ymax]", "parameters": {"type": "object", "properties": {"coordinates": {"type": "array", "description": "[xmin, ymin, xmax, ymax]", "items": {"type": "number", "minimum": 0}, "minItems": 4, "maxItems": 4}}, "required": ["coordinates"]}}}}
{"type": "function", "function": {"type": "function", "function": {"name": "zoom_in", "description": "Magnifies image area using interpolation while maintaining original data", "parameters": {"type": "object", "properties": {"scale": {"type": "number", "minimum": 1}}, "required": ["scale"]}}}}
{"type": "function", "function": {"type": "function", "function": {"name": "reuse", "description": "Reuses the image without any modifications to enhance visual memory"}}}
</tools>

For each function call, return a json object with function name and arguments within <function></function> XML tags:
<function>
{"name": <function-name>, "arguments": <args-json-object>}
</function>

Figure 8: System prompt for Simple o3. Simple o3 calls the corresponding tool during the inference stage according to the definition, thereby operating the image to complete interleaved visual-language reasoning.

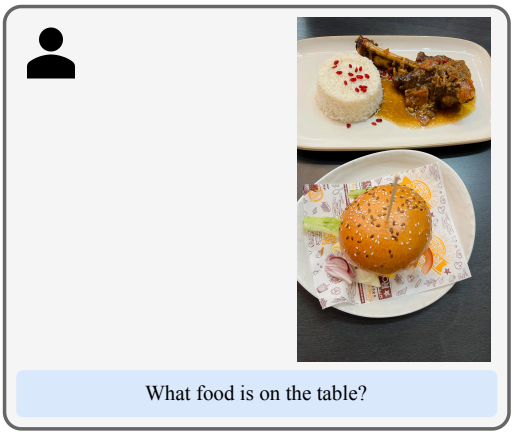
Tool Command Samples for Simple o3

#focus_area
{"name": "focus_area", "arguments": {"coordinates": [0, 386, 419, 643]}}

#zoom_in
{"name": "zoom_in", "arguments": {"scale": 2}}

#reuse
{"name": "reuse", "arguments": {}}

Figure 9: Examples of three tool commands in JSON format, including $focus_area$, $zoom_in$, and $reuse$.



What food is on the table?

<reasoning>
The question asks about the food on the table. The image shows two plates of food. To accurately identify all items, I will first focus on the top plate which contains what appears to be a main dish and rice.
</reasoning>
<function>
{"name": "focus_area", "arguments": {"coordinates": [0, 0, 923, 768]}}
</function>
<observation>

</observation>
<reasoning>
I have already identified the main dish as a lamb shank and rice from the previous observation. Now I need to identify the food on the bottom plate. I will focus on the bottom plate to get a clearer view of its contents.
</reasoning>
<function>
{"name": "focus_area", "arguments": {"coordinates": [0, 768, 923, 1679]}}
</function>
<observation>

</observation>
<reasoning>
I have successfully identified both dishes. The top plate has a lamb shank and rice. The bottom plate has a burger. I can now provide the complete answer.
</reasoning>
<answer>
There is a lamb shank and rice on the table, along with a burger.
</answer>

Figure 10: Multi-step interleaved vision-language reasoning of Simple o3 – Case 1. Simple o3 uses atomic reasoning steps to gradually analyze and accurately locate each entity.

Is there a cow in this image?

<reasoning>
The user is asking whether there is a cow in the image. I will first observe the entire image to get an overall sense of its content. The image appears to be a busy street scene with many buses and buildings. There are no obvious cows visible at first glance. To confirm this, I will reuse the image as it seems clear enough to make a judgment.
</reasoning>
<function>
{"name": "reuse", "arguments": {}}
</function>
<observation>

</observation>
<reasoning>
I have carefully observed the entire image and can confirm that there are no cows present. The image depicts a city street with numerous buses and buildings.
</reasoning>
<answer>
No
</answer>

Figure 11: Multi-step interleaved vision-language reasoning of Simple o3 – Case 2. Simple o3 can reconfirm the deduction by reusing the original image.

Is this a photo of Larvs kyrka?

<reasoning>
The question asks whether the image shows Larvs kyrka. I need to examine the text on the stone plaque to see if it belongs to Larvs Kyrka. The current image shows the entire plaque, but the text is small and difficult to read clearly. Therefore, I will zoom in on the text to make it legible.
</reasoning>
<function>
{"name": "zoom_in", "arguments": {"scale": 2}}
</function>
<observation>

</observation>
<reasoning>
I have already zoomed in once, but the text is still not clear enough to read. I need to zoom in further to be able to read the text on the stone plaque.
</reasoning>
<function>
{"name": "zoom_in", "arguments": {"scale": 4}}
</function>
<observation>

</observation>
<reasoning>
I have successfully zoomed in twice. Now the text on the stone plaque is much clearer, which is indeed the stone plaque on Larvs kyrka. This is a quote from the Bible, Psalm 144, verse 5. In English (KJV). This directly answers the question.
</reasoning>
<answer>
yes
</answer>

Figure 12: Multi-step interleaved vision-language reasoning of Simple o3 – Case 3. If Simple o3 detects that certain entities or fonts in an image are too small, it can enhance perception of the current image through iterative magnification.

Is this a photo of St Thomas Rest Park?

<reasoning>
The question asks whether the image shows St Thomas Rest Park. The image displays a large building with a distinctive architectural style, surrounded by greenery. To determine if it's St Thomas Rest Park, I need to identify specific features that might indicate its identity. Since there are no clear signs or distinguishing features visible in the initial view, I will first focus on the entire building to get a better overall sense of its characteristics.
</reasoning>
<function>
{"name": "focus_area", "arguments": {"coordinates": [0, 0, 531, 391]}}
</function>
<observation>

</observation>
<reasoning>
The image shows a large, old-looking building with multiple stories and dormer windows. It appears to be made of stone and brick, with some parts covered in ivy. There's a paved path leading up to the entrance. Given the question about "St Thomas Rest Park," which sounds like a park or recreational area, and not a typical historical building, I need to look for any clues that might link this building to such a place. However, based on the current view, there are no obvious signs or contextual elements that would definitively identify it as St Thomas Rest Park. Therefore, I will reuse the image as there is no specific area to focus on or zoom into that would provide more information to answer the question.
</reasoning>
<function>
{"name": "reuse", "arguments": {}}
</function>
<observation>

</observation>
<reasoning>
The image depicts an old, multi-story building with a distinctive architectural style, possibly a chateau or manor house, situated in a natural setting with grass and trees. There are no discernible features or signs within the image that explicitly identify it as "St Thomas Rest Park." The architecture does not align with common representations of a rest park, which typically feature open spaces, recreational areas, and modern infrastructure rather than grand historical buildings. Without additional context or information beyond what is visually present, it is impossible to confirm if this is St Thomas Rest Park.
</reasoning>
<answer>
No
</answer>

Figure 13: Multi-step interleaved vision-language reasoning of Simple o3 – Case 4. Simple o3 extends both text tokens and visual tokens simultaneously, further enhancing its reasoning capabilities.