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

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have achieved strong performance on vision–language tasks, yet often fail to preserve and effectively leverage visual evidence throughout generation. We identify three fundamental types of visual grounding failures: *Long-Context Grounding Error*, where visual embeddings gradually decay over long sequences; *Fine-Grained Grounding Error*, where low-resolution or degraded inputs hinder the recovery of detailed visual information; and *Regional Grounding Error*, where spatially diffuse attention weakens region-level vision–language alignment. To address these issues, we propose a tool-augmented reasoning framework with three targeted compensation strategies: *reuse*, which re-injects the original image to mitigate visual forgetting; *focus_area*, which constrains attention to task-relevant regions; and *zoom_in*, which enhances visual resolution for fine-grained perception. We further construct the TWI-Tools-146K dataset and develop SIMPLE-O3, a tool-augmented MLLM that interleaves visual and textual tokens. Extensive experiments show that each tool yields targeted improvements for its corresponding grounding error, while their combination produces synergistic gains in visual reasoning. Beyond performance, our analysis provides mechanistic insights into how tool-based interventions improve visual grounding, pointing toward more reliable multimodal reasoning.

1 Introduction

Multimodal large language models (MLLMs) have achieved impressive performance across a wide range of vision–language tasks, such as image captioning, visual question answering, and multimodal reasoning (Bai et al., 2025a; Chen et al., 2024b; Liu et al., 2024a; Li et al., 2024a; Team et al., 2025). However, despite these advances, existing MLLMs still suffer from systematic failures in basic visual understanding (Tong et al., 2024; Liu et al., 2025a; Sun et al., 2025b). As shown in Figure 1, models may misjudge spatial relations in traffic scenes, confuse relative positions of objects, or inaccurately count visible instances, revealing fundamental weaknesses in visual grounding.

These errors indicate that the primary bottleneck is not language generation, but the model's inability to consistently ground its reasoning in visual evidence (Shiri et al., 2024; Qharabagh et al., 2024; Tamarapalli et al., 2025; Guan et al., 2024). In most MLLMs, visual information is injected only once at the beginning of the multimodal sequence (Jiang et al., 2025; Arif et al., 2025; Chung et al., 2025). Although visual tokens remain accessible via attention, their influence diminishes as textual generation proceeds, making it difficult to preserve perceptual information over long reasoning chains (Jung et al., 2025; Liu et al., 2024c).

Visual grounding failures also arise in scenarios requiring high visual precision or accurate spatial localization (Tang et al., 2025; Bousselham et al., 2024). Low-resolution or degraded inputs can obscure fine-grained details, leading to incorrect answers even for simple visual queries (Usama et al., 2025). Moreover, attention over visual tokens is often spatially diffuse rather than concentrated on task-relevant regions (Esmaeilkhani and Latecki, 2025; Kang et al., 2025; Yu et al., 2024a), weakening region-level alignment between visual instances and textual references (Zeng et al., 2024b; Ouyang et al., 2025).

Taken together, these observations suggest that visual grounding failures in MLLMs are not monolithic, but arise from distinct mechanisms (Xiao et al., 2024a). We categorize them into three major types: *Long-Context Grounding Error*, where visual representations fade over extended generation (Sun et al., 2025a); *Fine-Grained Grounding Error*, where limited resolution or visual degradation prevents recovery of detailed information (He et al., 2025); and *Regional Grounding Error*, where

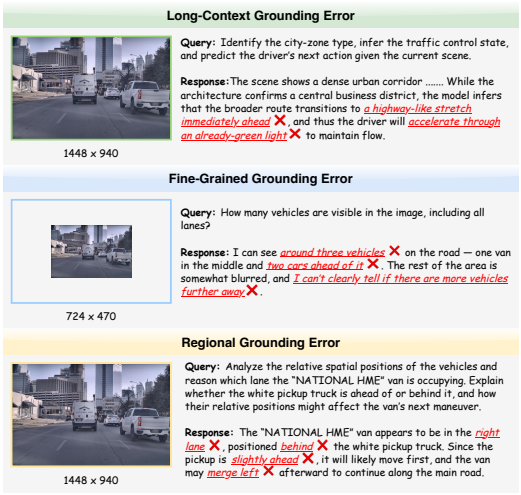

Figure 1: Examples of three categories of visual grounding errors in multimodal reasoning. The erroneous text highlighted in red.

spatially dispersed attention undermines precise region-level alignment (Cheng et al., 2024; Ma et al., 2024). These categories capture complementary failure modes in how models perceive and utilize visual evidence.

To mitigate these issues, we propose a tool-augmented compensation paradigm that directly targets each type of grounding error. We introduce three complementary strategies: *reuse*, which reinjects the original image to counteract visual forgetting; *zoom_in*, which increases visual resolution to support fine-grained perception; and *focus_area*, which constrains attention to task-relevant regions to enhance spatial alignment. Integrated into a deliberate, slow-thinking reasoning process, these tools enable MLLMs to preserve visual evidence throughout extended multimodal inference.

To facilitate systematic evaluation, we construct the TWI-Tools-146K dataset and develop SIMPLE-O3, a tool-augmented MLLM that interleaves visual and textual tokens. Extensive experiments show that each tool yields targeted improvements for its corresponding grounding error, while their combination leads to consistent and complementary gains. Beyond performance, our study provides mechanistic insights into visual grounding failures in MLLMs and offers practical guidance toward more reliable multimodal reasoning.

2 Related Work

2.1 Multimodal 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., 2024a). 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., 2025b; 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., 2025b; 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., 2025; Yu et al., 2025; Kumar et al., 2025). This lack of adaptability fails to meet

the demands of diverse scenarios, where different tasks may require more flexible visual processing strategies (Zhu et al., 2025; Chen et al., 2025).

3 Tool-wise Preliminary Analysis

To analyze the linguistic characteristics of the three grounding error types, we examine token-level text–image alignment using attention maps during decoding, as visualized in Figure 2. In the case of long-form generation, we compare the attention maps corresponding to the first and last occurrences of the token "building". While early decoding exhibits strong alignment with the relevant building regions, this alignment gradually weakens as generation progresses, with attention becoming dispersed, indicating a degradation of vision-language grounding consistency.

In addition, to study resolution-related grounding errors, we compare attention aggregation for the phrase "front car" under low- and high-resolution visual inputs, and find that low-resolution images lead to coarser and less accurate grounding, whereas higher-resolution inputs result in more precise token-level alignment. For region-level grounding, we analyze the attention of "white van" observing that well-performing models correctly localize attention on the target van, while weaker models are more susceptible to visual similarity and attend to semantically irrelevant regions such as white trucks.

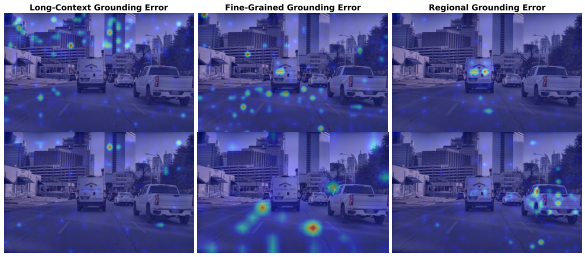

Figure 2: Attention maps revealing the root causes of the three types of grounding errors. Images below each column show deficiencies in the corresponding grounding capability.

We conduct controlled preliminary experiments to study three basic visual operation tools, namely *reuse*, *focus_area*, and *zoom_in*, under an isolated single-tool setting. Using the COCO (Chen et al., 2015) subset of LLaVA-CoT-100K (Xu et al., 2024), we construct tool-specific reasoning data where each sample consists of a single atomic reasoning step followed by a single tool invocation. For each tool, we curate 20K samples and perform supervised fine-tuning on Qwen2.5-VL, resulting in three tool-specialized models.

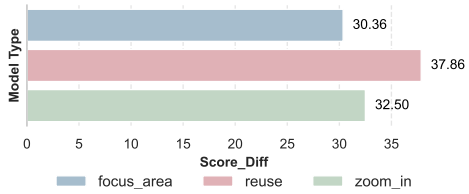

Figure 3: Improvement in performance of specific tool models over the base model on MME reasoning tasks.

Compared with the base model in Figure 3, all three tools consistently improve overall reasoning scores. In particular, the reuse model yields the strongest overall reasoning improvements, suggesting that explicit visual revisitation helps stabilize visual evidence during reasoning. These results indicate that even single-tool, single-step supervision can introduce a beneficial inductive bias for multimodal reasoning, motivating our subsequent exploration of multi-step and compositional tool-augmented reasoning.

4 Method

In this section, we propose visual grounding enhancement tools in Section 4.1. Then, we describe the scalable data generation pipeline in Section 4.2, as well as the training and multi-step inference procedures of Simple o3 in Section 4.3.

4.1 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. The model dynamically produces JSON-formatted tool commands C_t during reasoning, as shown in Figure 14, specifying function names and arguments. The executor parses these functions, performs operations on the original image, and appends the resulting image to the 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. 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. When focused attention on specific entities or regions is required, the model calls *focus_area*

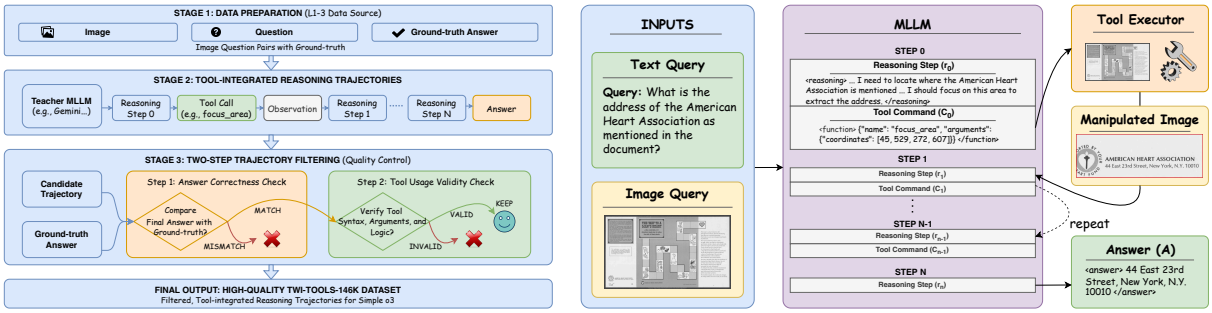

Figure 4: Overview of our scalable data generation pipeline (left) and Simple o3 (right). Left: The data generation process consists of data preparation, reasoning trajectory sampling, and a two-stage verification based on answer correctness and semantic validity. Right: The reasoning content $R_t = (r_t, p_t)$, comprising atomic reasoning steps r_t and visual planning p_t, is encapsulated within <*reasoning*> tags. Tool command $C_t = (T_n, \phi_t)$, specifying both operation names and input arguments, are marked by <*function*> tags.

with a bbox parameter, and the executor returns images of key areas. By adaptively integrating these visual compensation mechanisms throughout the reasoning process, the system achieves significantly enhanced visual grounding performance.

4.2 Scalable Data Generation Pipeline

We design a scalable data generation pipeline that integrates a reasoning-path generator with a two-stage verification module, as depicted in the left panel of Figure 4. During the generation phase, the MLLM acts as a controller that dynamically analyzes task requirements and coordinates three predefined visual manipulation tools: *focus_area*, *zoom_in*, and *reuse*.

Reasoning Path Generation We employ teacher MLLMs 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 denotes the reasoning chain comprising serialized atomic steps s, and the final answer A. Each step output $s_t = (R_t, C_t, I_t)$ includes the reasoning content R_t, the executed tool command C_t, and the resulting visual state I_t. The model samples each step conditioned on the query Q, input image I, and reasoning history H_{t-1}, enabling interleaved multimodal reasoning as the process unfolds. To ensure structural consistency, the pipeline encodes all reasoning components between XML-style tags.

Verification and Quality Control The verification module verifies answer correctness and tool usage validity through a two-step trajectory filtering process, ensuring both geometric and semantic validity of the generated samples. Answer correctness is evaluated by comparing generated responses against reference answers using a high-confidence MLLM-based evaluator. Besides, for *zoom_in* and *reuse*, we assess semantic consistency between the issued tool command C_t and its corresponding visual plan p_t. For *focus_area*, we retain instances where the predicted bounding box (bbox) accurately covers or fully contains the target entity. For each question–image pair, the system constructs a multi-step reasoning chain, invoking visual tools as needed, and performs up to two attempts per sample to ensure reliability. Detailed prompt templates are provided in Appendix A.6.

4.3 Training and Inference for Simple o3

We construct the TWI-Tools-146K dataset by curating data from MATHV360K (Shi et al., 2024), LLaVA-CoT-100K, and ArxivQA (Li et al., 2024b) through the scalable tool-integrated data generation pipeline. The details are provided in Appendix A.3. The model is trained 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 supervised 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

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 ⟨function⟩ ⟨/function⟩ 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 ⟨answer⟩ ⟨/answer⟩ detected in y **then** ▷ Terminate generation
    return final generated response y
  end if
end while
return final generated response y

with a red bbox as input, while during inference, the image is cropped using the corresponding bbox to inject focused visual tokens. We discuss this training–inference discrepancy in Appendix A.5 and Section 5.3.1. As shown in Algorithm 1 and on the right of Figure 4, 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. Some reasoning examples of Simple o3 are provided in Appendix A.7.

5 Experiments

5.1 Main Experiments

General Performance By delving into the performances in Table 1, we conclude that Simple o3 achieves significant improvements on multimodal reasoning benchmarks. It outperforms the base model by nearly 50 points on the MME (Fu et al., 2024) reasoning subset and surpasses GPT-4o (Hurst et al., 2024) by 27 points, underscoring the efficacy of 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 2, 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., 2024b) validate its advanced multimodal integration and complex 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, 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., 2025c) and VStarBench (Wu and Xie, 2024), highlighting its fine-grained perception capability in high-resolution scenarios where target objects occupy minimal pixels. As shown in Table 3, Simple o3 also excels in spatial reasoning tasks, particularly on subtasks of MMVet and LogicVista (Xiao et al., 2024b), 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.

Original Data Comparison To validate the effectiveness of tool integration, we compare our proposed TWI-Tools-146K dataset with the original data using identical queries, in Table 4. Our experiments reveal that direct SFT with the original 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

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 1: 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 2: 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 3: The performance of Simple o3 on spatially relevant benchmarks.

gains across all evaluation datasets. This improvement substantiates that our three proposed tools collectively enhance grounding capabilities, where iterative interactions between textual reasoning and critical visual information progressively strengthen the model's reasoning capacity.

Comparison with RL-based methods 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 5. 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 in POPE (Li et al., 2023b), it still improves on two hallucination benchmarks, especially by 2.3% in HallusionBench (Guan et al., 2024).

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/ Original	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 4: Performance comparison of dataset variants: original 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 5: Comparative performance of Simple o3 with RL-based interleaved vision-language reasoning methods: DeepEyes and Chain-of-Focus.

5.2 Ablation Study

Visual Grounding Tool Composition We progressively activate visual grounding tools in Simple o3 via system prompt customization to quantify their cumulative impact, as summarized in Table 6. While all components contribute positively, the *zoom_in* operation yields only marginal gains. In contrast, *focus_area* plays a critical role, particularly in fine-grained perception benchmarks such as HR-Bench 4K and VStarBench, where it improves performance by 4.9% and 6.9%, respectively. By enabling precise visual localization, *focus_area* filters irrelevant visual content while emphasizing task-relevant regions, leading to consistent performance improvements. Overall, these results demonstrate that the coordinated integration of visual grounding tools collectively elevates the model's reasoning upper bound.

Heterogeneous Training Data Effects We examine how heterogeneous training data influences model capability in Table 7. We first establish

a baseline by training on the general CoT reasoning dataset described in Appendix A.3, which yields modest but consistent gains. To selectively enhance specific abilities, we then progressively incorporate three specialized subsets from TWI-Tools-146K: MATHV360K-Tools, LLaVA-CoT-100K-Tools, and ArxivQA-Tools.

Introducing MATHV360K-Tools substantially broadens the data distribution with complex VQA, chart understanding, and logical reasoning tasks, resulting in a notable performance increase. We attribute this to strengthened modeling of inter-object relations, which refines the model's strategic planning for visual tool usage. However, this training mixture contains relatively fewer perceptual samples, limiting gains on fine-grained perception benchmarks. To address this gap, we further add LLaVA-CoT-100K-Tools, which significantly enhances fine-grained perceptual capability without degrading reasoning performance. Finally, despite its narrow distribution, ArxivQA-Tools directly reinforces atomic reasoning skills through challenging abstract diagrams from Arxiv papers, yielding substantial improvements in logical reasoning. Synergistically integrating heterogeneous data sources effectively strengthens the model's higher-order iterative reasoning ability.

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 6: Performance comparison of tool implementations variants in Simple o3: Only *reuse*, plus *zoom_in*, and with *focus_area*.

5.3 Further Analysis

5.3.1 Mechanisms Behind Visual Grounding Improvements

Beyond the preliminary experiments, we isolate *reuse* through system prompt control (Table 6) and find that visual reuse mitigates long-context degradation by counteracting visual token drift during autoregressive decoding. Re-encoding the image and reinjecting the original visual tokens restores degraded perceptual representations, re-aligns visual and linguistic contexts, and suppresses hallucination accumulation over long reasoning chains.

We further study the effect of input resolution in Table 8, which determines the effectiveness boundary of fine-grained grounding and directly constrains the behavior of *zoom_in*. Performance improves monotonically with higher resolution, confirming that high-quality visual input is a prerequisite for fine-grained perception. Under low resolution limits, *zoom_in* behaves similarly to *reuse* due to image compression; however, with sufficiently high resolution budgets, enlarged regions are injected without compression, providing dense visual evidence and fully activating fine-grained grounding. This interpolation-based resampling densifies pixel distributions and alleviates recognition errors caused by discretization artifacts.

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

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 8: 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.

Finally, we evaluate three execution modes of *focus_area* in Table 9. Cropping the specified region consistently yields the best performance, as it enforces spatial selectivity and injects task-relevant visual evidence. This localized visual focus shifts the model from global scanning to region-centered reasoning, strengthening regional grounding and factual accuracy.

5.3.2 Model Learns to Call Tools as Needed

As shown in Figure 5 and Figure 6, we have computed the average calling frequency of each type of tool and the average reasoning steps, both on the training data and on the reasoning trajectories from 6 benchmarks. Results reveal substantial variation in tool-calling behavior across different benchmark categories. For instance, HRBench-4K requires the model to identify small key entities within high-resolution images, resulting in a high frequency of

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

invoking the cropping tool.

In contrast, on InfoVQA, the model's fine-grained visual perception is often sufficient to support reasoning directly on the global image. In some cases, the model even prefers to reason purely at the textual level without invoking any tools, demonstrating the adaptive efficiency of Simple o3. This behavior is consistent with the principle articulated in (Wang et al., 2025a) that effective agents should solve tasks with the minimum necessary tool usage. Consequently, the tool-calling pattern during inference does not merely mirror the frequency distribution observed in the training data, indicating that the model has learned to invoke tools only when needed and can generalize this behavior across tasks.

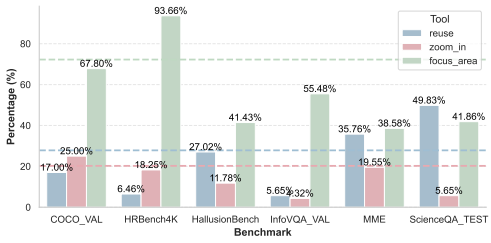

Figure 5: Average tool usage frequency of Simple o3 across different benchmarks, with the dashed line representing the corresponding frequency from the training data.

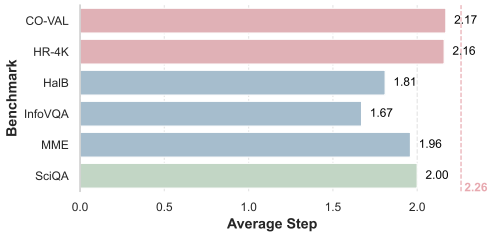

Figure 6: Average inference steps of Simple o3 across different benchmarks, with the dashed line representing the average steps from the training data.

5.3.3 Detailed Comparisons and Case Studies

Here, DeepEyes is re-run on several benchmarks under the same experimental setup. Under this controlled re-evaluation, Simple o3 demonstrates an even larger advantage over DeepEyes in Table 10. A key issue we observe is that DeepEyes frequently produces malformed or abnormal bbox, as shown in Figure 20, which repeatedly interrupts tool execution and leads to evaluation failures on MME and HallusionBench. In contrast, Simple o3 does not exhibit this failure mode and maintains stable tool invocation. While prior work suggests RL-based methods may generalize better in certain settings (Chu et al., 2025; Myers et al., 2025), we argue that such generalization does not necessarily translate into visual reasoning tasks that require precise bbox or structured outputs. We hypothesize that the abnormal bbox output in DeepEyes may reflect reward hacking induced by format-centric rewards commonly used in verifiable-reward RL.

Despite consistent improvements over the base model, as shown in Figure 19, Simple o3 still exhibits failures in certain scenarios, shown in Figure 21 to Figure 24. In fine-grained OCR cases, the model may correctly extract text during intermediate reasoning but fail to consistently use it in the final decision, suggesting insufficient interleaved image-text OCR-centric trajectories in training data. Incorporating more OCR-based reasoning samples may improve this failure mode. Additionally, the model sometimes over-relies on perceptual evidence while under-utilizing its prior knowledge. For example, when it fails to locate "1998" in the image, it may become overly cautious and answer incorrectly, indicating brittle behavior when salient visual cues are missing. We also present two case studies in which the model fails without using these tools, but succeeds when the tools are employed in Figure 25 and Figure 26.

6 Conclusion

In this work, we investigate how MLLMs lose visual grounding fidelity during multimodal reasoning and identify three fundamental failure modes: *Long-Context Grounding Error*, *Fine-Grained Grounding Error*, and *Regional Grounding Error*. To mitigate these issues, we introduce *reuse*, *zoom_in*, and *focus_area* to respectively preserve visual representations throughout the reasoning process. Built upon the TWI-Tools-146K dataset, our model SIMPLE-O3 achieves consistent improvements in visual grounding and multimodal reasoning. We provide a principled framework for diagnosing and mitigating grounding degradation in MLLMs, and lays the groundwork for developing more reliable multimodal reasoning systems.

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.

Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, and Chris Thomas. 2025. Paint: Paying attention to informed tokens to mitigate hallucination in large vision-language model. *arXiv preprint arXiv:2501.12206*.

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*.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. 2024. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837.

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*.

Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. 2025. Sifthinker: Spatially-aware image focus for visual reasoning. *arXiv preprint arXiv:2508.06259*.

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*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

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.

Parsa Esmaeilkhani and Longin Jan Latecki. 2025. Direct visual grounding by directing attention of visual tokens. *arXiv preprint arXiv:2511.12738*.

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.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. 2025. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. *arXiv preprint arXiv:2501.15140*.

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*.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. 2025. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. *arXiv preprint arXiv:2502.01419*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350.

Sunil Kumar, Bowen Zhao, Leo Dirac, and Paulina Varshavskaya. 2025. Reinforcing vlms to use tools for detailed visual reasoning under resource constraints. *Preprint*, arXiv:2506.14821.

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.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and 1 others. 2025a. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9087–9097.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. 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*.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer.

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.

Vivek Myers, Catherine Ji, and Benjamin Eysenbach. 2025. Horizon generalization in reinforcement learning. *arXiv preprint arXiv:2501.02709*.

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*.

Shuyi Ouyang, Ziwei Niu, Hongyi Wang, Yen-Wei Chen, and Lanfen Lin. 2025. Region-aware anchoring mechanism for efficient referring visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24192–24202.

Muhammad Fetrat Qharabagh, Mohammadreza Ghofrani, and Kimon Fountoulakis. 2024. Lvlm-count: Enhancing the counting ability of large vision-language models. *arXiv preprint arXiv:2412.00686*.

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*.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455.

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

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. 2025a. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. *arXiv preprint arXiv:2503.13360*.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. 2025b. Math blind: Failures in diagram understanding undermine reasoning in mllms. *Preprint*, arXiv:2503.20745.

Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, and Sahiti Yerramilli. 2025. Countqa: How well do mllms count in the wild? *arXiv preprint arXiv:2508.06585*.

Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. 2025. Visual position prompt for mllm based visual grounding. *arXiv preprint arXiv:2503.15426*.

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*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. 2025. Analysing the robustness of vision-language-models to common corruptions. *arXiv preprint arXiv:2504.13690*.

Hongru Wang, Cheng Qian, Manling Li, Jiahao Qiu, Boyang Xue, Mengdi Wang, Heng Ji, and Kam-Fai Wong. 2025a. Toward a theory of agents as tool-use decision-makers. *arXiv preprint arXiv:2506.00886*.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025b. 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. 2025c. 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.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2024a. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024b. 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*.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024a. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pages 251–268. Springer.

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

Xuan Yu, Dayan Guan, and Yanfeng Gu. 2025. Zoom-refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. *arXiv preprint arXiv:2506.01663*.

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. 2024a. 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.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. 2024b. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14151.

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 $< |$*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.

A.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). From MATHV360K, we first exclude subtasks involving pure mathematical problem solving, as geometry-oriented reasoning primarily emphasizes the extension of textual tokens. Using proportional sampling over the remaining sub-datasets, we initially collect 150K candidate samples. After rigorous verification, we retain approximately 100K high-quality samples from MATHV360K, covering perceptual VQA, knowledge-based reasoning, chart analysis, and logical reasoning, representing a comprehensive collection of challenging 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. Moreover, 19K proportionally balanced samples are extracted from ArxivQA to further enhance abstract figure interpretation and reasoning, covering all subcategories. The resulting dataset comprises approximately 146K samples. For consistency, all bbox coordinates are converted to the Qwen2.5-VL-compatible format.

A.4 Setups

A.4.1 Implementation Details

We perform supervised 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.

Model	MME	CharXiv	COCOC	HR4K	V*	POPE	MMVet	HalB	InfoVQA	VisLog
Qwen2.5-VL	652.5	37.6	15.0	68.8	77.5	86.0	68.2	42.1	80.0	26.0
Simple o3	**702.1**	**41.8**	**18.1**	**76.2**	**84.3**	87.6	**69.7**	**44.4**	**82.0**	**29.3**
DeepEyes	F	37.3	17.2	75.4	73.3	**87.9**	67.7	F	81.3	27.0

Table 10: Re-evaluation of DeepEyes under the same experimental setting and comparison with Simple o3. **F** indicates evaluation failure: an excessive number of abnormal predicted bbox results in too many invalid samples, preventing the evaluation from being completed.

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.4.2 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., 2025c), 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., 2024b). 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.5 Manual Evaluation of Synthesized Samples

We manually evaluated 300 synthesized samples, and all sample answers matched the ground truth. However, 24 samples did not fully contain the corresponding entities, as shown in Figure 7. Although the bbox center pointed to the key entity, directly cropping would result in partial or incomplete entities, leading to the loss of necessary visual information. This confirms that providing the full image with bbox during training is necessary. Otherwise, the model may become confused when receiving incomplete entities. Regarding semantic consistency, defined as whether the reasoning content matches the function name, 7 samples exhibited confusion between *zoom_in* and *focus_area*. In these cases, the reasoning intended to crop the key entity, but *zoom_in* was selected.

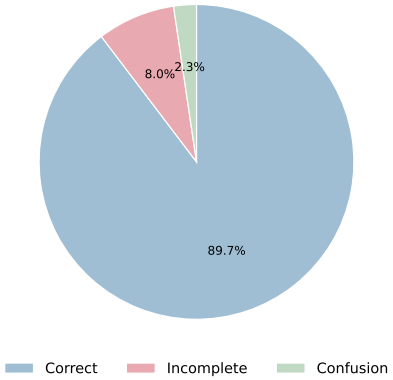

Figure 7: Statistics of the manual evaluation on the synthesized data. Although all answers are correct, the red portion represents samples where the *focus_area* bbox fail to fully cover the target entities, leading to potential visual information loss when cropped. The blue portion denotes inconsistencies between the reasoning and the invoked tool.

A.6 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 8: 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 9: Tool definitions in OpenAI's dialogue format for the reasoning path generator.

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 10: 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 11: 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 12: Answer verification prompt. The model judges whether the the system output is accurate based on the question and ground truth.

A.7 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 13: 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 14: 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 15: 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 16: 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 17: 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 18: 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.

A.8 Detailed Case Study and Comparison

Qwen2.5-VL vs Simple o3

Query: Is this photo taken in a place of youth hostel? Please answer yes or no.\n<image>
Ground Truth: No

Response:
Qwen-2.5-VL-7B: Yes
Simple o3:
<reasoning>
The question asks whether the photo was taken in a youth hostel. To determine this, I need to look for specific features commonly found in hostels, such as shared rooms, basic amenities, or signs indicating a budget lodging establishment. The current view shows a single bed in what appears to be a private room. There's nothing immediately obvious that suggests it's a youth hostel. Therefore, I will reuse the image to confirm if there are any other clues.
</reasoning>
<function>
{\"name\": \"reuse\", \"arguments\": {}}
</function>
<reasoning>
Based on the observation of a single bed in a room with a window and a small table, there is no visual evidence to suggest that this is a youth hostel. Youth hostels typically have multiple beds in shared rooms, communal areas, and often a more utilitarian or budget-oriented appearance. This room appears to be a private bedroom, which is not characteristic of a youth hostel.
</reasoning>
<answer>
No
</answer><|im_end|>

Figure 19: The experiment shows that grounding errors are improved in Simple o3 compared to Qwen2.5-VL.

DeepEyes Error Cases

Error Case 1:
raise ValueError(msg)
ValueError: Coordinate 'lower' is less than 'upper'

Error Case 2:
raise ValueError(msg)
ValueError: absolute aspect ratio must be smaller than 200, got 316.0

Figure 20: DeepEyes outputting incorrect bbox leads to errors. Case 1: The parsed bbox has a width or height of 0. Case 2: Abnormal aspect ratio.

Common Failure Case 1

Query: Is the word in the logo \"angle's\"? Please answer yes or no.\n<image>
Ground Truth: No

Simple o3 Response:
<reasoning>
The question asks whether the word in the logo is \"angle's\". To verify this, I need to clearly see the text in the logo. The current image shows the entire logo, but the text might be small. Therefore, I will focus on the area containing the text.
</reasoning>
<function>
{\"name\": \"focus_area\", \"arguments\": {\"coordinates\": [56, 55, 278, 393]}}
</function>
<reasoning>
The previous operation focused on the text area. Now, I can clearly read the word \"ANGIE'S\" in the logo. This directly answers the question.
</reasoning>
<answer>
Yes
</answer><|im_end|>

Figure 21: Common Failure Mode 1: Recognition errors in OCR scenarios (Case 1). *Observation* in the model response has been omitted.

Common Failure Case 2

Query: Is the word in the logo \"beavertalls pastry\"? Please answer yes or no.\n<image>
Ground Truth: No

Simple o3 Response:
<reasoning>
The question asks whether the word in the logo is \"beavertails pastry\". To answer this, I need to clearly read the text on the sign. The current image shows the entire sign, but the text might be small. Therefore, I will focus on the area of the sign where the text is located.
</reasoning>
<function>
{\"name\": \"focus_area\", \"arguments\": {\"coordinates\": [123, 159, 426, 278]}}
</function>
<reasoning>
I have already focused on the relevant area. Now, I will zoom in further to ensure that the text \"BEAVERTAILS\" and \"PASTRY\" is clearly legible.
</reasoning>
<function>
{\"name\": \"zoom_in\", \"arguments\": {\"scale\": 2}}
</function>
<reasoning>
The image now clearly displays the words \"BEAVERTAILS\" and \"PASTRY\" on the sign. This confirms that the word in the logo is indeed \"beavertails pastry\".
</reasoning>
<answer>
Yes
</answer><|im_end|>

Figure 22: Common Failure Mode 1: Recognition errors in OCR scenarios (Case 2). *Observation* in the model response has been omitted.

Common Failure Case 3

Query: Is this movie directed by adam robitel? Please answer yes or no.\n<image>
Ground Truth: No

Simple o3 Response:
<reasoning>
The question asks whether the movie is directed by Adam Robitel. To answer this, I need to identify the director of the movie shown in the image. The image displays a movie poster for \"Jurassic Park III\". I will focus on the title of the movie to see if there is any information about the director.
</reasoning>
<function>
{\"name\": \"focus_area\", \"arguments\": {\"coordinates\": [87, 56, 299, 135]}}
</function>
<reasoning>
The previous observation was not sufficient to determine the director's name. I need to zoom in further on the title to see if there is any text indicating the director.
</reasoning>
<function>
{\"name\": \"zoom_in\", \"arguments\": {\"scale\": 2}}
</function>
<reasoning>
The previous attempts to find the director's name from the movie poster were unsuccessful. Since the image does not contain the director's name, I cannot answer the question based solely on the provided image. Therefore, I will state that the information is not available.
</reasoning>
<answer>
unanswerable
</answer><|im_end|>

Figure 23: Common Failure Mode 2: Over-reliance on perceptual reasoning (Case 3). *Observation* in the model response has been omitted.

Common Failure Case 4

Query: Is this movie titled apt pupil (1998)? Please answer yes or no.\n<image>
Ground Truth: Yes

Simple o3 Response:
<reasoning>
The question asks whether the movie title is \"Apt Pupil (1998)\". The OCR output clearly shows \"Apt Pupil\" as part of the text on the poster. However, it does not explicitly state \"(1998)\". To confirm if the year is present, I need to focus on the area around the title.
</reasoning>
<function>
{\"name\": \"focus_area\", \"arguments\": {\"coordinates\": [76, 235, 256, 302]}}
</function>
<reasoning>
The previous observation confirmed that \"Apt Pupil\" is indeed visible. However, the year \"(1998)\" is not present in the current view. Therefore, I need to zoom out to check the entire poster for the year.
</reasoning>
<function>
{\"name\": \"reuse\", \"arguments\": {}}
</function>
<reasoning>
The previous attempts to find the year have been unsuccessful. Upon closer inspection of the original image, there is no explicit mention of the year \"(1998)\" anywhere on the poster. The title \"Apt Pupil\" is clearly visible, but the year is missing from the provided information.
</reasoning>
<answer>
No
</answer><|im_end|>

Figure 24: Common Failure Mode 2: Over-reliance on perceptual reasoning (Case 4). *Observation* in the model response has been omitted.

w/o Tools vs w/ Tools in Simple o3 Case 1

Query: What instrument does Yo-Yo play?\nOptions:\nA. the violin\nB. the cello\nPlease select the correct answer from the options above. \n<image>
Ground Truth: B

Before using tools:
<reasoning>
The man in the picture is holding a violin. This man is Yo-Yo, so the answer should be A.
</reasoning>
<answer>
A. the violin
</answer>

After using tools:
<reasoning>
The question asks what instrument Yo-Yo plays. The image shows a man holding a large string instrument next to him. To identify the instrument, I need to focus on it.
</reasoning>
<function>
{\"name\": \"focus_area\", \"arguments\": {\"coordinates\": [168, 129, 353, 559]}}
</function>
<reasoning>
The previous observation allowed me to see the instrument more clearly. It appears to be a cello. Therefore, I can confidently answer the question.
</reasoning>
<answer>
B. the cello
</answer><|im_end|>

Figure 25: Case 1 in which if not using these tools, the model fails, and using tools leads to success.

w/o Tools vs w/ Tools in Simple o3 Case 2

Query: Is it appropriate to translate the Chinese in the image into English 'a difficult work' in the picture? Please answer yes or no.\n<image>
Ground Truth: Yes

Before using tools:
<reasoning>
The user asked if the Chinese text in the image was suitable to be translated into the English phrase "a difficult work," and the image contained some Chinese characters.
</reasoning>
<answer>
No
</answer>

After using tools:
<reasoning>
The user is asking if the translation \"a difficult work\" is appropriate for the Chinese text in the image. To answer this question, I need to understand what the Chinese text means. The current image shows the entire text clearly, so I will reuse the image to confirm the meaning.
</reasoning>
<function>
{\"name\": \"reuse\", \"arguments\": {}}
</function>
<reasoning>
The Chinese characters in the image are \"\u56f0\u96be\u7684\u5de5\u4f5c\". This translates to \"difficult work\" or \"hard work\". Therefore, the translation \"a difficult work\" is appropriate.
</reasoning>
<answer>
Yes
</answer><|im_end|>

Figure 26: Case 2 in which if not using these tools, the model fails, and using tools leads to success.