SimpleOCR: Rendering Visualized Questions to Teach MLLMs to Read

Yibo Peng[1,2†*], **Peng Xia**[1*], **Ding Zhong**[1,3†*], **Kaide Zeng**[1*], **Siwei Han**[1]
Yiyang Zhou[1], **Jiaqi Liu**[1], **Ruiyi Zhang**[4], **Huaxiu Yao**[1]
[1]UNC-Chapel Hill, [2]Carnegie Mellon University, [3]University of Michigan, [4]Adobe Research
yibop@andrew.cmu.edu,dingdd@umich.edu,{pxia,kdzeng,huaxiu}@cs.unc.edu

Abstract

Despite the rapid advancements in Multimodal Large Language Models (MLLMs), a critical question regarding their visual grounding mechanism remains unanswered: do these models genuinely "read" text embedded in images, or do they merely rely on parametric shortcuts in the text prompt? In this work, we diagnose this issue by introducing the Visualized-Question (VQ) setting, where text queries are rendered directly onto images to structurally mandate visual engagement. Our diagnostic experiments on Qwen2.5-VL reveal a startling capability-utilization gap: despite possessing strong OCR capabilities, models suffer a performance degradation of up to 12.7% in the VQ setting, exposing a deep-seated "modality laziness." To bridge this gap, we propose SimpleOCR, a plug-and-play training strategy that imposes a structural constraint on the learning process. By transforming training samples into the VQ format with randomized styles, SimpleOCR effectively invalidates text-based shortcuts, compelling the model to activate and optimize its visual text extraction pathways. Empirically, SimpleOCR yields robust gains without architectural modifications. On four representative OOD benchmarks, it surpasses the base model by 5.4% and GRPO based on original images by 2.7%, while exhibiting extreme data efficiency, achieving superior performance with 30x fewer samples (8.5K) than recent RL-based methods. Furthermore, its plug-and-play nature allows seamless integration with advanced RL strategies like NoisyRollout to yield complementary improvements. Code is available at https://github.com/aiming-lab/SimpleOCR.

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in visual reasoning by integrating vision encoders with large language models (Liu et al., 2023; Bai et al., 2023, 2025; Hurst et al., 2024; Comanici et al., 2025). Central to this capability is optical character recognition (OCR), i.e., the ability to extract and interpret text embedded in images, which underpins performance on chart understanding (Masry et al., 2022; Wang et al., 2024c), document analysis (Mathew et al., 2021; Han et al., 2025; Mathew et al., 2022), and geometry-centric reasoning (Lu et al., 2021, 2023). While current MLLMs achieve strong performance on standalone OCR benchmarks, a fundamental question remains underexplored: *do these models actually leverage their OCR capabilities when solving downstream tasks?*

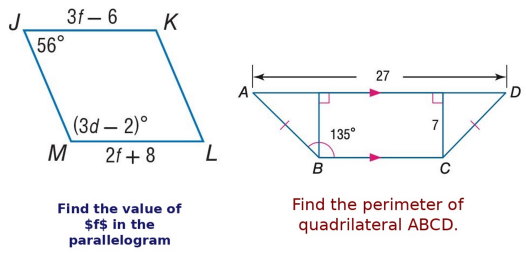
To investigate this, we introduce a controlled diagnostic intervention called the *visualized-question* (VQ) format. In standard evaluation, models receive questions via text, which may allow reasoning based on linguistic priors or parametric shortcuts rather than visual evidence. In the VQ setting, we render the question text directly onto the image and provide only a generic instruction (e.g., "`Please answer the question in the image`"), forcing the model to ground its reasoning in visual text. If a model fully utilizes its OCR capabilities, performance under both settings should be comparable. However, our experiments reveal a striking *capability–utilization gap*.

As shown in Figure 1, Qwen2.5-VL-7B suffers substantial degradation under the VQ setting, with an average absolute drop of 6.9% across four multimodal reasoning benchmarks and a maximum drop of 12.7% on WeMath (Qiao et al., 2024). This phenomenon aligns with recent observations of "modality laziness" (Lin et al., 2023; Fu et al., 2025; Yao et al., 2025), where models systematically underweight visual evidence when informative text prompts are available.

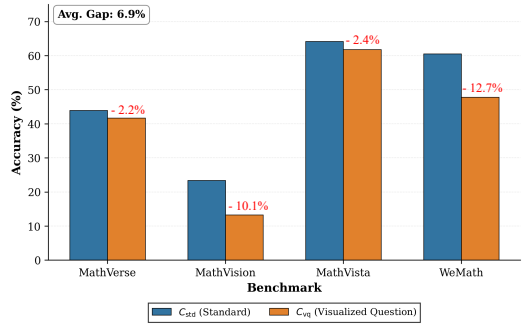
Motivated by this diagnosis, we propose **SimpleOCR**, a training strategy that addresses this gap through *structural constraint*. Rather than

*Equal Contribution. †Work was done during the internship at UNC

Findings of the Association for Computational Linguistics: ACL 2026, pages 10697–10710



(a) Visualized-Question (VQ) Format



(b) Capability-Utilization Gap on Benchmarks

Figure 1: (a) **Visualized-Question (VQ) Format.** We render the question text into the image as the only question source, removing text-channel shortcuts and requiring visual reading. (b) **Capability–Utilization Gap.** On Qwen2.5-VL-7B, performance drops markedly under VQ versus standard inputs, indicating that OCR capability is not reliably utilized during reasoning.

auxiliary losses or architectural modifications (Yu et al., 2025a; Cao et al., 2025; Sarch et al., 2025), SimpleOCR operates purely through input transformation: all training samples are converted to VQ format with randomized visual styles, eliminating text-based shortcuts entirely. Notably, SimpleOCR introduces zero additional computational overhead or inference latency. By embedding questions directly into the visual space, it forces the model to decode image-based prompts prior to reasoning, thereby drastically improving OCR-based understanding. As a plug-and-play strategy, SimpleOCR can be seamlessly incorporated into any VLM training framework, enhancing model robustness and reasoning by enriching the visual distribution of training data.

Empirically, SimpleOCR induces robust performance gains across both in-domain (ID) and out-of-distribution (OOD) scenarios. When trained on Geo3K and MMK12, SimpleOCR achieves a 6.6% improvement over the base model on ID test sets, and achieves 8.5% compared to GRPO based on original images. The generalization capability of our approach is substantiated by results on challenging OOD benchmarks. On MathVerse, MathVision, MathVista, WeMath, and HallusionBench, SimpleOCR surpasses the base model by 5.4% and GRPO based on original images by 2.7%. Notably, SimpleOCR exhibits extreme data efficiency: with only 8.5K training samples, it outperforms RL-based methods (Zhang et al., 2025a; Yang et al., 2025b) that require over 260K samples, demonstrating a 30x reduction in data dependency. Furthermore, SimpleOCR is a plug-and-play strategy that requires no modifications to the model architecture or training paradigms. It integrates seamlessly with existing VLM training frameworks. For instance, when combined with RL methods like NoisyRollout (Liu et al., 2025b), it yields complementary gains, confirming that SimpleOCR enhances a unique and orthogonal dimension of multimodal reasoning.

Our primary contribution is SimpleOCR, a plug-and-play training strategy designed to bridge the OCR *capability–utilization* gap. By imposing structural constraints, SimpleOCR forces models to actively engage with visual text, effectively addressing the performance degradation (up to 12.7%) seen when text shortcuts are removed. Empirical results across multiple multimodal reasoning benchmarks demonstrate that our approach significantly enhances out-of-distribution generalization. Furthermore, we verify the effectiveness of our structural components and demonstrate the broad compatibility of SimpleOCR with existing multimodal architectures.

2 Related Work

Reinforcement Learning for MLLMs. Reinforcement Learning from Verifiable Rewards (RLVR) advances multimodal reasoning by utilizing programmatic signals rather than subjective preferences, extending the RLHF paradigm (Ouyang et al., 2022; Yu et al., 2024; Wang et al., 2025a; Tu et al., 2025; Xia et al., 2025a,b; Liu et al., 2025a; Su et al., 2025; Xia et al., 2026; Yang et al., 2025a). The GRPO algorithm (Shao et al., 2024) has powered frontier models like DeepSeek-R1 (Guo et al., 2025), with recent adaptations refining the framework through diverse mechanisms. Specifically, R1-Onevision (Yang et al., 2025b) and Vision-R1 (Huang et al., 2025) optimize cross-modal formalization and training dynamics, respectively, while R1-VL (Zhang et al., 2025a)

and VLAA-Thinker (Chen et al., 2025) introduce step-wise rewards and mixed perception-cognition signals. To enhance stability, MM-Eureka (Meng et al., 2025) and ThinkLite-VL (Wang et al., 2025b) employ data-centric strategies such as rejection sampling and MCTS-based selection. Then NoisyRollout (Liu et al., 2025b) targets policy diversity by mixing distorted trajectories. However, these methods primarily focus on logical derivation or robustness, lacking explicit constraints to enforce visual text reading against shortcut learning.

Visual Grounding in Text-Rich Contexts. The paradigm for text-rich understanding has shifted from modular OCR pipelines to unified end-to-end architectures (Bai et al., 2025; Zeng et al., 2025; Li et al., 2024a; Zhang et al., 2025b). To circumvent resolution constraints, Monkey (Li et al., 2024b) and TextMonkey (Liu et al., 2024) introduced patch-division strategies, while VisInContext (Wang et al., 2024a) leveraged visual tokens to efficiently scale context length. Subsequently, architectures like GOT (Wei et al., 2024) and Donut (Blecher et al., 2023) unified perception and reasoning. Current state-of-the-art models, including Qwen2.5-VL (Bai et al., 2025), MiniCPM-V 4.5 (Yu et al., 2025b), and HunyuanOCR (Team et al., 2025a), leverage large-scale OCR corpora (Geng et al., 2025) and native-resolution ViTs (Dosovitskiy, 2020) to handle complex layouts. Despite these advances in *capability acquisition*, a critical dichotomy remains: models possess strong OCR capabilities but suffer from systematic "modality laziness" (Fu et al., 2025; Yao et al., 2025), failing to utilize visual evidence during reasoning. Unlike prior works, our work targets *capability utilization*, ensuring the model actively grounds its reasoning in visual text evidence.

3 Preliminaries

In this section, we will provide a brief overview of MLLMs and GRPO algorithm. We build upon Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a reinforcement learning framework designed to improve the reasoning ability of large language models.

Given a multimodal question q, consisting of an image x_{img} and a text prompt q_{text}, the policy model π_θ generates a reasoning response o. For each question q, GRPO samples a group of G candidate responses $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. Each response o_i is assigned a reward r_i (e.g., from a reward model or rule-based verifier). The group-relative advantage $\hat{A}_i$ for each response is then computed by:

$$\hat{A}_i = \frac{r_i - \frac{1}{G}\sum_{j=1}^{G} r_j}{\text{std}(r_1, \ldots, r_G)}, \tag{1}$$

which centers and normalizes the rewards within the group, effectively removing question-level biases.

The policy model π_θ is updated by maximizing the GRPO objective, which incorporates a PPO-style clipped surrogate loss and a KL divergence penalty against a frozen reference model π_{ref}:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Bigg(\\ \min\Big(r_i(\theta)\hat{A}_i, \text{clip}\left(r_i(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\Big)\\ -\beta D_{KL}(\pi_\theta\|\pi_{\text{ref}})\Bigg)\Bigg],\end{aligned} \tag{2}$$

where $r_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ denotes the probability ratio. The hyperparameters ϵ and β represent the clipping range and the KL divergence penalty weight, respectively. By bypassing the value function and utilizing group-relative advantages, GRPO significantly optimizes memory usage and training efficiency while maintaining robust performance.

4 SimpleOCR: Addressing the Gap Through Visual Question Training

4.1 Visual Question Setting

Given a training sample $S = (\mathbf{x}_{\text{img}}, q_{\text{text}})$, we define two informationally equivalent yet structurally distinct input contexts.

Standard Context C_{orig}. This context preserves the conventional multimodal schema, $C_{\text{orig}} = (\mathbf{x}_{\text{img}}, q_{\text{text}})$, where the question is provided via the text channel.

Visual Question Context C_{vq}. To structurally enforce visual grounding, we introduce a transformation $\mathcal{T}_{\text{render}}$ that embeds the semantic content of q_{text} directly into the visual modality:

$$C_{\text{vq}} = \left(\mathcal{T}_{\text{render}}(\mathbf{x}_{\text{img}}, q_{\text{text}}),\quad p_{\text{prompt}}\right) \tag{3}$$

where p_{prompt} is a generic instruction (e.g., "`Answer the question in the image`"). By removing q_{text}

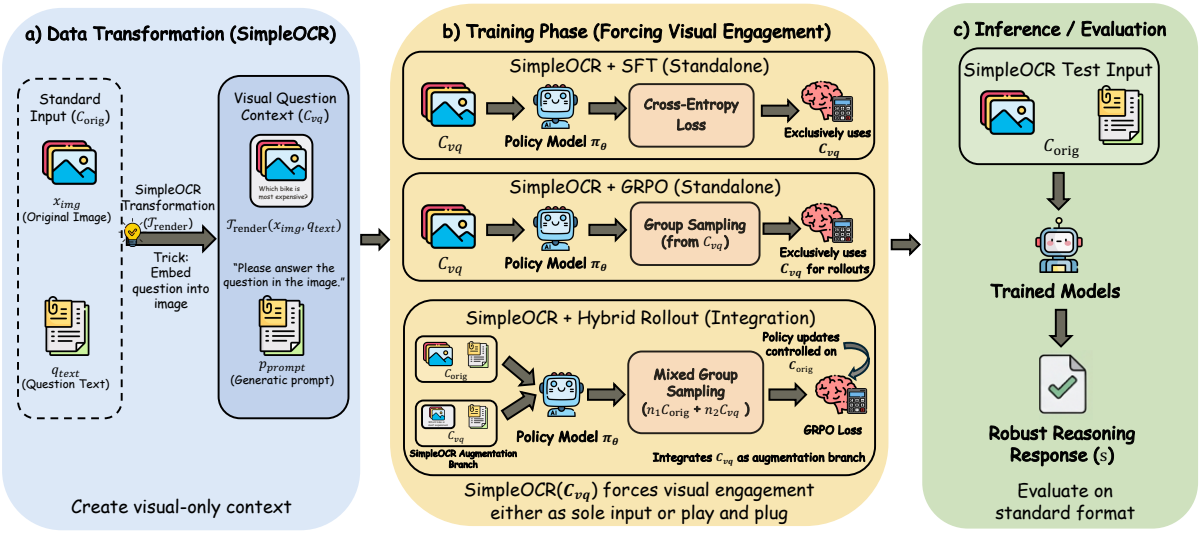


Figure 2: The SimpleOCR framework. During training, all inputs are transformed into visual question contexts C_{vq}, where question text is rendered onto images. This structurally eliminates text-based shortcuts and forces visual OCR engagement. At inference, models trained this way demonstrate robust performance on standard inputs C_{orig}. The method integrates seamlessly as an augmentation branch in existing RL frameworks.

Algorithm 1 Visual Question Rendering ($\mathcal{T}_{\text{render}}$)

```
# x: original image, q: question text
def render(x, q):
    # Sample random style (language-aware)
    font, color ← random_style()
    size ← random.randint(18, 42)

    # Wrap text and create canvas
    lines ← wrap(q, width=x.width, size=size)
    h ← len(lines) × line_height(size)
    canvas ← Image.new((x.width, x.height + h), white)

    # Paste original image and draw text
    canvas.paste(x, (0, 0))
    draw(canvas, lines, font, size, color, y=x.height)
    return canvas
```

from the text channel, C_{vq} eliminates the possibility of text-based shortcuts, making visual text reading structurally necessary.

As detailed in Algorithm 1, $\mathcal{T}_{\text{render}}$ appends the question text to a canvas region below the original image, ensuring all original visual features are preserved. To prevent the model from overfitting to specific layouts, we employ a randomized rendering strategy: parameters such as font family (with CJK support), color, and size (dynamically scaled between 18–42pt) are sampled stochastically during training. This diversity ensures that the learned OCR capabilities are robust to varying visual presentations.

Algorithm 2 SimpleOCR Training Strategy

Require: Dataset $\mathcal{D}$, Policy π_θ, Reference π_{θ_0}, Renderer $\mathcal{T}_{\text{render}}$
Ensure: Optimized Policy π_θ

for each batch $(x_{\text{img}}, q_{\text{text}}, a) \in \mathcal{D}$ **do**
  ▷ *1. Construct Visual Question Context*
  $x_{\text{render}} \leftarrow T_{\text{render}}(x_{\text{img}}, q_{\text{text}})$
  $C_{\text{vq}} \leftarrow (x_{\text{render}}, p_{\text{prompt}})$
  ▷ *2. Group Sampling (Visual Exploration)*
  Sample G outputs from visual context: $\{s_1, \ldots, s_G\} \sim \pi_\theta(\cdot|C_{\text{vq}})$
  ▷ *3. Advantage Computation*
  for $k = 1$ to G **do**
    Compute reward r_k comparing s_k to ground-truth a
  end for
  $\hat{A}_k = \frac{r_k - \text{mean}(\mathbf{r})}{\text{orig}(\mathbf{r}) + \epsilon}$ // Group-relative advantage
  ▷ *4. Policy Update*
  Compute GRPO loss on C_{vq}: $\mathcal{L} = -\frac{1}{G}\sum_{k=1}^{G}\left[\hat{A}_k \log \pi_\theta(s_k|C_{\text{vq}}) - \beta \mathbb{D}_{\text{KL}}\right]$
  Update θ using gradient descent
end for

4.2 Training Strategy

SimpleOCR trains models exclusively on visual question format. All training samples undergo the $\mathcal{T}_{\text{render}}$ transformation which is no mixing of standard and visual question formats during training. This design eliminates text channel shortcuts entirely, forcing every training update to engage the visual text reading pathway.

Our approach is implemented purely as data preprocessing via $\mathcal{T}_{\text{render}}$, requiring no architectural changes and no modification to standard training objectives. For RL training, as illustrated in Alg. 2,

we follow the standard GRPO algorithm while conditioning generation on C_{vq}: we first construct x_{render} and C_{vq}, sample a group of G responses, compute rewards and group-relative advantages, and update the policy using the GRPO objective with the KL regularizer unchanged.

Critically, while training uses exclusively C_{vq}, evaluation employs standard format C_{orig}. This forces models to develop format-agnostic reasoning capabilities rather than format-specific patterns, learning to extract and process question content regardless of presentation modality.

4.3 Plug-and-Play Integration

Beyond standalone training, SimpleOCR integrates seamlessly into existing training frameworks. We demonstrate this with NoisyRollout (Liu et al., 2025b).

NoisyRollout employs a hybrid rollout strategy: for each sample, it generates n_1 rollouts from clean images $(\mathbf{x}_{\text{img}}, q_{\text{text}})$ and n_2 rollouts from perturbed images $(T_\alpha(\mathbf{x}_{\text{img}}), q_{\text{text}})$, where T_α applies image distortion with strength α. All rollouts contribute to computing group-relative advantages, improving policy exploration and visual robustness.

We integrate SimpleOCR by substituting the perturbation branch with visual question samples. Specifically, we generate n_1 rollouts from the standard context C_{orig} and n_2 rollouts from the visual question context C_{vq}. All rollouts contribute to group-relative advantage computation as in standard NoisyRollout. Policy updates remain conditioned on C_{orig} following NoisyRollout's original design. This integration requires no algorithmic modifications, as we simply substitute one augmentation strategy for another. The combination proves effective because the two methods target orthogonal objectives: NoisyRollout enhances visual robustness through image perturbations, while SimpleOCR specifically addresses OCR utilization through visual text reading.

5 Experiments

5.1 Experiment Settings

Dataset. We train on Geometry3K (Lu et al., 2021) (2.1K instances) and MMK12 (Meng et al., 2025) (6.4K instances), totaling 8.5K instances.

Evaluation. We evaluate on two dimensions: (1) *in-domain* performance on Geometry3K and MMK12 test sets, and (2) *out-of-distribution* generalization on MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024b), MathVista (Lu et al., 2023), and HallusionBench (Guan et al., 2024). We additionally evaluate on OCR-intensive benchmarks: InfographicVQA (Mathew et al., 2022) (InfoVQA) and ChartQA (Masry et al., 2022). All evaluations utilize greedy decoding, followed by a hybrid judging pipeline combining symbolic verification (Math-Verify[1]) and LLM-based assessment (GPT-4o (Hurst et al., 2024)). We detail the full protocol in Appendix E.

5.2 Main Results

Robust Transfer via Zero-Shot Generalization. SimpleOCR trains exclusively on VQ inputs but evaluates on standard inputs, creating a severe distributional shift that rigorously tests visual capability. Rather than suffering the expected degradation from format mismatch, SimpleOCR achieves robust zero-shot transfer. As shown in Table 1, it matches the baseline's in-domain performance (52.9% vs. 53.1%) while strictly outperforming it on out-of-distribution generalization (52.6% vs. 51.2%). This transfer is most potent on visually demanding tasks like MathVision, where we observe a 10.7% gain. These results prove that the model has not merely memorized the VQ format, but has internalized a fundamental visual text extraction capability that persists even when text shortcuts are restored.

Gains Correlate with Visual-Text Dependency. The performance improvements are structurally non-uniform. MathVision exhibits the most significant boost (24.9% vs. 22.5%), followed by MathVista (68.7% vs. 66.9%) and MathVerse (47.7% vs. 46.4%). Crucially, these benchmarks share a dependency on visual information density: they require extracting critical data or text embedded directly within figures. In contrast, performance slightly regresses on Geometry3K (43.4% vs. 44.3%), a benchmark governed more by abstract geometric logic than by visual text reading. This divergence confirms that SimpleOCR specifically sharpens the visual-text extraction pathway rather than offering a generic reasoning boost. We consider this a strategic trade-off: a marginal dip in pure geometry is exchanged for robust generalization on tasks where visual grounding is paramount.

Superiority on OCR-Intensive Benchmarks. Figure 3 (see Appendix Table 7) confirms that

[1]https://github.com/huggingface/Math-Verify

Method	Data Size	In-Domain			Out-of-Distribution				
		Geo3K	MMK12	Avg.	MathVerse	MathVision	MathVista	HallusionBench	Avg.
Open-source Baselines (SFT / General)									
InternVL-2.5-8B-Instruct* (Chen et al., 2024)	-	-	-	-	39.5	19.7	64.4	67.3	47.7
LLaVA-OneVision-7B* (Li et al., 2024a)	-	-	-	-	26.2	-	63.2	48.4	-
Kimi-VL-16B* (Team et al., 2025b)	-	-	-	-	44.9	21.4	**68.7**	66.2	50.3
Mulberry-7B* (Yao et al., 2024)	-	-	-	-	-	-	63.1	-	-
Math-LLaVA (Shi et al., 2024)	360K	-	-	-	22.9	15.7	46.6	-	-
RL-Optimized Models (R1-series)									
R1-VL-7B (Zhang et al., 2025a)	260K+10K	34.3	39.1	36.7	37.5	19.1	61.6	62.8	45.3
R1-OneVision-7B (Yang et al., 2025b)	155K+10K	37.4	47.2	42.3	43.6	20.9	63.1	65.6	48.3
ThinkLite-7B-VL (Wang et al., 2025b)	1.1K	38.8	56.8	47.8	46.7	24.2	66.9	66.1	51.0
VLAA-Thinker-7B (Chen et al., 2025)	25K	37.6	56.7	47.2	<u>46.9</u>	<u>24.4</u>	<u>67.6</u>	68.1	51.8
MM-Eureka-8B* (Meng et al., 2025)	15K	-	-	-	40.4	22.2	67.1	65.3	48.8
Our Methods									
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	-	37.6	53.6	45.6	43.9	23.4	64.2	68.2	49.9
+ GRPO	**8.5K**	**44.3**	<u>61.9</u>	**53.1**	46.4	22.5	66.9	<u>68.9</u>	<u>51.2</u>
+ SimpleOCR	**8.5K**	<u>43.4</u>	**62.3**	<u>52.9</u>	**47.7**	**24.9**	**68.7**	**69.1**	**52.6**

Table 1: Performance on mathematical reasoning and visual perception benchmarks. Models marked with "*" are cited from original papers. **Bold** and <u>underlined</u> numbers indicate the best and second-best performance, respectively. Data sizes for SFT and RL are respectively marked in blue and red.

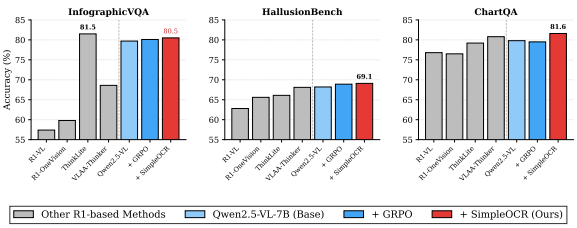


Figure 3: Performance on OCR-intensive benchmarks. SimpleOCR demonstrates superior performance, achieving 81.6% on ChartQA and 69.1% on HallusionBench.

SimpleOCR excels on tasks requiring explicit visual text recognition. On ChartQA, while standard GRPO slightly degrades performance (79.8% → 79.5%), SimpleOCR reverses this trend, reaching 81.6%. Consistent improvements are observed on InfographicVQA and HallusionBench, reaching 80.5% and 69.1%, respectively. This establishes a clear hierarchy of efficacy: gains are pronounced on OCR-centric tasks (e.g., ChartQA) and visually grounded math (e.g., MathVision), but negligible on pure geometry (e.g., Geometry3K). This distribution confirms that SimpleOCR functions as a targeted enhancer of visual-text utilization rather than a generic regularizer.

5.3 Analysis

Plug-and-Play Compatibility. Table 2 demonstrates the compatibility of SimpleOCR with advanced training strategies like NoisyRollout (Liu et al., 2025b). On Qwen2.5-VL-7B, SimpleOCR outperforms the GRPO baseline by 2.7%. This trend is consistent at the 3B scale: SimpleOCR delivers a 5.3% boost in average OOD accuracy, which is further amplified by the inclusion of Noisy-Rollout. The consistent gains confirm that the methods target distinct reasoning dimensions: SimpleOCR provides semantic grounding, while Noisy-Rollout improves perceptual robustness. This orthogonality validates SimpleOCR as a flexible plug-and-play augmentation compatible with existing training paradigms.

Consistency Across Model Scales. We further investigate scaling behavior in Table 2. On Qwen2.5-VL-7B, SimpleOCR delivers a robust 2.7% over the GRPO baseline (52.6% vs. 51.2%), validating its efficacy beyond small-scale models. While the gain margin naturally narrows compared to the 3B model (an expected consequence of *performance saturation* in larger models), the consistent positive trajectory confirms that "modality laziness" is a fundamental architectural tendency irrespective of capacity. SimpleOCR effectively mitigates this tendency regardless of scale, serving as a scalable corrective mechanism.

5.4 Ablation Study

Optimization Conflict in Mixed Strategies. To better understand the interaction between standard inputs and VQ training, we evaluated a mixed strategy (Partial Exposure). Figure 4 reveals a distinct U-shaped performance trajectory. On average across four representative OOD benchmarks (detailed in Appendix Table 6), the mixed setting (50% VQ) unexpectedly dips below the baseline

Table 2: **Analysis of Integration & Scalability:** SimpleOCR integrates seamlessly with HybridRollout across model scales. The combination yields consistent gains, particularly on the 3B model, validating that SimpleOCR (focused on text reading) and HybridRollout (focused on visual robustness) are orthogonal and complementary.

	In-Domain			Out-of-Distribution				
Method Configuration	Geo3K	MMK12	Avg.	MathVerse	MathVision	MathVista	Hallusion	Avg.
Qwen2.5-VL-3B-Instruct	26.0	45.9	36.0	32.5	18.2	50.8	59.1	40.2
+ GRPO (baseline) ($n = 6$)	35.1	51.2	43.2	36.6	18.5	53.0	58.6	41.7
+ SimpleOCR ($n = 6$)	**36.1**	**53.6**	**44.9**	40.4	20.1	53.3	**61.7**	43.9
+ HybridRollout ($n_1 = 3, n_2 = 3$)	34.8	**53.6**	44.2	**41.4**	**20.6**	**57.3**	58.7	**44.5**
Qwen2.5-VL-7B-Instruct	37.6	53.6	45.6	43.9	23.4	64.2	68.2	49.9
+ GRPO (baseline) ($n = 6$)	**44.3**	61.9	**53.1**	46.4	22.5	66.9	68.9	51.2
+ SimpleOCR ($n = 6$)	43.4	62.3	52.9	**47.7**	**24.9**	**68.7**	**69.1**	**52.6**
+ HybridRollout ($n_1 = 3, n_2 = 3$)	41.1	**65.0**	**53.1**	47.6	**24.9**	**68.7**	68.0	52.3

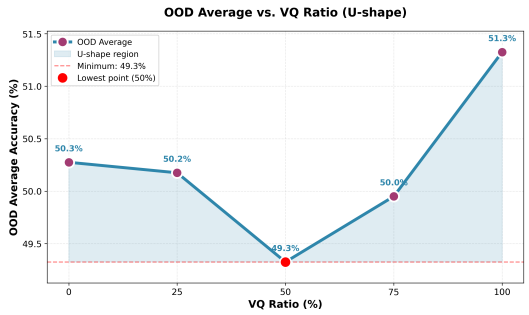


Figure 4: The "U-Shaped" Optimization Conflict. We report the average performance across four representative OOD benchmarks. The mixed strategy (50% VQ) results in a net performance loss, illustrating that contradictory modality signals hinder generalization.

Table 3: Ablation on rendering style. Randomization prevents overfitting to specific visual patterns. (Note: "Random style" corresponds to the full SimpleOCR method used in main results.)

	In-Domain		Out-of-Distribution			
Rendering Strategy	Geo3K	MMK12	M-Verse	M-Vision	M-Vista	WeMath
Fixed style	41.4	61.3	46.9	23.4	65.9	61.6
Random style	**43.4**	**62.3**	**47.7**	**24.9**	**68.7**	**64.0**

Table 4: Impact of Group Sampling Size n. We analyze the effect of the number of generations per prompt during GRPO training.

	In-Domain			Out-of-Distribution					
Configuration	**Geo3K**	**MMK12**	**Avg.**	**M-Verse**	**M-Vision**	**M-Vista**	**WeMath**	**Hallusion**	**Avg.**
$n = 3$	40.4	**62.5**	51.5	46.2	24.0	67.5	60.5	**70.4**	53.7
$n = 6$	**43.4**	62.3	**52.9**	**47.7**	**24.9**	**68.7**	**64.0**	69.1	**54.9**
$n = 9$	41.4	63.0	52.2	47.4	24.6	66.4	61.6	67.9	53.6

(49.3% vs. 50.3%), creating a generalization valley. This degradation is particularly pronounced on reasoning-heavy tasks like WeMath (-4.4%) and MathVista (-2.8%).

We attribute this to a fundamental optimization conflict. When exposed to mixed formats, the model receives contradictory learning signals: standard inputs encourage reliance on the text encoder (the path of least resistance), while VQ inputs demand active visual engagement. Rather than converging on a robust joint strategy, the model oscillates between these modalities, failing to master either. The SimpleOCR (100% VQ) setting resolves this by enforcing a structural constraint. By completely blocking text-based shortcuts, the model is compelled to optimize the visual extraction pathway. Paradoxically, this "forced commitment" yields representations that are modality-agnostic, enabling superior zero-shot transfer (51.3% average accuracy).

Robustness via Randomization. Table 3 validates the efficacy of our randomized rendering strategy. Compared to a static rendering style (e.g., fixed font and color), applying stochastic styles (varying font, size, and color) yields consistent gains, most notably a 2.8% improvement on MathVista and 2.4% on WeMath. We attribute the limitations of the fixed setting to *feature overfitting*. When text always appears with a deterministic visual style, the model tends to memorize low-level texture cues (e.g., specific font patterns) rather than performing generalizable OCR. Randomization disrupts these shortcuts. By diversifying the stylistic presentation, we compel the model to actively decode text regardless of its visual variations. This strategy effectively prevents the model from relying on nuisance variables (such as font type or color), ensuring that the learned grounding capability is genuinely robust.

Sensitivity to Group Sampling Size. We investigate the impact of the group size n (the number of rollouts generated per prompt) on SimpleOCR training dynamics in Table 4, employing the 7B

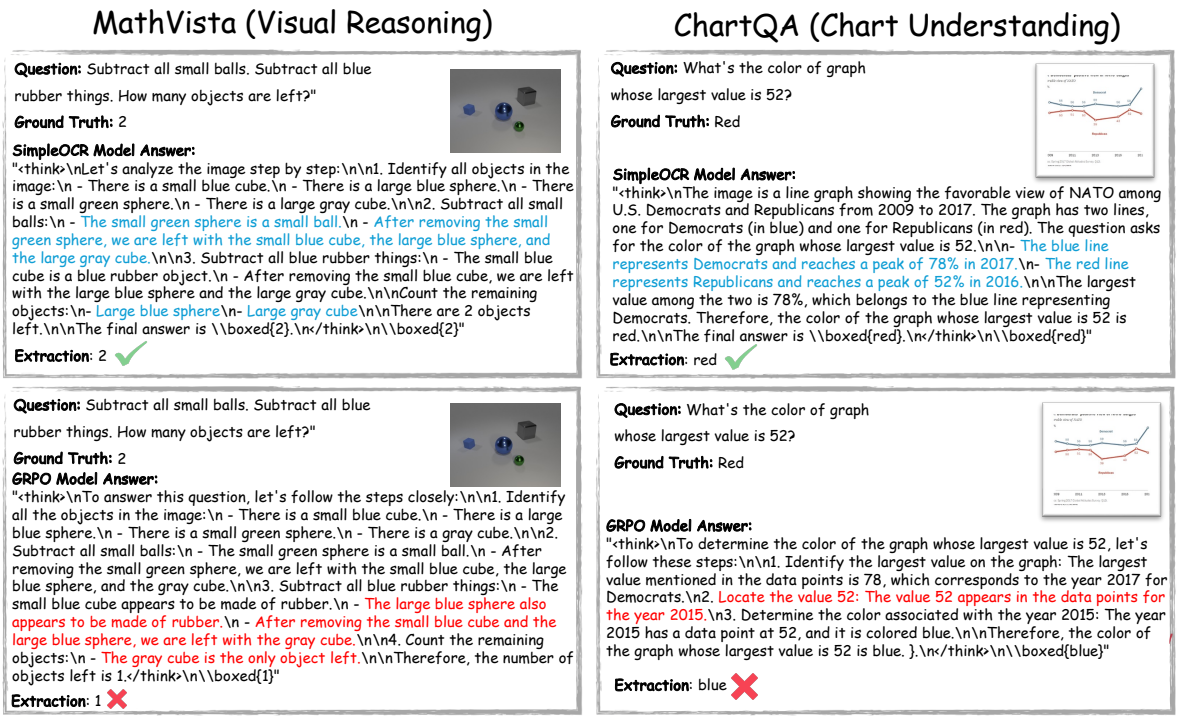


Figure 5: Left: On MathVista, the GRPO baseline is misled by hallucinated semantic priors, while SimpleOCR correctly identifies material properties. Right: On ChartQA, the baseline relies on superficial keyword spotting, whereas SimpleOCR performs holistic visual analysis. Blue: correct grounding; red: heuristic errors.

model as the backbone. Standard RL scaling laws typically suggest that larger group sizes improve gradient estimation. However, our results reveal an inverted U-shaped trend. Increasing the group size from $n = 3$ to $n = 6$ yields a robust 2.2% gain in average OOD performance, confirming that sufficient exploration is critical for learning complex visual grounding. Crucially, further scaling to $n = 9$ does not yield additional benefits; instead, performance suffers a slight 2.4% regression. We hypothesize that in the context of VQ training, excessively large groups may introduce "reward hacking" on noisy visual samples or optimization instability. Consequently, we adopt $n = 6$ as the optimal trade-off between computational efficiency and reasoning performance.

5.5 Qualitative Analysis

Figure 5 illustrates the behavioral shift. In visual reasoning (MathVista), the baseline GRPO model succumbs to semantic priming, associating the text "blue" with a prominent sphere despite conflicting visual evidence (metallic luster), whereas SimpleOCR discriminates texture correctly. Similarly, on ChartQA, the baseline relies on superficial keyword spotting, matching "52" without comprehending the structural condition "largest value", while SimpleOCR successfully parses the chart topology. These cases validate that the capability-utilization gap is not a deficit of perception but of execution preference. Standard models default to spurious text shortcuts, but SimpleOCR structurally blocks this path, compelling the model to engage in grounded visual reasoning.

6 Conclusion

In this paper, we identified and quantified the "modality laziness" in MLLMs, where models bypass visual evidence in favor of text-based shortcuts. Our diagnostic VQ setting revealed a significant capability-utilization gap, which we addressed through SimpleOCR. By structurally enforcing visual engagement via randomized text rendering, SimpleOCR effectively transforms the model's reliance from parametric priors to grounded visual perception. Empirically, SimpleOCR delivers consistent improvements across both in-domain and out-of-distribution benchmarks. Notably, it achieves these gains with extreme data efficiency (using 30× less data than comparable RL methods) and seamless plug-and-play compatibility with existing frameworks.

Acknowledgments

This work was partially supported by the Amazon Research Award, the Cisco Faculty Research Award.

Limitations

While SimpleOCR effectively bridges the capability-utilization gap, we identify two primary limitations. First, our method operates as an *elicitation strategy* rather than a fundamental capability builder. It relies on the base MLLM having latent OCR capabilities (i.e., a strong vision encoder) to recognize the rendered text. Second, our approach is bounded by *visual resolution constraints* when handling extremely long queries. Unlike text encoders that scale efficiently to long contexts, rendering extensive text prompts (e.g., multi-paragraph instructions) onto a single image is limited by the vision encoder's input resolution.
Potential Risks. Enhanced visual text extraction could theoretically be leveraged to bypass visual security measures (e.g., CAPTCHA solvers) or to automate the extraction of sensitive personal information from natural images (e.g., reading documents or screens in the background of photos). However, our method functions as an activation strategy for existing base models rather than introducing new, specialized attack capabilities. The risks are inherently bound by the safety alignment and capabilities of the underlying foundation models.

References

Jinze Bai, Shuai Bai, Shusheng Yang, and 1 others. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. 2025. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. pages 14375–14385.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

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*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. 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.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*.

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

Jiaqi Liu, Kaiwen Xiong, Peng Xia, Yiyang Zhou, Haonian Ji, Lu Feng, Siwei Han, Mingyu Ding, and Huaxiu Yao. 2025a. Agent0-vl: Exploring self-evolving agent for tool-integrated vision-language reasoning. *arXiv preprint arXiv:2511.19900*.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025b. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Pan Lu, Hritik Bansal, Tony Xia, and 1 others. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, pages 6774–6786.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, and 1 others. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ACL*.

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. *WACV*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Runqi Qiao, Qiuna Tan, Guanting Dong, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. 2025. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680.

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

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, and 1 others. 2025a. Hunyuanocr technical report. *arXiv preprint arXiv:2511.19575*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025b. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Aaron Tu, Weihao Xuan, Heli Qi, Xu Huang, Qingcheng Zeng, Shayan Talaei, Yijia Xiao, Peng Xia, Xiangru Tang, Yuchen Zhuang, and 1 others. 2025. Position: The hidden costs and measurement gaps of reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2509.21882*.

Alex Jinpeng Wang, Linjie Li, Yiqi Lin, Min Li, Lijuan Wang, and Mike Zheng Shou. 2024a. Leveraging visual tokens for extended text contexts in multi-modal

learning. *Advances in Neural Information Processing Systems*, 37:14325–14348.

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

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

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

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

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, and 1 others. 2026. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, and 1 others. 2025a. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. *arXiv preprint arXiv:2506.00555*.

Peng Xia, Kaide Zeng, Jiaqi Liu, Can Qin, Fang Wu, Yiyang Zhou, Caiming Xiong, and Huaxiu Yao. 2025b. Agent0: Unleashing self-evolving agents from zero data via tool-integrated reasoning. *arXiv preprint arXiv:2511.16043*.

Xinyu Yang, Junlin Han, Rishi Bommasani, Jinqi Luo, Wenjie Qu, Wangchunshu Zhou, Adel Bibi, Xiyao Wang, Jaehong Yoon, Elias Stengel-Eskin, and 1 others. 2025a. Reliable and responsible foundation models. *Transactions on Machine Learning Research*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Guanyu Yao, Qiucheng Wu, Yang Zhang, Zhaowen Wang, Handong Zhao, and Shiyu Chang. 2025. Rethinking the text-vision reasoning imbalance in mllms through the lens of training recipes. *arXiv preprint arXiv:2510.22836*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, and 1 others. 2025a. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025b. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

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

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?

Yifan Zhang, Liang Hu, Haofeng Sun, Peiyu Wang, Yichen Wei, Shukang Yin, Jiangbo Pei, Wei Shen, Peng Xia, Yi Peng, and 1 others. 2025b. Skywork-r1v4: Toward agentic multimodal intelligence through interleaved thinking with images and deepresearch. *arXiv preprint arXiv:2512.02395*.

A Dataset Details

A.1 Training Data

Our training set consists of two high-quality mathematical reasoning datasets, totaling 8.5K instances. Detailed statistics are provided in Table 5.

Table 5: **Training Data Statistics.** We combine geometry-focused and general K-12 math datasets to construct a diverse training corpus.

Dataset	Source	Domain	Size
Geometry3K	(Lu et al., 2021)	Plane Geometry	2,100
MMK12	(Meng et al., 2025)	K-12 Mathematics	6,400
Total	-	**Mixed**	**8,500**

Geometry3K (Lu et al., 2021). A high-quality geometry problem-solving dataset containing formal geometric diagrams and corresponding problem descriptions. We utilize the training split (2.1K samples) to enhance the model's spatial reasoning and geometric calculation capabilities.
MMK12 (Meng et al., 2025). A comprehensive multimodal dataset derived from K-12 mathematics curriculum. It covers a wide range of topics including algebra, arithmetic, and function analysis. The subset used (6.4K samples) provides diverse visual-text reasoning scenarios essential for general mathematical grounding.

A.2 Evaluation Benchmarks

To rigorously assess generalization capabilities, we evaluate on five mathematical reasoning benchmarks and two OCR-intensive tasks.

Mathematical Reasoning.

- **MathVista** (Lu et al., 2023). A comprehensive benchmark integrating diverse mathematical reasoning tasks. It serves as a primary gauge for general multimodal mathematical capability.
- **MathVision** (Wang et al., 2024b). A large-scale benchmark designed to evaluate MLLMs across diverse mathematical domains and complex visual contexts.
- **MathVerse** (Zhang et al., 2024). A dataset specifically curated to diagnose whether MLLMs truly interpret visual diagrams or rely on text shortcuts. This aligns perfectly with our study's motivation to detect "modality laziness".
- **WeMath** (Qiao et al., 2024). A benchmark focusing on human-like reasoning processes in complex mathematical problems, testing the depth of the model's logical derivation.
- **HallusionBench** (Guan et al., 2024). An advanced diagnostic suite for detecting visual hallucinations and illusions. We use it to verify faithful visual grounding and resistance to perceptual interference.

OCR-Intensive Tasks. To verify the transfer of visual text reading skills, we include two specific benchmarks:

- **ChartQA** (Masry et al., 2022). A dataset requiring reasoning over charts with data labels, titles, and legends, serving as a direct test of the model's ability to extract and integrate fine-grained visual text.
- **InfographicVQA** (Mathew et al., 2022). A benchmark challenging models to understand complex document layouts and infographics with high-density text.

B System Prompts

We utilize the standard system prompt from the `verl` framework to elicit structured reasoning (Chain-of-Thought) and formatted answers.

System Prompt for Reasoning

"Solve the question. The user asks a question, and you solve it. You first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process MUST BE enclosed within `<think> </think>` *tags. The answer is in latex format and wrapped in* `$...$`*. The final answer must be wrapped using the* `\boxed{}` *command."*

B.1 Evaluated Models

We include a comprehensive set of state-of-the-art multimodal models in our evaluation, categorized into general-purpose open-source baselines and recent RL-optimized models.

Open-Source Baselines.

- **Qwen2.5-VL-3/7B-Instruct** (Bai et al., 2025). The latest iteration of the Qwen-VL series, featuring state-of-the-art OCR and visual understanding capabilities trained on massive-scale datasets. We utilize these as our primary base models to demonstrate the effectiveness of SimpleOCR.

- **InternVL-2.5-8B-Instruct** (Chen et al., 2024). A powerful MLLM that expands performance boundaries through model and test-time scaling, known for its strong general-purpose visual perception.
- **LLaVA-OneVision-7B** (Li et al., 2024a). A model designed for easy visual task transfer, utilizing a unified architecture to handle diverse vision-language scenarios efficiently.
- **Kimi-VL-16B** (Team et al., 2025b). A large-scale open-weights model utilizing a Mixture-of-Experts (MoE) architecture, demonstrating competitive performance on chart and document understanding benchmarks.
- **Mulberry-7B** (Yao et al., 2024). An MLLM empowered with OpenAI-o1-like reasoning capabilities via collective Monte Carlo Tree Search (MCTS), focusing on enhanced logical deduction.

RL-Optimized & R1-Series Models.

- **Math-LLaVA.** (Shi et al., 2024) A specialized model bootstrapped for mathematical reasoning, serving as a strong baseline for SFT-based mathematical capability.
- **R1-VL-7B.** (Zhang et al., 2025a) A pioneering model trained via step-wise Group Relative Policy Optimization (GRPO), explicitly rewarding intermediate reasoning steps to improve logical consistency.
- **R1-OneVision-7B.** (Yang et al., 2025b) An extension of the R1 series that advances generalized multimodal reasoning through cross-modal formalization techniques.
- **ThinkLite-7B-VL.** (Wang et al., 2025b) A data-efficient model achieving state-of-the-art performance with fewer samples, utilizing MCTS-guided sample selection for self-improvement.
- **VLAA-Thinker-7B.** (Chen et al., 2025) A model investigating the trade-offs between SFT and RL in R1-like reasoning, providing insights into training recipes for reasoning-heavy MLLMs.
- **MM-Eureka-8B.** (Meng et al., 2025) A model exploring the frontiers of multimodal reasoning using rule-based reinforcement learning, emphasizing verified feedback signals.

Table 6: Impact of VQ Training Ratio (Detailed Breakdown). We report the performance on four reasoning-heavy OOD benchmarks. The mixed strategy (50% VQ) consistently underperforms or stagnates compared to the baseline (Avg. 49.3 vs 50.3), supporting the hypothesis of optimization conflict. Only the full VQ strategy (SimpleOCR) achieves robust generalization gains (Avg. 51.3).

VQ Ratio	MathVerse	MathVision	MathVista	WeMath	Avg.
Standard (0% VQ)	46.4	22.5	66.9	65.3	50.3
Mixed (25% VQ)	47.8	22.9	67.2	62.8	50.2
Mixed (50% VQ)	46.2	23.7	65.0	62.4	49.3
Mixed (75% VQ)	48.0	23.9	65.7	62.2	50.0
SimpleOCR (100% VQ)	**47.7**	**24.9**	**68.7**	64.0	**51.3**

C Detailed Ablation Results

In Section 5.4, we discussed the optimization conflict observed in mixed training strategies. Table 6 provides the detailed performance breakdown across four representative out-of-distribution benchmarks.

As shown, the mixed strategy (50% VQ) fails to improve over the baseline in most reasoning-intensive tasks (e.g., WeMath, MathVista), confirming that the conflicting modality signals hinder model convergence. In contrast, the pure SimpleOCR strategy (100% VQ) achieves the best average performance across the board.

D OCR-Intensive Benchmarks

We provide the exact numerical breakdown for OCR-intensive tasks in Table 7. A key observation is that standard GRPO can lead to negative transfer on fine-grained visual tasks like ChartQA (dropping from 79.8% to 79.5%), likely due to the model overfitting to textual reasoning shortcuts. In contrast, SimpleOCR consistently yields improvements across all metrics (81.6% on ChartQA), confirming its effectiveness in preserving and enhancing visual grounding capabilities without compromising general reasoning.

Table 7: **Performance on OCR-Intensive Benchmarks.** Exact numbers corresponding to Figure 3.

Method	ChartQA	HallusionBench	InfoVQA
Base Model	79.8	68.2	79.7
GRPO (Original images)	79.5	68.9	80.1
GRPO + SimpleOCR	**81.6**	**69.1**	**80.5**

E Evaluation Protocol Details

Inference and Extraction. For all experiments, we perform inference using greedy decoding (`temperature=0`) to ensure reproducibility. To isolate the final answer from the Chain-of-Thought (CoT) rationale, we employ a rule-based extraction parser. Specifically, we extract the content within the last occurrence of the `\boxed{...}` delimiter in the model output. If no such delimiter is found, the raw output is passed to the subsequent evaluation stages.

Hierarchical Judging Pipeline. We implement a two-stage cascaded evaluation strategy to balance strict symbolic correctness with semantic flexibility:

1. **Stage 1: Symbolic Verification (Math-Verify).** We first employ the `math-verify` library for symbolic equivalence checks. This tool parses mathematical expressions into canonical forms (e.g., standardizing fractions, square roots, and units) to determine correctness. If `math-verify` returns a positive match, the sample is marked as correct immediately.

2. **Stage 2: LLM-based Fallback Judge.** For samples where symbolic verification fails or is inconclusive (e.g., complex textual reasoning or format mismatches), we employ `gpt-4o-2024-08-06` as a fallback evaluator. We construct a meta-evaluation prompt containing the question, the ground truth, and the student's answer. The LLM is strictly instructed to:

   - Ignore superficial formatting differences (e.g., Markdown styling).
   - Check for mathematical equivalence rather than string matching.
   - Allow a relative numerical tolerance of $\pm 1\%$ (unless specified otherwise).
   - For multiple-choice questions, verify that the selected option letter matches the ground truth.

   The LLM outputs a binary score (0 or 1) based on these criteria.

Benchmark-Specific Protocols.

- **HallusionBench:** We strictly adhere to the official evaluation protocol, utilizing its dataset-specific LLM judge to handle the unique "uncertain" label requirements.
- **Geometry3K:** Due to the strict formatting of this dataset, we rely primarily on symbolic verification, enforcing exact matches for geometric values and units.

F Supplementary Implementation Details

We provide the detailed hyperparameter configurations used in our experiments in Table 8.

Table 8: **Summary of hyperparameter configurations.**

Parameter	Configuration
Model Base	Qwen2.5-VL-Instruct
Vision Encoder	Frozen
Global Batch Size	128
Rollout Batch Size	512
Rollout Temperature	1.0
Learning Rate	1×10^{-6}
Optimizer	AdamW
Total Training Steps	200
CPU Memory	512GB
GPU	RTX 6000 Pro Blackwell