Thinking with Deltas: Incentivizing Reinforcement Learning via Differential Visual Reasoning Policy

Shujian Gao[1,2] **Yuan Wang**[3] **Jiangtao Yan**[4] **Zuxuan Wu**[1,2*] **Yu-Gang Jiang**[1*]
[1]Institute of Trustworthy Embodied AI, Fudan University [2]Shanghai Innovation Institute
[3]Zhejiang University [4]Wuhan University
{sjgao26,zxwu,ygj}@fudan.edu.cn yuanwang@zju.edu.cn jtyan@whu.edu.cn

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has significantly advanced reasoning capabilities in Large Language Models. However, adapting RLVR to multimodal domains suffers from a critical *perception-reasoning decoupling*. Existing paradigms, driven by text-centric outcome rewards, reasoning in language medium, inadvertently encourage models to bypass visual perception. We empirically validate this issue on representative RLVR policies through blind experiments: **the evaluated policies exhibit only small drops, and in some cases even improve, when visual inputs are removed or replaced.** This reveals that these models degenerate into *blind reasoners*, exploiting linguistic priors to generate plausible answers instead of attending to visual evidence. In response, we propose **Thinking with Deltas**, a framework driven by a **Differential Visual Reasoning Policy (DVRP)**. DVRP introduces intrinsic supervision via visual triplets, comprising original, masked, and perturbed inputs. It optimizes the model to maximize reasoning divergence from masked inputs (enforcing *visual sensitivity*) while minimizing divergence from perturbed inputs (ensuring *visual robustness*). By aligning reasoning variations strictly with the *Delta* of visual information, DVRP inherently bolsters visual understanding capabilities and significantly outperforms state-of-the-art methods on both general and medical benchmarks, without requiring external annotations or auxiliary tools.

1 Introduction

Recent advancements in Large Language Models (LLMs) (Chang et al., 2024; Brown et al., 2020) have been substantially driven by Chain-of-Thought (CoT) prompting and the curation of high-quality reasoning data (Kojima et al., 2022; Ye et al., 2025). Building on this foundation, Reinforcement Learning with Verifiable Rewards (RLVR) (Ouyang et al., 2022; Shao et al., 2024) has emerged as a promising post-training paradigm. By incentivizing self-correction and scaling test-time compute, RLVR enables models to generate rigorous and self-consistent solution paths, establishing a new standard for complex problem-solving in textual domains (Wang et al., 2025e).

However, a fundamental misalignment arises when directly applying the RLVR paradigm to multimodal domains (Liu et al., 2025a). Predominant approaches typically mirror text-centric methodologies (DeepSeek-AI et al., 2025) and rely heavily on textual outcome rewards such as accuracy or format constraints (Wang et al., 2025f). Furthermore, attempts to incorporate process supervision via external judges often inherit the biases and limitations of the judge models themselves (Luo et al., 2025). Conceptually, if we analogize a Multimodal LLM (MLLM) to a human observer, the visual encoder functions as the eyes. Optimizing rewards based solely on the output text sequence is equivalent to evaluating speech without verifying vision. This approach neither ensures the effective perception of visual signals nor stimulates deep visual understanding (Han et al., 2025b). Consequently, current frameworks neglect the visual dependency between visual inputs and reasoning outcomes.

As illustrated in Figure 1, we empirically validate this decoupling through **blind experiments**, where visual inputs are either removed (Text Only) or replaced with blank images (Black/White). We examine DAPO and GRPO under varying input configurations. For the GRPO baseline (Shao et al., 2024), the performance drop is negligible, despite the absence of essential visual context. More strikingly, DAPO (Yu et al., 2025) exhibits an *unexpected performance gain*: removing visual information *improves* accuracy (e.g., **+3.5%** on MathVista). These findings indicate latent *reward hacking*, where the policy degenerates into a *blind*

* Corresponding authors.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 986–1015
November 6-10, 2026 ©2026 Association for Computational Linguistics

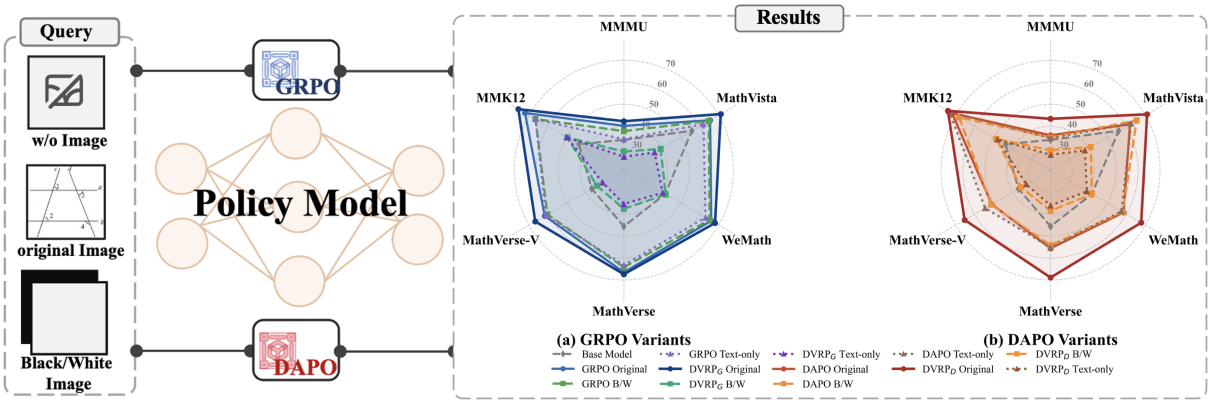


Figure 1: **Empirical Validation of Visual Decoupling.** Performance comparison in blind settings (Text Only or Blank Image) reveals latent reward hacking. The negligible performance drop in GRPO and the *unexpected performance gain* in DAPO (where removing visual inputs actually improves accuracy) indicate that policies degenerate into *blind reasoners* relying on linguistic shortcuts rather than visual evidence.

reasoner that exploits linguistic shortcuts rather than visual evidence. In contrast, our DVRP variants exhibit substantial degradation under the same blind settings, suggesting stronger behavioral dependence on visual inputs and reduced reliance on language-only shortcuts. Detailed analysis is provided in Section F.

To address this perception-reasoning decoupling without resorting to complex data construction (Liu et al., 2024a) or external reward engineering (Zhang et al., 2025d), we propose **Thinking with Deltas**, a novel framework centered around a **Differential Visual Reasoning Policy (DVRP)**. Rather than relying only on proxy textual rewards, DVRP introduces intrinsic visual supervision through a visual triplet input stream: (1) *Invariant* (the original image), (2) *Decremental* (masked visual input), and (3) *Incremental* (perturbed visual input). Our core insight is that visually dependent reasoning should respond to meaningful reductions in visual information while remaining stable under non-semantic perturbations. Specifically, DVRP encourages the policy to diverge between the Invariant and Decremental states, providing a training signal for visual sensitivity, while aligning the Invariant and Incremental states to promote visual robustness. Optimizing these differential signals encourages the model to make greater use of visual evidence without requiring expensive external tools or dense annotations.

Our contributions are summarized as follows:

- We identify the perception-reasoning decoupling in current multimodal RL as a primary bottleneck, arguing that standard objectives optimize for linguistic plausibility rather than visual grounding.
- We introduce the **DVRP**, an algorithm that leverages visual triplets to construct intrinsic *Delta supervision signals* for both visual sensitivity and robustness, encouraging stronger visual perception and reasoning capabilities.
- Extensive experiments on two backbones spanning domains from **mathematical reasoning** to **specialized medical diagnostics** demonstrate that DVRP significantly outperforms state-of-the-art RLVR methods like GRPO and DAPO, effectively enhancing visual perception capabilities.

2 Related Work

2.1 Multimodal Reasoning

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly extended the reasoning capabilities of LLMs to visual domains (Wang et al., 2024a). Pioneering works such as LLaVA (Liu et al., 2024a) and Qwen-VL (Bai et al., 2023) demonstrated that visual instruction tuning could effectively align visual encoders with LLMs, enabling strong performance on general visual-oriented benchmarks (Yue et al., 2024b). Following this, the community has focused on enhancing the reasoning depth of these models. Inspired by the efficacy of CoT in LLMs, the integration of Multimodal-CoT (Zhang et al., 2024c; Sarch et al., 2025) and the curation of high-quality reasoning datasets (Chen et al., 2024b, 2023; Dong et al., 2025; Sun et al.) have emerged as piv-

otal strategies for bolstering visual understanding, thereby significantly mitigating linguistic hallucinations (Huang et al., 2024; Leng et al., 2024).

To substantially enhance visual understanding, a novel paradigm, which termed *visual reasoning in action* has emerged, advocating for thinking with images by explicitly incorporating visual content within CoT processes (Su et al., 2025). Representative strategies employ visual programming via code generation (Surís et al., 2023; Gupta and Kembhavi, 2023; Lin et al., 2025), leverage external visual utilities (e.g., crop, zoom, rotate) (Zheng et al., 2025b; Zhang et al., 2025a; Qi et al., 2024; Liu et al., 2023), or orchestrate diverse expert models to tackle multimodal tasks (Shen et al., 2023; Lu et al., 2023). Further explorations extend to tool-integrated, multi-turn, and multi-agent system for interleaved reasoning (Wu et al., 2025; Chen et al., 2024d; Wang et al., 2025b; Man et al., 2025) and unified frameworks (Han et al., 2025a; Li et al., 2025a). *However, these methods rely heavily on external sources such as expensive trajectory data, agentic tools, or advanced teacher models. This incurs significant overhead and does not fundamentally improve the model's intrinsic visual perception and reasoning capabilities.*

2.2 Advancements in RLVR Frameworks

Recent research on RLVR focuses on optimizing three critical components: **Data**, **Reward**, and **Rollout**.

Data Curation. High-quality Chain-of-Thought initialization is a prerequisite for stable RL training (Chen et al., 2025b; Huang et al., 2025a; Zhang et al., 2025c; Liang et al., 2025). Complementing this, recent studies employ dynamic data selection strategies, such as filtering trajectories based on value estimation (Wang et al., 2025d) or applying perplexity- and difficulty-based correction mechanisms (Kong et al., 2024; Jiang et al., 2025) to enhance sample efficiency and diversity.

Reward Engineering. Granular reward design serves as a pivotal mechanism in RLVR (DeepSeek-AI et al., 2025). Dominant approaches rely on *outcome-based* signals, including rigorous accuracy verification (Liu et al., 2025b), strict formatting and length constraints (Parthasarathi et al., 2025; Zhang and Zuo, 2025), caption alignment metrics (Gou et al., 2025b) for visual-semantic understanding, and grounding constraints (Zhang et al., 2025d). To further mitigate reasoning shortcuts, recent methods augment these with *process-level constraints* (Jiang et al., 2025), aiming to bolster the procedural correctness and reliability of intermediate reasoning chains.

Rollout Optimization. Enhancing the efficiency of the exploration phase is another active frontier. Recent strategies focus on optimizing the rollout space by dynamically adjusting sampling temperatures (Liao et al., 2025), expanding the sampling space (Liu et al., 2025a; Li et al., 2025b), polishing reasoning processes (Fan et al., 2025), and employing tree-search algorithms to diversify generation paths (Ji et al., 2025). These methods aim to balance exploration and exploitation, ensuring the policy covers a broader solution space without diverging into incoherence.

However, these strategies relegate the visual modality to an auxiliary role relative to text, neglecting the critical need to enhance visual robustness and sensitivity in MLLMs. In contrast, our proposed DVRP intrinsically fosters the model's visual perception capabilities.

3 Method

3.1 Problem Formulation

Preliminaries on GRPO. We consider a multimodal reasoning task where a policy model π_θ, parameterized by θ, takes a visual input I and a textual query q to generate a reasoning chain followed by a final answer, denoted as o. The training dataset is represented as $\mathcal{D} = \{(I, q, a)\}$, where a is the ground truth answer.

Recent advancements in RLVR have largely adopted GRPO (Shao et al., 2024) to enhance reasoning capabilities without the need for a separate value network. Formally, for each input instance (I, q), GRPO samples a group of G outputs $\{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{old}}$. A rule-based reward function $r(o, a)$ evaluates the correctness of each output (e.g., format compliance and answer accuracy), assigning a reward R_i. To reduce variance, the advantage $\hat{A}_i$ for the i-th output is computed by normalizing the rewards within the group:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon}, \tag{1}$$

where ϵ is a small constant for numerical stability. The optimization objective of GRPO is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}\sim\pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \mathcal{L}_{\text{clip}}(o_i, \hat{A}_i) \right], \tag{2}$$

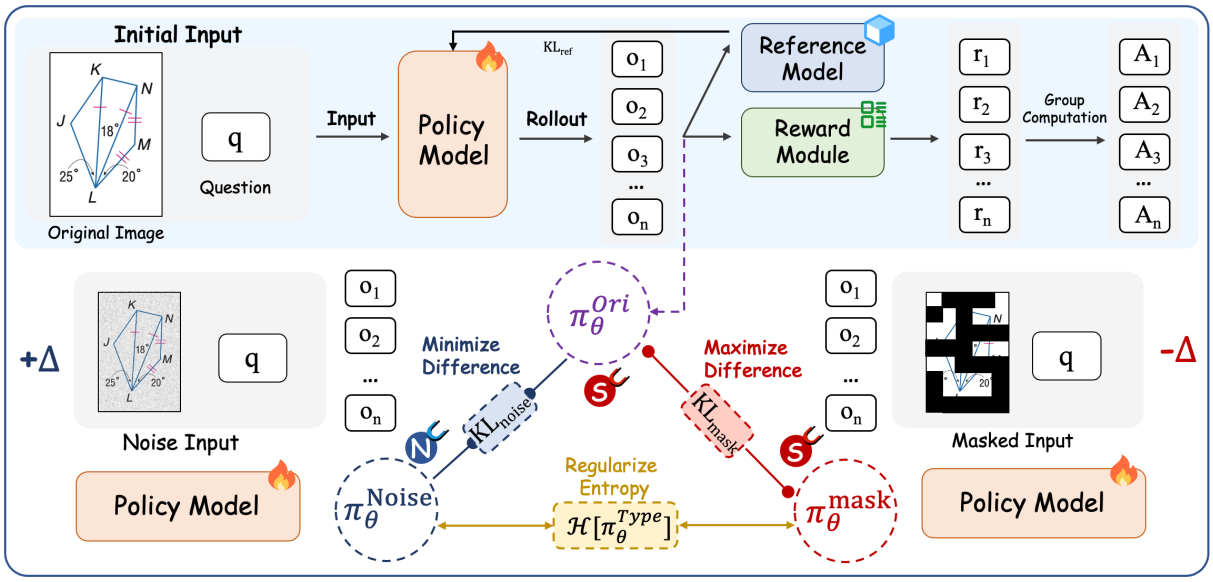


Figure 2: The framework of DVRP. Our method mitigates perception-reasoning decoupling via a visual triplet objective. The upper stream represents the standard reasoning rollout. The lower streams encourage two complementary behaviors: (1) Visual Robustness: minimizing the KL divergence between predictions on original and noise-perturbed inputs; and (2) Visual Sensitivity: maximizing the KL divergence when visual information is partially removed by masking. An entropy regularization term prevents distribution collapse.

where $\mathcal{L}_{\text{clip}}$ denotes the standard PPO clipped surrogate loss:

$$\mathcal{L}_{\text{clip}} = \min\big(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\epsilon_{clip}, 1+\epsilon_{clip})\hat{A}_i\big), \tag{3}$$

with $\rho_i = \frac{\pi_\theta(o_i|I,q)}{\pi_{\theta_{old}}(o_i|I,q)}$ representing the probability ratio. Following recent practices (Yu et al., 2025), we remove the KL divergence penalty term to encourage sufficient exploration.

The Perception-Reasoning Decoupling. A critical limitation of current multimodal RLVR paradigms is their failure to enforce the *visual dependency* between visual perception and reasoning paths (Jiang et al., 2025). By relying solely on outcome-based textual rewards and reasoning in language medium, these methods treat visual inputs as optional context rather than necessary evidence. Consequently, existing frameworks largely overlook the model's **sensitivity to critical visual semantics** and **robustness against visual perturbations**, inadvertently incentivizing linguistic shortcuts over visually grounded reasoning (Han et al., 2025b).

3.2 Visual Triplets as Intrinsic Supervision

To bridge this gap, we propose **DVRP**, which introduces intrinsic supervision via a *Visual Triplet* mechanism. As illustrated in Figure 2 , for each training instance, we construct three distinct views:

- **Invariant View (I):** The original visual input, serving as the anchor for standard reasoning. The policy distribution under this view is denoted as $\pi_\theta^{Ori} = \pi_\theta(\cdot|I,q)$.
- **Decremental View (I_{mask}):** A "$-\Delta$" state generated by masking visual regions randomly. This creates a counterfactual scenario where visual evidence is lost. The corresponding policy is $\pi_\theta^{Mask} = \pi_\theta(\cdot|I_{mask},q)$.
- **Incremental View (I_{noise}):** A "$+\Delta$" state generated by injecting non-semantic Diffusion noise. This simulates environmental instability while preserving semantics. The corresponding policy is $\pi_\theta^{Noise} = \pi_\theta(\cdot|I_{noise},q)$.

This triplet structure allows us to operationalize two complementary learning signals: *Visual Sensitivity* (diverging from π_θ^{Mask}) and *Visual Robustness* (aligning with π_θ^{Noise}). We conducted comprehensive experiments to investigate the insights and applicability of our approach, which are detailed in Appendix E.

3.3 Optimization Objective

Formally, DVRP optimizes a unified objective that balances task performance with intrinsic visual

Method	General Multimodal Reasoning								Medical Multimodal Reasoning						Overall	
	Geo3k	Vista	WeMath	MVerse	MVerse-V	MMKI2	AVG	Δ_{rel}	Slake	Path	Rad	PMC	AVG	Δ_{rel}	AVG	Δ_{rel}
	Qwen2.5-VL-3B Backbone															
Base Model	20.6	40.6	23.9	30.9	28.2	34.8	29.8	*Ref.*	48.7	59.2	40.3	46.8	48.7	*Ref.*	37.4	*Ref.*
GRPO	28.7	59.3	58.9	55.3	52.2	57.2	51.9	↑74.2%	70.9	74.5	71.2	53.4	67.5	↑38.6%	58.2	↑55.6%
DAPO	31.2	60.9	60.0	56.3	53.0	**66.8**	54.7	↑83.6%	71.3	72.4	70.8	56.6	67.8	↑39.2%	59.9	↑60.2%
NoiseRollout	32.5	63.0	60.1	56.9	53.5	63.4	54.9	↑84.2%	72.5	76.1	75.0	61.0	71.2	↑46.2%	61.4	↑64.2%
PAPO	30.9	61.3	60.1	57.1	53.9	57.3	53.4	↑79.2%	72.1	75.9	74.8	60.7	70.9	↑45.6%	60.4	↑61.5%
DVRP$_D$ (Ours)	**35.1**	64.9	**60.5**	**58.1**	**54.8**	60.9	**55.7**	**↑86.9%**	76.3	**78.9**	75.9	**62.2**	**73.3**	**↑50.5%**	**62.7**	**↑67.6%**
DVRP$_G$ (Ours)	34.5	**65.5**	60.3	57.7	54.5	61.2	55.6	↑86.6%	**74.1**	77.5	**76.2**	61.5	72.3	↑48.5%	62.3	↑66.6%
	Qwen2.5-VL-7B Backbone															
Base Model	33.8	55.9	41.8	45.6	36.9	43.7	43.0	*Ref.*	63.7	60.6	61.3	52.2	59.5	*Ref.*	49.6	*Ref.*
GRPO	40.2	65.5	66.1	66.5	61.7	72.1	62.0	↑44.2%	74.7	75.7	74.9	56.2	70.4	↑18.3%	65.4	↑31.9%
DAPO	35.9	61.9	58.5	55.6	51.0	71.9	55.8	↑29.8%	74.9	76.2	67.8	57.2	69.0	↑16.0%	61.1	↑23.2%
NoiseRollout	39.7	67.8	65.3	66.1	62.8	70.5	62.0	↑44.2%	72.8	74.1	72.2	59.9	69.8	↑17.3%	65.1	↑31.3%
PAPO	40.2	69.5	66.7	68.4	64.9	72.5	63.7	↑48.1%	76.9	73.4	77.5	61.6	72.4	↑21.7%	67.2	↑35.5%
DVRP$_D$ (Ours)	**43.4**	70.9	67.8	**68.9**	65.3	74.1	65.1	↑51.4%	80.3	**79.7**	**80.5**	64.2	76.2	↑28.1%	69.5	↑40.1%
DVRP$_G$ (Ours)	42.3	**71.1**	**68.1**	67.4	**66.7**	**75.6**	**65.2**	**↑51.6%**	**81.7**	78.3	79.9	**65.5**	**76.4**	**↑28.4%**	**69.7**	**↑40.5%**

Table 1: Performance (**avg@8** acc %) comparison of Qwen2.5-VL-3B and 7B backbones. We compare our **DVRP** against baselines including GRPO, DAPO, and PAPO. Subscripts $_D$ and $_G$ denote optimizations using DAPO and GRPO, respectively. The **Base Model** (highlighted) serves as the reference for calculating relative improvements ($\Delta_{rel} = \frac{\text{Method}-\text{Base}}{\text{Base}} \times 100\%$). For a fair comparison, all baselines are reproduced in the same environment and evaluated using the identical suite.

grounding. We integrate visual triplet supervision into the GRPO framework **to enforce** two complementary constraints: (1) **Visual Sensitivity:** The policy must diverge from the decremental view (π_θ^{Mask}) where visual evidence is absent; (2) **Visual Robustness:** The policy must remain consistent with the incremental view (π_θ^{Noise}) despite perturbations.

However, auxiliary branches under corrupted or masked views may produce overly diffuse predictive distributions, which weakens the training signal carried by the KL terms. We therefore add an entropy penalty on the auxiliary branches to discourage high-uncertainty predictions and stabilize the differential constraints. Consequently, the final optimization objective $\mathcal{J}_{\text{DVRP}}$ is formulated as:

$$\begin{aligned}\mathcal{J}_{\text{DVRP}}(\theta) = \quad & \underbrace{\mathcal{J}_{\text{GRPO}}(\theta)}_{\text{Maximize Reward}} \\ & \underbrace{+\lambda_{nec} \cdot \mathbb{D}_{KL}\left(\pi_\theta^{Ori} \parallel \pi_\theta^{Mask}\right)}_{\text{Visual Sensitivity (Max Difference)}} \\ & \underbrace{-\lambda_{rob} \cdot \mathbb{D}_{KL}\left(\pi_\theta^{Ori} \parallel \pi_\theta^{Noise}\right)}_{\text{Visual Robustness (Min Difference)}} \\ & \underbrace{-\lambda_{ent} \cdot \mathbb{E}\left[\mathcal{H}(\pi_\theta^{Noise}) + \mathcal{H}(\pi_\theta^{Mask})\right]}_{\text{Entropy Penalty (Prevent Collapse)}},\end{aligned} \tag{4}$$

where $\pi_{\theta_{old}}$ (implicitly used in $\mathcal{J}_{\text{GRPO}}$) serves as the frozen reference distribution to stabilize the reward maximization, while the triplet objectives enforce constraints on the current policy manifold.

4 Experiments

4.1 Experiment Setup and Training Details

Implementation Framework. Our experimental framework is built upon Easy-R1 (Zheng et al., 2025a), implemented using Python 3.10 and PyTorch 2.4.0 (Paszke et al., 2019) with CUDA 12.4 support. Following the training paradigm established by DeepSeek-R1 (DeepSeek-AI et al., 2025), we employed Direct Reinforcement Fine-Tuning (RFT) to optimize the Qwen2.5-VL-7B and 3B backbones (Qwen et al., 2025). All experiments were conducted on a computational cluster equipped with 4× NVIDIA A800 GPUs.

Training Datasets. To foster robust reasoning across domains, we employ ViRL39K (Wang et al., 2025a) for mathematical training, while for the medical domain, we construct a composite dataset by amalgamating the training splits of Slake (Liu et al., 2021), PathVQA (He et al., 2020), Rad-VQA (Lau et al., 2018), and PMC-VQA (Zhang et al., 2024b).

Training Details. To demonstrate the versatility of the proposed *DVRP*, we integrated DVRP with two fundamental RLVR algorithms, DAPO (Yu et al., 2025) and GRPO (Shao et al., 2024), and evaluated its impact on the Qwen2.5-VL-7B and 3B backbones (Qwen et al., 2025). We conducted extensive ablation studies to determine the optimal masking ratios and noise levels (Jiang et al., 2025). Details are provided in Appendix D. Following the methodology of NoiseRollout (Liu et al., 2025a),

Method	In-Domain Datasets				Out-of-Domain Datasets	
	PMCVQA	VQA-RAD	SLAKE	PathVQA	MedXpertQA	MMMU-Med
Proprietary Models						
Gemini-2.0-flash-lite	50.8	59.4	73.1	64.9	22.3	58.7
GPT-4.1-Nano	53.1	61.8	73.1	70.6	22.7	60.6
General-purpose Multimodal VLMs						
Qwen-VL-Chat (Wang et al., 2024b)	36.6	47.0	56.0	55.1	18.5	32.7
Yi-VL-34B (AI et al., 2025)	39.5	53.0	58.9	47.3	19.2	41.5
LLaVA-v1.6-7B (Liu et al., 2024b)	35.5	52.6	57.9	47.9	–	33.1
LLaVA-v1.6-13B (Liu et al., 2024b)	36.6	55.8	58.9	51.9	20.1	39.3
LLaVA-v1.6-34B (Liu et al., 2024b)	44.4	58.6	67.3	59.1	22.8	48.8
LLaVA-v1.5-LLaMA3-8B (Contributors, 2023)	36.4	54.2	59.4	54.1	18.7	38.2
Medical Multimodal VLMs						
Med-Flamingo (Moor et al., 2023)	23.3	45.4	43.5	54.7	22.1	28.3
RadFM (Wu et al., 2023)	25.9	50.6	34.6	38.7	23.4	27.0
LLaVA-Med-7B (Li et al., 2023)	24.7	51.4	48.6	56.8	20.8	36.9
LLaVA_Med-LLaMA3-8B (Contributors, 2023)	46.6	60.2	61.2	54.5	19.8	41.1
PubMedVision-8B (Chen et al., 2024a)	52.7	63.8	74.5	59.9	21.2	49.1
HuatuoGPT-Vision-34B (Chen et al., 2024a)	58.2	68.1	76.9	63.5	22.1	54.4
MedVLThinker-7B (Huang et al., 2025b)	57.5	63.7	67.8	65.2	20.9	57.0
CAPO-7B (Jiang et al., 2025)	55.5	78.5	79.1	68.9	22.5	60.0
Medical Agentic Systems						
MedAgents (Tang et al., 2024)	59.5	65.6	67.9	63.2	–	49.7
MDAgents (Kim et al., 2024)	62.8	66.8	68.2	65.4	–	52.3
AFlow (Zhang et al., 2025b)	61.5	67.3	68.9	66.4	–	53.6
MMedAgent-RL-7B (Xia et al., 2025)	62.4	71.5	76.2	72.3	–	**66.4**
Ours						
Base Model *(Qwen2.5-VL-3B)*	46.8	40.3	48.7	59.2	20.7	31.5
+ DVRP$_G$	61.5	76.2	74.1	77.5	23.2	42.4
	+14.7	+35.9	+25.4	+18.3	+2.5	+10.9
+ DVRP$_D$	62.2	75.9	76.3	<u>78.9</u>	<u>24.1</u>	44.1
	+15.4	+35.6	+27.6	+19.7	+3.4	+12.6
Base Model *(Qwen2.5-VL-7B)*	52.2	61.3	63.7	60.6	20.1	54.7
+ DVRP$_G$	**65.5**	<u>79.9</u>	**81.7**	78.3	24.0	<u>66.2</u>
	+13.3	+18.6	+18.0	+17.7	+3.9	+11.5
+ DVRP$_D$	<u>64.2</u>	**80.5**	<u>80.3</u>	**79.7**	**25.7**	65.9
	+12.0	+19.2	+16.6	+19.1	+5.6	+11.2

Table 2: Comprehensive performance comparison on 2D medical VQA benchmarks. **Bold** and <u>underline</u> denote best and second-best scores. The green values indicate the absolute improvement over the corresponding Base Model. Subscripts $_D$ and $_G$ denote optimizations using DAPO and GRPO, respectively.

we adopted a sigmoid function for diffusion noise annealing (Detailed in Section B.4). We also adopt random patch masking as the visual sensitivity operation in this work, a choice that has been discussed in (Wang et al., 2025g). Specifically, the KL divergence terms are computed as the summation of **token-level KL divergences** between the categorical output distributions of the compared policies at each generation step, calculated over the trajectories sampled from the invariant policy π_θ^{Ori}. Detailed hyperparameter configurations are provided in Appendix B.2.

Evaluation Protocol. We conduct a comprehensive evaluation across benchmarks spanning both in-domain and out-of-domain settings. Our comparative analysis includes a wide range of baselines, comprising closed-source commercial models, open-source general-purpose models, and domain-specific reasoning models. For detailed experimental settings, please refer to Appendix B.1.

4.2 Main Experiments

4.2.1 RLVR Experiments

We benchmark DVRP against representative RLVR baselines (Liu et al., 2025a) (Table 1). On the 7B scale, our approach consistently establishes

Method	Mathematical Multimodal Reasoning						
	Geo3k	Vista	WeMath	MVerse	MVerse-V	MMKI2	AVG
Proprietary Models							
GPT-4o-mini	–	59.9	56.3	42.3	45.1	51.9	51.1
Gemini-2.0-flash	–	70.4	47.4	47.8	48.7	65.2	55.9
General-purpose Multimodal VLMs							
Qwen2.5-VL-72B (Wang et al., 2024b)	–	**74.2**	49.1	47.3	48.6	70.5	57.9
InternVL2.5-8B (Chen et al., 2025c)	41.2	64.9	44.9	37.0	40.2	46.8	45.8
InternVL2.5-VL-78B (Chen et al., 2025c)	–	64.9	44.9	37.0	40.2	59.8	49.4
LLaVA-OneVision-7B (Li et al., 2024)	38.5	58.5	44.1	45.2	–	–	46.6
LLaVA-OneVision-72B (Li et al., 2024)	–	67.1	32.0	27.2	30.1	–	39.1
LLaVA-OneVision-1.5-8B (An et al., 2025)	43.8	69.6	61.5	–	–	56.4	57.8
LLaVA-Critic-R1-7B (Wang et al., 2025c)	35.4	68.7	62.6	58.9	53.1	57.4	56.0
R1-OneVision-7B (Yang et al., 2025)	30.6	64.9	55.2	61.7	44.3	43.3	50.0
Math-Specific Multimodal VLMs							
MM-Eureka-7B (Meng et al., 2025a)	36.4	59.1	45.3	57.6	56.4	60.6	52.6
MM-Eureka-7B-CPGD (Liu et al., 2025c)	37.6	64.2	64.3	63.7	59.2	64.7	59.0
ADORA-7B (Gui and Ren, 2025)	41.2	61.1	53.0	45.2	41.8	49.8	48.7
R1-VL-7B (Zhang et al., 2025c)	31.9	63.5	56.1	42.0	43.2	55.3	48.7
VLAA-Thinker-7B (Chen et al., 2025a)	24.2	67.4	65.9	47.9	52.0	63.2	53.4
VL-Rethinker-7B (Wang et al., 2025a)	33.6	61.3	66.5	64.0	60.8	59.8	57.7
RACRO-7B-CRO-GRPO (Gou et al., 2025b)	41.4	61.7	**68.9**	65.7	61.7	70.5	61.7
ThinkLite-7B-VL (Wang et al., 2025d)	34.4	68.9	63.5	49.5	46.0	56.2	53.1
Ours							
Base Model *(Qwen2.5-VL-3B)*	20.6	40.6	23.9	30.9	28.2	34.8	29.8
+ DVRP$_D$	35.1	64.9	60.5	58.1	54.8	60.9	55.7
	+14.5	+24.3	+36.6	+27.2	+26.6	+26.1	+25.9
+ DVRP$_G$	34.5	65.5	60.3	57.7	54.5	61.2	55.6
	+13.9	+24.9	+36.4	+26.8	+26.3	+26.4	+25.8
Base Model *(Qwen2.5-VL-7B)*	33.8	55.9	41.8	45.6	36.9	43.7	43.0
+ DVRP$_D$	**43.4**	70.9	67.8	**68.9**	<u>65.3</u>	<u>74.1</u>	<u>65.1</u>
	+9.6	+15.0	+26.0	+23.3	+28.4	+30.4	+22.1
+ DVRP$_G$	<u>42.3</u>	<u>71.1</u>	<u>68.1</u>	<u>67.4</u>	**66.7**	**75.6**	**65.2**
	+8.5	+15.2	+26.3	+21.8	+29.8	+31.9	+22.2

Table 3: Performance comparison on mathematical multimodal reasoning benchmarks. **Bold** and <u>underline</u> denote the best and second-best performance, respectively. The green values indicate the absolute improvement over the corresponding Base Model. Subscripts $_D$ and $_G$ denote optimizations using DAPO and GRPO, respectively.

new state-of-the-art results. Implementing DVRP on GRPO (DVRP$_G$) yields an overall accuracy of 69.7%, representing a **40.5%** relative improvement over the base model and surpassing the strongest baseline PAPO by 2.5 points. In the medical domain, DVRP$_G$ reaches 76.4% accuracy with a relative gain of **28.4%**, demonstrating differential visual constraints effectively enhance robustness in specialized reasoning tasks. For further case studies and reasoning consistency evaluation, please refer to Appendix J.

The performance gain is even more pronounced on the smaller 3B backbone. DVRP$_D$ boosts the overall accuracy from 37.4% to 62.7%, achieving a remarkable **67.6%** relative improvement. These results validate that DVRP effectively compensates for limited model capacity by maximizing the utilization of visual evidence. Additionally, we verified our method on QWEN3-VL (Bai et al., 2025)(Detailed in Appendix H).

4.2.2 Domain Foundation Model Experiments

We evaluate DVRP against a wide spectrum of proprietary, general-purpose, and domain-specific

Method	General Multimodal Reasoning							Medical Multimodal Reasoning					Overall
	Geo3k	Vista	WeMath	MVerse	MVerse-V	MMKI2	AVG	Slake	Path	Rad	PMC	AVG	AVG
	Qwen2.5-VL-7B Backbone												
Base Model (Original)	33.8	55.9	41.8	45.6	36.9	43.7	43.0	63.7	60.6	61.3	52.2	59.5	49.6
Baseline (GRPO)	40.2	65.5	66.1	66.5	61.7	72.1	62.0	74.7	75.7	74.9	56.2	70.4	65.4
+ Visual Sensitivity	41.5	68.8	67.2	**67.8**	64.8	73.9	64.0	78.5	**79.2**	78.2	61.4	74.3	68.1
+ Visual Robustness	40.9	67.2	66.6	67.0	63.2	73.0	63.0	76.2	76.5	**80.2**	59.1	73.0	67.0
DVRP$_G$ (Full)	**42.3**	**71.1**	**68.1**	67.4	**66.7**	**75.6**	**65.2**	**81.7**	78.3	79.9	**65.5**	**76.4**	**69.7**

Table 4: Ablation study of **Visual Sensitivity** (via masking) and **Visual Robustness** (via noise injection) on the Qwen2.5-VL-7B backbone. The **Base Model** (gray) represents the original pre-trained weights, while the **Baseline** utilizes the standard GRPO algorithm. While specific constraints (e.g., masking on PathVQA or noise on VQA-RAD) may yield marginal gains on individual datasets, the full **DVRP**$_G$ framework (green) achieves the best overall performance, demonstrating the synergy of the combined objectives.

foundation models across medical and mathematical benchmarks. In the medical domain, as detailed in Table 2, our 7B model establishes new state-of-the-art performance on in-domain tasks, consistently outperforming specialized baselines like CAPO-7B (Jiang et al., 2025) and HuatuoGPT-Vision-34B (Chen et al., 2024a). Notably, on PathVQA, DVRP$_D$ achieves **79.7%** accuracy, surpassing the proprietary models and rivaling complex agentic systems like MMedAgent-RL without requiring external tools or trajectory data (Zhang et al., 2025b).

Turning to the mathematical domain in Table 3, DVRP demonstrates superior generalization capabilities. Our 7B model attains an average accuracy of **65.2%**, significantly outperforming the 72B-parameter Qwen2.5-VL at 57.9%. It also establishes a clear advantage over math-specific baselines such as RACRO-7B (Gou et al., 2025a) and MM-Eureka (Meng et al., 2025a), which score 61.7% and 59.0% respectively. These results indicate that by explicitly enforcing differential visual sensitivity and robustness, DVRP ensures consistent generalization across diverse datasets and model scales. Crucially, this approach empowers parameter-efficient models to rival or even exceed the reasoning capabilities of significantly larger counterparts and more sophisticated agentic reasoning systems (Xia et al., 2025). Additional evaluations about perception and spatial understanding are presented in Appendix C.

4.3 Ablation Experiments

To investigate the individual contributions of the proposed components, we conduct an ablation study on the Qwen2.5-VL-7B backbone using GRPO as the baseline (Table 4). The introduction of *Visual Sensitivity* yields a substantial improvement over the baseline, confirming that penalizing blind reasoning is critical for enforcing stronger visual dependence. Similarly, the *Visual Robustness* term independently enhances performance by fostering stability against perturbations. While individual components may occasionally outperform the unified model on specific tasks, the full DVRP framework achieves the best overall performance. This demonstrates that sensitivity and robustness constraints are complementary, effectively synergizing to maximize reasoning reliability.

Furthermore, we explore the impact of perturbation intensity across domains. Our experiments reveal a domain-specific dichotomy: general multimodal reasoning (e.g., Math) benefits from aggressive perturbations ($P_{mask} = 0.6, T_{init} = 500$) to enforce structural dependency, whereas medical reasoning requires milder regularization ($P_{mask} = 0.2, T_{init} = 100$) to preserve fine-grained pathological features. We provide a detailed discussion of these hyperparameter sensitivities in Appendix D.

5 Conclusion

In this work, we address the critical perception-reasoning decoupling in current multimodal RLVR paradigms by proposing **Thinking with Deltas**. Driven by the Differential Visual Reasoning Policy, our framework leverages self-supervised visual triplets to introduce intrinsic training signals that encourage reasoning behavior to depend on the presence and stability of visual evidence. This approach enhances perception without relying on external dependencies. Extensive experiments across mathematical and medical domains demonstrate that DVRP effectively bridges the perception-reasoning decoupling and empowers parameter-efficient models to achieve sota performance, often rivaling significantly larger commercial baselines.

Limitations

Despite the robust performance and generalization capabilities our method demonstrates across diverse domains and comparative settings, several limitations remain. First, our empirical evaluation is currently confined to models with small parameters. Due to computational resource constraints, we have not yet scaled the proposed framework to larger foundation models (e.g., 70B+ parameters). But we have empirically verified its efficacy on two different backbones. Second, our current approach focuses on intrinsic policy optimization and has not yet explored integration with agentic systems. Synergizing DVRP with multi-agent frameworks or interactive tool-use pipelines to achieve more sophisticated reasoning strictly grounded on visual evidence remains an unexplored frontier. We hope this work inspires future research into differential constraints as a minimalist yet powerful paradigm for building trustworthy and visually grounded multimodal systems. Finally, because our evaluation includes medical VQA tasks, DVRP should be viewed only as a research framework rather than a clinical decision-support system, and any real-world medical use would require expert validation, bias assessment, and safeguards against harmful diagnostic errors.

Acknowledgements

This work is supported the National Natural Science Foundation of China (Grant No. 62472098), the Science and Technology Commission of Shanghai Municipality (No. 25511106100).

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, and 14 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

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

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *Preprint*, arXiv:2406.19280.

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, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024c. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, and 1 others. 2024d. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025b. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. 2025. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *Preprint*, arXiv:2505.17018.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. 2025a. Perceptual decoupling for scalable multi-modal reasoning via reward-optimized captioning. *arXiv preprint arXiv:2506.04559*.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. 2025b. Reasoning-aligned perception decoupling for scalable multi-modal reasoning. *Preprint*, arXiv:2506.04559.

Lujun Gui and Qingnan Ren. 2025. Training reasoning model with dynamic advantage estimation on reinforcement learning. https://www.notion.so/Training-Reasoning-Model-with-Dynamic-Advantage-Estimation-on-Reinforcement-Learning-1a830cc0904681fa9df3e076b6557a3e. Notion Blog.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962.

Feng Han, Yang Jiao, Shaoxiang Chen, Junhao Xu, Jingjing Chen, and Yu-Gang Jiang. 2025a. Controlthinker: Unveiling latent semantics for controllable image generation through visual reasoning. *arXiv preprint arXiv:2506.03596*.

Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. 2025b. Learning to see before seeing: Demystifying llm visual priors from language pre-training. *Preprint*, arXiv:2509.26625.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

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

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025b. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. 2025. Tree search for llm agent reinforcement learning. *arXiv preprint arXiv:2509.21240*.

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Sibo Song, Yang Feng, Jimeng Sun, Jian Wu, and Zuozhu Liu. 2025. Capo: Reinforcing consistent reasoning in medical decision-making. *Preprint*, arXiv:2506.12849.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Preprint*, arXiv:2404.15155.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Keyi Kong, Xilie Xu, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2024. Perplexity-aware correction for robust alignment with noisy preferences. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, and 1 others. 2025a. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Guilin Li, Bo Wang, Linghe Kong, Lichao Sun, and Weiran Huang. 2025b. Revisiting visual understanding in multimodal reasoning through a lens of image perturbation. *Preprint*, arXiv:2506.09736.

Yiqing Liang, Jielin Qiu, Wenhao Ding, Zuxin Liu, James Tompkin, Mengdi Xu, Mengzhou Xia, Zhengzhong Tu, Laixi Shi, and Jiacheng Zhu. 2025. Modomodo: Multi-domain data mixtures for multimodal llm reinforcement learning. *Preprint*, arXiv:2505.24871.

Mengqi Liao, Xiangyu Xi, Ruinian Chen, Jia Leng, Yangen Hu, Ke Zeng, Shuai Liu, and Huaiyu Wan. 2025. Enhancing efficiency and exploration in reinforcement learning for llms. *Preprint*, arXiv:2505.18573.

Kevin Qinghong Lin, Yuhao Zheng, Hangyu Ran, Dantong Zhu, Dongxing Mao, Linjie Li, Philip Torr, and Alex Jinpeng Wang. 2025. Vcode: a multimodal coding benchmark with svg as symbolic visual representation. *Preprint*, arXiv:2511.02778.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *Preprint*, arXiv:2102.09542.

Haogeng Liu, Quanzeng You, Xiaotian Han, Yongfei Liu, Huaibo Huang, Ran He, and Hongxia Yang. 2024a. Visual anchors are strong information aggregators for multimodal large language model. *arXiv preprint arXiv:2405.17815*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2023. Llava-plus: Learning to use tools for creating multimodal agents. *Preprint*, arXiv:2311.05437.

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

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual agentic reinforcement fine-tuning. *Preprint*, arXiv:2505.14246.

Zongkai Liu, Fanqing Meng, Lingxiao Du, Zhixiang Zhou, Chao Yu, Wenqi Shao, and Qiaosheng Zhang. 2025c. Cpgd: Toward stable rule-based reinforcement learning for language models. *Preprint*, arXiv:2505.12504.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

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 *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14268–14280.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025a. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2503.07365.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, and 1 others. 2025b. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: a multimodal medical few-shot learner. *Preprint*, arXiv:2307.15189.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. 2025. Grpo-λ: Credit assignment improves llm reasoning. *Preprint*, arXiv:2510.00194.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and 1 others. 2024. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

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*.

Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and 1 others. 2025b. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, and 1 others. 2024a. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. 2025c. Llava-critic-r1: Your critic model is secretly a strong policy model. *Preprint*, arXiv:2509.00676.

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

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2025e. Reinforcement learning for reasoning in large language models with one training example. *Preprint*, arXiv:2504.20571.

Yuan Wang, Shujian Gao, Jiaxiang Liu, Songtao Jiang, Haoxiang Xia, Xiaotian Zhang, Zhaolu Kang, Yemin Wang, and Zuozhu Liu. 2025f. Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation. *arXiv preprint arXiv:2512.02710*.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. 2025g. Perception-aware policy optimization for multimodal reasoning. *Preprint*, arXiv:2507.06448.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Preprint*, arXiv:2308.02463.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. 2025. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*.

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

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

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024b. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. 2025a. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025b. Aflow: Automating agentic workflow generation. *Preprint*, arXiv:2410.10762.

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

Jixiao Zhang and Chunsheng Zuo. 2025. Grpo-lead: A difficulty-aware reinforcement learning approach for

concise mathematical reasoning in language models. *Preprint*, arXiv:2504.09696.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Yuting Zhang, Kaishen Yuan, Hao Lu, Yutao Yue, Jintai Chen, and Kaishun Wu. 2025d. Medtvt-r1: A multimodal llm empowering medical reasoning and diagnosis. *Preprint*, arXiv:2506.18512.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024c. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025a. Easyr1: An efficient, scalable, multi-modality rl training framework.

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

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *Preprint*, arXiv:2501.18362.

A Appendix

This appendix provides implementation details, additional quantitative ablations, generalization studies, computational analysis, and extensive qualitative examples to support the findings in the main paper. The organization is as follows:

- **§B Implementation Details**. We provide a detailed breakdown of our experimental setup, including:
  - **§B.1 Evaluation Details**: Detailed protocols for mathematical and medical benchmarks.
  - **§B.2 Hyperparameters**: Training hyperparameters, inference settings, and DVRP coefficients.
  - **§B.3 Prompt Template**: The standardized system prompt and reasoning template used across all experiments.
  - **§B.4 Noise Scheduling**: The variance-preserving diffusion formulation and sigmoid annealing schedule used to modulate perturbation intensity.
  - **§B.5 Computational Profile**: Runtime and memory analysis showing that DVRP reuses the invariant-view rollout and avoids a $3\times$ autoregressive generation cost.
- **§C Generalization to Spatial and Real-World Reasoning**. We evaluate DVRP on GQA and MMStar to verify transfer beyond structured math and medical benchmarks, especially under compositional spatial reasoning and vision-indispensable settings.
- **§D Ablation on Perturbation Parameters**. We present the sensitivity analysis for masking ratios (P_{mask}) and noise steps (T_{init}), highlighting the domain-specific perturbation trade-off between structural robustness and fine-grained visual fidelity.
- **§E Visual Robustness and Sensitivity Analysis**. We analyze the contrasting behaviors of MLLMs under diffusion noise and semantic masking, motivating the robustness and sensitivity objectives in DVRP.
- **§F RLVR Visual Dependency Experiments**. We detail the blind experiments, including Text-Only and Blank Image settings, used to diagnose whether RLVR policies genuinely rely on visual evidence.
- **§G Comparison with Explicit Data Augmentation**. We compare DVRP with a Tri-View Explicit-Augmentation baseline that uses the same original, masked, and noised views as ordinary training samples, isolating whether the gains come from differential KL constraints rather than additional exposure to perturbed visual inputs.
- **§H Generalization to Newer Vision-Language Backbones**. We evaluate DVRP on Qwen3-VL-4B and Qwen3-VL-8B to test whether the gains transfer to stronger and more recent vision-language backbones without hyperparameter re-tuning.
- **§I Qualitative Analysis of Visual Dependency**. We provide case studies showing how

baseline policies exploit linguistic shortcuts or hallucinate visual evidence under blind settings.

- **§J Qualitative Comparison of RLVR Algorithms**. We compare reasoning trajectories from GRPO, DAPO, and DVRP-D, focusing on visual faithfulness, logical consistency, and perception-reasoning coupling.

B Implementation Details

B.1 Evaluation Details

Evaluation Protocols. To comprehensively assess model performance, we conducted evaluations across both general (mathematical) and medical domains, distinguishing between in-domain and out-of-domain (OOD) settings.

- **Mathematical Evaluation:** We employed Geo3k (Lu et al., 2021), Vista (Lu et al., 2024), WeMath (Qiao et al., 2024), MVerse (Zhang et al., 2024a), MVerse-V (Zhang et al., 2024a), and MMKI2 (Meng et al., 2025b) as OOD benchmarks to test generalization capabilities. Furthermore, we utilized `mathruler.grader` to facilitate precise evaluation.
- **Medical Evaluation:** We utilized the test splits of Slake (Liu et al., 2021), PathVQA (He et al., 2020), RadVQA (Lau et al., 2018), and PMC-VQA (Zhang et al., 2024b) for in-domain evaluation. Furthermore, MedXpertQA (Zuo et al., 2025) and MMMU-Med (Yue et al., 2024a) were employed to assess OOD performance.

To ensure the statistical reliability of our results, we report the Average Accuracy over 8 runs (**AVG@8 Acc**). Concretely, for each evaluation sample we draw 8 independent stochastic rollouts from the policy under the inference configuration described above (temperature 1.0, top-p 0.9), score each rollout with the rule-based reward function $r(o, a)$, and report the arithmetic mean of the 8 binary accuracies. We deliberately adopt this expectation-style metric rather than Pass@N (best-of-N) or majority voting in order to faithfully account for the sampling variance induced by stochastic decoding and to avoid inflating performance via test-time aggregation. AVG@8 reduces decoding variance, but full multi-seed RL training is computationally expensive; therefore reliability is supported by repeated evaluation, controlled baselines, ablations, and blind dependency diagnostics. For inference, we deployed the vLLM engine to accelerate generation (Kwon et al., 2023). For fair comparison, all models utilized a unified system prompt (reasoning template) to elicit chain-of-thought reasoning, with the temperature set to 1.0 and top-p to 0.9. The specific templates used are detailed in Appendix B.3.

Baselines. To conduct a comprehensive comparative analysis, we evaluate our proposed method against baselines categorized into two distinct dimensions: RL-based optimization methods and foundational MLLMs.

RL-based Methods. We benchmark our approach against state-of-the-art RL strategies designed for reasoning or visual alignment, including **GRPO** (Shao et al., 2024), **DAPO** (Yu et al., 2025), **PAPO** (Wang et al., 2025g), and **NoiseRollout** (Liu et al., 2025a). To ensure a fair comparison, we re-trained these RL methods using the identical training set and evaluation suite. For PAPO, we reproduce it with base GRPO and DAPO, and report with higher results.

Foundation Models. We compare against a diverse spectrum of baselines ranging from proprietary commercial systems to specialized open-source models:

- *Medical-Specific Models:* We evaluate performant open-source models tailored for the biomedical domain, such as **HuatuoGPT-Vision-34B** (Chen et al., 2024a).
- *Math-Specific Models:* For mathematical reasoning tasks, we include domain-expert models like **MM-Eureka** (Meng et al., 2025a) and **ThinkLite-7B** (Wang et al., 2025d).
- *General-Purpose Models:* We also include widely adopted general open-source benchmarks, such as **Qwen-VL** (Bai et al., 2023) and **LLaVA-v1.6** (Liu et al., 2024b).

B.2 Hyperparameter

To ensure reproducibility, we provide a comprehensive overview of the training and evaluation hyperparameters in Table 5 and Table 6. All models are optimized using AdamW with a constant learning rate of 1e-6 and a global batch size of 128. Consistent with recent reinforcement learning practices, we remove the KL penalty relative to the reference model to facilitate broader policy exploration (Liu et al., 2025a; Yu et al., 2025).

Hyperparameter	Value
General Training	
Optimizer	AdamW (BF16)
Learning Rate	1e-6
LR Schedule	Constant
Total Epochs	3
RL Process (GRPO)	
Global Batch Size	128
Rollout Batch Size	384
Group Size (G)	5
KL Penalty (Ref Model)	None (Disabled)
Reward Signal	Accuracy
DVRP Objectives	
Visual Robustness Loss (λ_{rob})	0.01
Visual Sensitivity Loss (λ_{nec})	0.01
Entropy Regularization (λ_{ent})	0.05
Visual Perturbation	**Math / Medical**
Mask Probability (P_{mask})	0.6 / 0.2
Patch Size (P_{mask})	14 / 14
Noise Steps (T_{init})	500 / 100
Inference	
Temperature	1.0
Top-p	0.90
Max New Tokens	2048
Resources & Efficiency	
Compute	4× NVIDIA A800
Step Time	∼65–70s

Table 5: Key hyperparameters for training and evaluation. DVRP parameters vary by domain (Math vs. Medical). Notations correspond to Eq. (4).

Hyperparameter	Value
General Training	
Optimizer	AdamW (BF16)
Learning Rate	1e-6
LR Schedule	Constant
Total Epochs	3
RL Process (DAPO)	
Global Batch Size	128
Rollout Batch Size	384
Mini-Rollout Batch Size	128
Group Size (G)	5
KL Penalty (Ref Model)	None (Disabled)
Reward Signal	Accuracy
DVRP Objectives	
Visual Robustness Loss (λ_{rob})	0.01
Visual Sensitivity Loss (λ_{nec})	0.01
Entropy Regularization (λ_{ent})	0.05
Visual Perturbation	**Math / Medical**
Mask Probability (P_{mask})	0.6 / 0.2
Patch Size (P_{mask})	14 / 14
Noise Steps (T_{init})	500 / 100
Inference	
Temperature	1.0
Top-p	0.90
Max New Tokens	2048
Resources & Efficiency	
Compute	4× NVIDIA A800
Step Time	∼65–70s

Table 6: Key hyperparameters for training and evaluation based on the DAPO + DVRP configuration.

Crucially, our **DVRP** framework introduces three auxiliary objectives regulated by specific coefficients: Visual Robustness ($\lambda_{rob} = 0.01$), Visual Sensitivity ($\lambda_{nec} = 0.01$), and Entropy Regularization ($\lambda_{ent} = 0.05$). The above loss weights are consistent with previous works (Wang et al., 2025g; Liu et al., 2025a). To address the varying information density of visual modalities, we employ domain-adaptive perturbation strategies. As detailed in Table 5, we apply a higher mask probability ($P_{mask} = 0.6$) and longer noise injection steps ($T_{init} = 500$) for mathematical reasoning, while adopting milder perturbations ($P_{mask} = 0.2, T_{init} = 100$) for medical tasks to preserve fine-grained pathological details.

B.3 Prompt Template

For the prompt selection, we employ a standardized template across all training and evaluation stages to elicit stable Chain-of-Thought (CoT) reasoning. As illustrated in Figure 3, the template explicitly instructs the model to generate an internal monologue within `<think>` tags before producing the final answer, ensuring the output format remains consistent for automated parsing and reward calculation.

B.4 Noise Scheduling

We adopt a variance-preserving (VP) diffusion process to construct the incremental view I_{noise}, simulating environmental instability while maintaining semantic consistency. Given an original image I and a noise intensity coefficient $\beta \in [0, 1]$, the perturbed input is formulated as:

$$I_{\text{noise}} = \sqrt{1-\beta} \cdot I + \sqrt{\beta} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (5)$$

where ϵ represents standard Gaussian noise. The coefficient β determines the signal-to-noise ratio; a higher β introduces stronger perturbation.

To balance structural robustness learning with convergence stability, we introduce a curriculum learning strategy via a Sigmoid decay schedule.

Method / Setting	General Multimodal Reasoning							Medical Multimodal Reasoning					Overall
	Geo3k	Vista	WeMath	MVerse	MVerse-V	MMKI2	AVG	Slake	Path	Rad	PMC	AVG	AVG
Qwen2.5-VL-7B Backbone													
Base Model (Original)	33.8	55.9	41.8	45.6	36.9	43.7	43.0	63.7	60.6	61.3	52.2	59.5	49.6
Baseline (GRPO)	40.2	65.5	66.1	66.5	61.7	72.1	62.0	74.7	75.7	74.9	56.2	70.4	65.4
Ablation I: Sensitivity to Masking Ratio (P_{mask})	(Math favors high mask, Medical favors low mask)												
+ Masking ($P = 0.2$)	41.0	68.2	66.8	66.9	63.5	73.5	63.3	78.5	**79.2**	78.2	61.4	**74.3**	67.7
+ Masking ($P = 0.4$)	41.8	69.5	67.5	67.1	65.1	74.8	64.3	76.8	78.1	77.4	60.2	73.1	67.8
+ Masking ($P = 0.6$)	**42.3**	**71.1**	**68.1**	**67.4**	**66.7**	**75.6**	**65.2**	74.5	76.5	75.8	58.9	71.4	67.7
Ablation II: Robustness to Noise Injection (T_{init})	(Math needs strong noise, Medical needs weak noise)												
+ Noise ($T = 100$)	41.2	68.5	66.9	67.0	64.2	73.9	63.6	76.2	76.5	**80.2**	59.1	**73.0**	67.4
+ Noise ($T = 300$)	41.9	70.1	67.6	67.2	65.8	74.9	64.6	75.1	75.8	78.4	57.8	71.8	67.5
+ Noise ($T = 500$)	**42.3**	**71.1**	**68.1**	**67.4**	**66.7**	**75.6**	**65.2**	73.5	74.2	76.1	56.5	70.1	67.2
DVRP$_G$ (Optimal)	**42.3**	**71.1**	**68.1**	67.4	**66.7**	**75.6**	**65.2**	**81.7**	78.3	79.9	**65.5**	**76.4**	**69.7**

Table 7: Comprehensive ablation study of the DVRP framework on Qwen2.5-VL-7B. We systematically vary the **Masking Ratio** ($P_{mask} \in \{0.2, 0.4, 0.6\}$) and **Noise Steps** ($T_{init} \in \{100, 300, 500\}$). The results highlight a domain-specific dichotomy: **General/Math** tasks benefit from aggressive perturbations ($P = 0.6, T = 500$) to enforce structural reasoning, whereas **Medical** tasks require milder regularization ($P = 0.2, T = 100$) to preserve fine-grained pathological features. The final **DVRP$_G$** model employs these domain-optimal settings.

Method / Setting	Medical Multimodal Reasoning				
	Slake	Path	Rad	PMC	AVG
Base Model (Original)	63.7	60.6	61.3	52.2	59.5
Baseline (GRPO)	74.7	75.7	74.9	56.2	70.4
Comparison of Noise Types	(Fixed scheduling, varying structure)				
+ Gaussian Noise ($\sigma = 0.1$)	75.2	76.0	76.5	57.0	71.2
+ Diffusion Noise ($T = 100$)	**76.2**	**76.5**	**80.2**	**59.1**	**73.0**

Table 8: Ablation study comparing noise injection strategies in the Medical domain. Under the DVRP framework, structured **Diffusion Noise** ($T = 100$) outperforms unstructured **Gaussian Noise** ($\sigma = 0.1$) by **+1.8%**, demonstrating that preserving semantic structure during perturbation is critical for medical visual reasoning.

Specifically, the noise intensity β is dynamically annealed based on the training progress. Let k be the current training step and K be the total steps. We compute the diffusion timestep t as:

$$t(k) = T_{\text{init}} \cdot \sigma\left(\gamma \cdot \left(0.5 - \frac{k}{K}\right)\right), \qquad (6)$$

where T_{init} is the initial noise step (e.g., 500), $\sigma(\cdot)$ is the sigmoid function, and γ is a scaling factor controlling the decay steepness (set to 10 in our experiments). The noise intensity is then derived as $\beta_k = t(k)/T_{\max}$. This schedule imposes high-variance perturbations in the early stages to enforce broad structural invariance, while gradually annealing the noise to zero to refine fine-grained visual consistency.

B.5 Computational Profile

Although DVRP introduces a visual triplet formulation, it does *not* incur a $3\times$ compute or memory blow-up. The key implementation insight is that the autoregressive rollout, which dominates wall-clock cost in RLVR, is performed **only once on the Invariant view** I. The Decremental (I_{mask}) and Incremental (I_{noise}) views are consumed exclusively during the actor update phase, where the model performs a single, fully parallelized forward pass over the *trajectory already sampled from* I to obtain the log-probabilities required by the differential KL terms. We thus avoid maintaining additional KV-caches or generation streams, and the augmented forward passes do not build extra autoregressive computation graphs.

We profile peak VRAM and wall-clock time per update step on a 4×NVIDIA A800 (80GB) cluster with the Qwen2.5-VL-7B backbone and FSDP enabled. As summarized in Table 9, DVRP introduces only an $\sim$18.6% increase in peak VRAM and an $\sim$26.1% increase in step time relative to standard GRPO, both of which fit comfortably within commodity 80GB GPU budgets. Empirically, this fractional overhead fundamentally re-shapes the optimization trajectory: simply training a single-view DAPO baseline for $\geq$26% additional steps does *not* close the gap to DVRP, indicating that the additional compute is invested in re-coupling perception and reasoning rather than in extra gradient updates over an already-saturated objective.

Method	AR Roll.	Fwd. Passes	Peak VRAM	Time/Step
GRPO	1	2 (C, R)	∼64.5 GB	∼65.0 s
DAPO	1	2 (C, R)	∼66.8 GB	∼68.5 s
DVRP (Ours)	1	4 (C, D, I, R)	∼76.5 GB (+18.6%)	∼82.0 s (+26.1%)

Table 9: Computational profile of DVRP versus RLVR baselines on Qwen2.5-VL-7B (4×A800-80GB, FSDP, BF16). **AR Roll.** = autoregressive rollout streams; **Fwd. Passes** counts forward passes in the actor update (C = Clean/Invariant, D = Decremental, I = Incremental, R = reference $\pi_{\theta_{\text{old}}}$). DVRP avoids the 3× blow-up by reusing the trajectory sampled from the Invariant view across all three branches.

Method	GQA	MMStar
Qwen-VL-Chat	57.5	37.5
LLaVA-Next-34B	67.1	52.1
GPT-4V (high)	65.4	57.1
Qwen2.5-VL-7B (Base)	64.5	53.2
+ GRPO	65.2	54.8
+ DVRP$_G$ (Ours)	**68.3** (+3.1)	**58.6** (+3.8)

Table 10: Out-of-domain generalization on **GQA** (compositional spatial reasoning) and **MMStar** (vision-indispensable benchmark eliminating linguistic shortcuts).

C Generalization to Spatial and Real-World Reasoning

The benchmarks in §4.2 stress structurally grounded or fine-grained perception. To verify that DVRP also transfers to less structured, real-world reasoning rather than merely tightening structural priors, we evaluate on two complementary out-of-domain benchmarks: **GQA** (Hudson and Manning, 2019), which probes compositional spatial reasoning over natural scenes, and **MMStar** (Chen et al., 2024c), a curated *vision-indispensable* benchmark whose questions are filtered to be unanswerable from textual priors alone, serving as a direct testbed for the blind-reasoner pathology in §1. Results are summarized in Table 10.

Compositional Spatial Reasoning. On GQA, DVRP$_G$ reaches 68.3%, surpassing the GRPO baseline by $+3.1\%$ and outperforming substantially larger counterparts (e.g., LLaVA-Next-34B) and the proprietary GPT-4V. This indicates the visual-robustness constraint preserves unstructured spatial relations and instead encourages alignment with object–attribute–relation cues that GQA explicitly probes.

Eliminating Linguistic Shortcuts. The $+3.8\%$ gain on MMStar (reaching 58.6%, exceeding GPT-4V) is particularly informative: by construction, MMStar removes the shortcuts exploited by standard RLVR policies in Figure 1. The fact that DVRP achieves its largest relative improvement precisely where shortcut exploitation is foreclosed provides direct evidence that the differential triplet objective re-couples reasoning with visual evidence.

Combined with the medical results in §4.2, these findings show that DVRP generalizes along two orthogonal axes of visual difficulty—*fine-grained perceptual dependence* and *commonsense spatial dependence*—rather than relying on domain-specific inductive biases.

D Ablation on Perturbation Parameters

We investigate the sensitivity of the DVRP framework to the intensity of visual perturbations. We perform ablations on the Qwen2.5-VL-7B (Qwen et al., 2025) backbone by varying the masking ratio $P_{mask} \in \{0.2, 0.4, 0.6\}$ and the noise injection steps $T_{init} \in \{100, 300, 500\}$. The results are detailed in Table 7.

Masking Ratio. We observe a distinct divergence in optimal masking ratios between domains. For **General Multimodal Reasoning** (e.g., Math), performance improves consistently as the masking ratio increases, peaking at $P_{mask} = 0.6$. This suggests that geometric diagrams and natural images contain high spatial redundancy; aggressive masking effectively forces the model to learn structural dependencies rather than relying on local textures. In contrast, **Medical Multimodal Reasoning** favors a conservative ratio ($P_{mask} = 0.2$). Performance drops significantly (from 74.3% to 71.4%) when the ratio is increased to 0.6. This indicates that medical diagnosis relies heavily on fine-grained visual details (e.g., small lesions or boundaries), which are easily destroyed by aggressive occlusion.

Noise Injection Level. A similar trend is observed for noise robustness. We implement noise injection via a diffusion process, utilizing a sigmoid schedule to modulate the noise intensity. In the general domain, the model benefits from stronger consistency regularization, achieving peak accuracy with higher diffusion timesteps ($T_{init} = 500$). Conversely, the medical domain is sensitive to high-intensity diffusion perturbations. Performance degrades as T_{init} increases, with the optimal setting

found at a lower timestep ($T_{init} = 100$). We hypothesize that excessive diffusion noise in medical images risks corrupting subtle pathological features or mimicking sensor artifacts, thereby confusing the reasoning policy.

Based on these observations, our final **DVRP**$_G$ model adopts domain-specific settings: we apply stronger perturbations ($P = 0.6, T = 500$) for math tasks to encourage robustness, and milder perturbations ($P = 0.2, T = 100$) for medical tasks to preserve visual fidelity.

Impact of Noise Structure. Table 8 further scrutinizes the efficacy of structured *Diffusion Noise* ($T = 100$) versus unstructured *Gaussian Noise* ($\sigma = 0.1$) in the medical domain ($P = 0.6$). The results demonstrate that Diffusion Noise consistently outperforms Gaussian Noise, achieving an average gain of **+1.8%**. This suggests that structured perturbations effectively regularize the policy while preserving essential semantic integrity, whereas unstructured Gaussian noise risks degrading the pixel-level fidelity required for fine-grained medical diagnostics.

E Visual Robustness and Sensitivity Analysis

To investigate the reliance of MLLMs on visual fidelity and their resilience to perturbations, we conduct a comprehensive robustness analysis on Qwen2.5-VL benchmarks (Qwen et al., 2025) across general and medical domains. We introduce two distinct types of visual perturbations: Diffusion Noise, which represents high-frequency corruption, and Patch Masking, which represents semantic information loss. As visualized in Figure 4, these perturbations induce contrasting reasoning behaviors: the model remains resilient and consistent under noise ($+\Delta$) but becomes confused and divergent when visual semantics are masked ($-\Delta$). This qualitative observation is quantitatively confirmed by the results in Table 11, serving as the empirical foundation for our proposed method.

Robustness to Diffusion Noise. As shown in the Noise rows of Table 11, MLLMs exhibit strong stability against diffusive perturbations. We implement noise injection via a diffusion scheduler controlled by a sigmoid function. Despite these perturbations, the Qwen2.5-VL-3B model maintains performance and even achieves a slight average gain of **+0.6%** in general reasoning tasks. Similarly, the 7B model shows negligible performance fluctuations ($\Delta < 0.3\%$) across most benchmarks. This indicates that current MLLMs possess inherent resilience to the high-frequency jitter introduced by the diffusion scheduler, likely due to the robust feature extraction capabilities of the vision encoder.

Insight I: This robustness validates the feasibility of our *Noise Consistency* objective. Since the model demonstrates invariance to diffusion noise, explicitly enforcing output consistency between clean and noisy views acts as a safe regularization term. This stabilizes the policy without degrading representation quality.

Sensitivity to Semantic Masking. Conversely, the models demonstrate high sensitivity to patch masking. When a portion of the visual input is occluded, performance degrades sharply. Notably, the 7B model suffers a severe average drop of **−20.9%** in the medical domain and **−15.4%** in general tasks. This degradation is significantly more pronounced than that caused by noise, indicating that reasoning depends heavily on specific visual patches rather than language priors alone.

Insight II: This sensitivity confirms that visual tokens are critical for correct reasoning. It motivates employing masking as a challenging view in the reinforcement learning loop. Exposing the policy to masked inputs forces the model to maximize information utilization from the remaining visible patches, thereby reducing hallucination and enhancing the grounding of the reasoning process.

Summary. MLLMs process these two perturbations through fundamentally different mechanisms. Diffusion noise acts as a low-level surface artifact to which pre-trained visual encoders exhibit strong invariance. In contrast, patch masking acts as a high-level semantic disruption that breaks the visual continuity required for logical deduction. This disparity highlights a critical insight: while visual representations are texturally robust, reasoning policies remain fragile to structural incompleteness. *This observation directly underpins our methodology, where we leverage diffusion noise for consistency regularization and masking for hardness-aware policy learning.*

F RLVR Visual Dependency Experiments

To rigorously investigate whether current RLVR methods genuinely leverage visual information or

SYSTEM:
You are a helpful assistant.

USER:
{question}

You first think through the reasoning process as an internal monologue, enclosed within <think> </think> tags. Then, provide your final answer enclosed within \boxed{}.

Figure 3: The standardized prompt template employed across all training and evaluation phases. To ensure consistent reasoning behaviors and facilitate automated answer extraction, we explicitly instruct the model to encapsulate its chain-of-thought within `<think>` tags and place the final result inside a `\boxed{}` command.

Setting	General Multimodal Reasoning							Medical Multimodal Reasoning				
	Geo3k	Vista	WeMath	MVerse	MVerse-V	MMK12	AVG	Slake	Path	Rad	PMC	AVG
	Qwen2.5-VL-3B Backbone											
Original	20.6	40.6	23.9	30.9	28.2	34.8	29.8	48.7	59.2	40.3	46.8	48.8
Noise	22.6$_{+2.0}$	41.4$_{+0.8}$	26.5$_{+2.6}$	30.8$_{-0.1}$	26.2$_{-2.0}$	35.1$_{+0.3}$	30.4$_{+0.6}$	48.1$_{-0.6}$	58.5$_{-0.7}$	40.6$_{+0.3}$	46.2$_{-0.6}$	48.4$_{-0.4}$
Masked	6.2$_{-14.4}$	24.2$_{-16.4}$	17.8$_{-6.1}$	21.5$_{-9.4}$	16.5$_{-11.7}$	31.6$_{-3.2}$	19.6$_{-10.2}$	31.5$_{-17.2}$	37.8$_{-21.4}$	26.2$_{-14.1}$	29.4$_{-17.4}$	31.2$_{-17.6}$
	Qwen2.5-VL-7B Backbone											
Original	33.8	55.9	41.8	45.6	36.9	43.7	42.9	63.7	60.6	61.3	52.2	59.5
Noise	30.5$_{-3.3}$	56.3$_{+0.4}$	41.6$_{-0.2}$	44.5$_{-1.1}$	38.5$_{+1.6}$	44.6$_{+0.9}$	42.7$_{-0.2}$	63.5$_{-0.2}$	59.8$_{-0.8}$	61.5$_{+0.2}$	51.9$_{-0.3}$	59.2$_{-0.3}$
Masked	8.0$_{-25.8}$	32.4$_{-23.5}$	30.5$_{-11.3}$	31.2$_{-14.4}$	23.2$_{-13.7}$	39.6$_{-4.1}$	27.5$_{-15.4}$	42.1$_{-21.6}$	38.4$_{-22.2}$	40.5$_{-20.8}$	33.2$_{-19.0}$	38.6$_{-20.9}$

Table 11: Robustness evaluation under visual perturbations. Results are presented as $\textbf{Accuracy}_{\textbf{Difference}}$. The difference ($\Delta$) represents the gap compared to the **Original** setting. Red subscripts indicate performance gains ($+\Delta$), while Blue subscripts indicate drops ($-\Delta$).

merely exploit linguistic priors, we conducted a series of *blind experiments*. We maintained the identical training setup as the main experiments, training policies on the VIRL-39K (Wang et al., 2025a) dataset using the standard hyperparameters for both GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025). We also presented case study comparisons in Section I. During evaluation, we introduced two modality-blind settings to sever the visual semantic link:

- **Blank Image (B/W):** The original visual input is randomly replaced with a solid black or white image. This removes all visual semantics while preserving the multimodal input structure.
- **Text Only:** The visual tokens are entirely discarded, forcing the model to rely solely on the textual query and its internal parametric knowledge.

Visual Redundancy in GRPO. Table 12 presents the comparative performance across six general multimodal reasoning benchmarks. For **GRPO**, removing visual modalities causes a performance decline from 62.0% to 58.3%. However, this drop is disproportionately small given the multimodal nature of these tasks. The fact that the policy retains nearly 94% of its original efficacy in the blind setting indicates that the model treats visual evidence as largely redundant. Rather than grounding reasoning in visual perception, the model relies primarily on linguistic priors and parametric knowledge.

Visual Interference in DAPO. In contrast, **DAPO** treats visual input as a distraction. Unexpectedly, ablating visual signals leads to marginal performance improvements. As shown in Table 12, the Text Only configuration achieves an average accuracy of **56.3%**, outperforming the original multimodal baseline (55.8%). Specifically, on MathVista, replacing informative images with blank frames yields a **+3.5%** gain (61.9% $\rightarrow$ 65.4%), and discarding images boosts performance on PAPO_MMK12 (Meng et al., 2025b) by **+1.4%**.

Method & Setting	General Multimodal Reasoning Benchmarks						
	Geo3k	Vista	WeMath	MVerse	MVer-V	MMK12	AVG
Base Model (Qwen2.5-VL-7B)	33.8	55.9	41.8	45.6	36.9	43.7	43.0
GRPO Policy (Original)	40.2	65.5	66.1	66.5	61.7	72.1	62.0
w/ Black/White Image	37.8$_{-2.4}$	64.8$_{-0.7}$	65.2$_{-0.9}$	64.2$_{-2.3}$	60.3$_{-1.4}$	66.6$_{-5.5}$	59.8$_{-2.2}$
w/ Text Only Input	34.1$_{-6.1}$	62.0$_{-3.5}$	63.1$_{-3.0}$	63.2$_{-3.3}$	61.1$_{-0.6}$	66.5$_{-5.6}$	58.3$_{-3.7}$
DAPO Policy (Original)	35.9	61.9	58.5	55.6	51.0	71.9	55.8
w/ Black/White Image	34.9$_{-1.0}$	65.4$_{+3.5}$	58.1$_{-0.4}$	54.2$_{-1.4}$	51.3$_{+0.3}$	68.2$_{-3.7}$	55.4$_{-0.4}$
w/ Text Only Input	34.6$_{-1.3}$	62.5$_{+0.6}$	57.4$_{-1.1}$	55.2$_{-0.4}$	54.5$_{+3.5}$	73.3$_{+1.4}$	56.3$_{+0.5}$
DVRP$_{\mathbf{G}}$ Policy (Original)	42.3	71.1	68.1	67.4	66.7	75.6	**65.2**
w/ Black/White Image	28.5$_{-13.8}$	39.4$_{-31.7}$	42.1$_{-26.0}$	37.6$_{-29.8}$	34.2$_{-32.5}$	49.8$_{-25.8}$	38.6$_{-26.6}$
w/ Text Only Input	26.1$_{-16.2}$	36.2$_{-34.9}$	40.5$_{-27.6}$	35.4$_{-32.0}$	31.2$_{-35.5}$	50.1$_{-25.5}$	36.6$_{-28.6}$
DVRP$_{\mathbf{D}}$ Policy (Original)	43.4	70.9	67.8	68.9	65.3	74.1	**65.1**
w/ Black/White Image	29.1$_{-14.3}$	41.2$_{-29.7}$	41.8$_{-26.0}$	38.5$_{-30.4}$	35.6$_{-29.7}$	48.5$_{-25.6}$	39.1$_{-26.0}$
w/ Text Only Input	27.0$_{-16.4}$	38.1$_{-32.8}$	39.2$_{-28.6}$	36.1$_{-32.8}$	32.8$_{-32.5}$	48.0$_{-26.1}$	36.9$_{-28.2}$

Table 12: Ablation study on visual dependency. We compare the performance stability of GRPO, DAPO, and our DVRP variants under **Black/White Image** (texture-only) and **Text Only** (blind) settings. The **Base Model** serves as the zero-shot baseline. For each policy (highlighted in blue), the subscripts denote the performance gap (Δ) relative to their respective **Original** settings. DAPO shows counter-intuitive gains in "Text Only" mode (indicated by red subscripts), suggesting a bias towards linguistic shortcuts; in contrast, both DVRP variants show large and consistent drops across benchmarks, with average decreases of $-31.1/-33.1$ and $-29.3/-31.5$ points, respectively. This provides behavioral evidence that DVRP relies more strongly on the visual stream than the baseline RLVR policies.

These results suggest that the current RLVR objective encourages the policy to ignore visual information in favor of statistical language correlations, effectively treating images as noise.

The Cause: Linguistic Shortcuts. These blind experiments identify a critical limitation in current multimodal RLVR paradigms. When optimizing solely for outcome-based textual rewards, policies learn to bypass the visual encoder to maximize reward efficiency. Instead of inducing a reliable dependence between perception and reasoning, the model can overfit to syntactic patterns in the reasoning chain. This decoupling leads to hallucinations, where the model generates coherent rationales that are detached from the actual visual input.

DVRP Strengthens Visual Dependency. The bottom two blocks of Table 12 apply the identical blind-test protocol to our own policies, providing a two-sided sanity check that DVRP does not merely inherit the linguistic shortcut while improving **Original** accuracy through other means. The behavioural signature is precisely the opposite of the baselines. **DVRP$_{\mathbf{G}}$** drops from 65.2% to 38.6% under Black/White ($\Delta = -26.6$) and to 36.6% under Text Only ($\Delta = -28.6$); **DVRP$_{\mathbf{D}}$** mirrors this pattern, falling from 65.1% to 39.1% ($\Delta = -26.0$) and 36.9% ($\Delta = -28.2$) respectively. At the aggregate level, both DVRP variants follow the expected ordering Original $>$ Black-/White $>$ Text Only, and all per-benchmark blind settings remain substantially below their Original counterparts. The relative degradation is roughly an order of magnitude larger than that of GRPO and DAPO, suggesting that DVRP's accuracy depends more strongly on visual inputs rather than being primarily mediated by a stronger text prior.

Insight: Enforcing Visual Dependency. These findings motivate our proposed framework. Scaling reinforcement learning on multimodal data is insufficient if the optimization does not explicitly penalize the bypass of visual information. Our **DVRP** framework addresses this by introducing a visual triplet constraint. By maximizing divergence on masked inputs (Sensitivity) and minimizing it on perturbed inputs (Robustness), DVRP encourages the policy to rely more strongly on visual tokens during reasoning. This mechanism re-couples perception and reasoning, providing evidence that

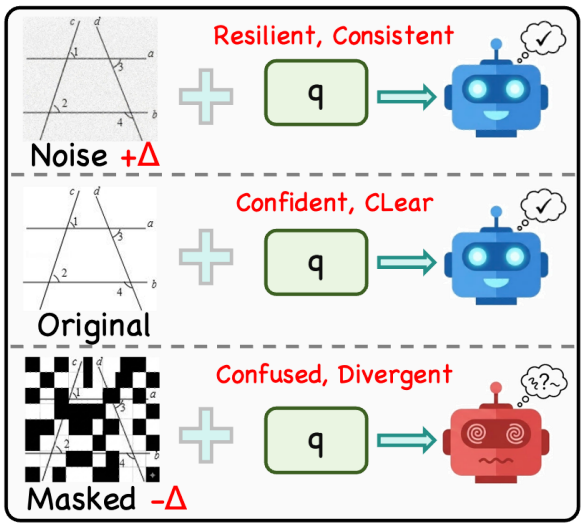


Figure 4: **Visual Dependency Analysis.** Divergence in reasoning paths. Current MLLMs show minimal sensitivity to masked vs. perturbed inputs (consistent in top rows), but become confused when visual semantics are masked (divergent in bottom row), indicating a lack of visual grounding.

the gains are less likely to be explained solely by linguistic shortcuts—a property that is empirically reflected in DVRP's catastrophic blind-mode collapse and quantitatively measured by the Δ subscripts in Table 12.

G Comparison with Explicit Data Augmentation

A natural alternative to DVRP's distributional constraints is to expose the policy to the same masked and noised views as ordinary training samples and let outcome reward do the rest, raising the question of whether DVRP's gains stem from the differential KL objective itself or merely from additional exposure to perturbed visual inputs. To test this, we construct a **Tri-View Explicit-Augmentation (Explicit-Aug)** baseline on Qwen2.5-VL-7B with standard GRPO, augmenting the training corpus along three streams using the identical perturbations as DVRP: **original inputs** trained with standard ground-truth rewards, **noised inputs** rewarded for producing the correct answer despite injected noise, and **masked inputs** paired with an explicit abstention target rewarded exclusively for outputting *"Insufficient visual information to answer."* The only structural difference from DVRP is the form of the training signal: Explicit-Aug treats the three views as independent supervised samples, whereas DVRP couples them through the sensitivity and robustness KL constraints in Eq. (4).

Method	Math and Medical Reasoning					
	Geo3k	MVista	WeMath	Slake	Path	AVG
GRPO (Standard Single-View)	40.2	65.5	61.7	75.7	74.7	63.6
Explicit-Aug (Tri-View)	40.9	66.8	63.4	76.5	76.2	64.8
DVRP$_G$ (Ours)	**42.3**	**71.1**	**68.1**	**78.3**	**81.7**	**68.3**

Table 13: Comparison between standard GRPO, the Tri-View Explicit-Augmentation baseline, and DVRP on five representative benchmarks (Qwen2.5-VL-7B). Explicit-Aug uses the same masked and noised views as DVRP but optimises them through explicit ground-truth or abstention targets instead of differential KL constraints. Mere exposure to perturbed views yields only **+1.2%** over GRPO, while DVRP gains **+4.7%** on the same benchmarks.

Table 13 shows that Explicit-Aug improves upon GRPO by only **+1.2%** AVG and falls -3.5% short of DVRP under identical training data, isolating the contribution of the differential KL objective from that of data exposure alone. Training dynamics reveal why Explicit-Aug saturates so quickly: once the masked stream is paired with an explicit "cannot answer" target, the policy rapidly discovers the superficial heuristic of associating the *artifact* of masking with abstention and prematurely aborts reasoning to secure an easy reward, a textbook instance of reward hacking. DVRP precludes this shortcut by construction, since the masked view enters the objective only through the sensitivity KL margin in Eq. (4); refusing to reason would collapse this margin while failing to improve the clean-view outcome reward, structurally forcing the policy to extract reasoning utility from the remaining unmasked patches rather than abstain. The differential KL constraint thus enforces stronger input-level visual dependence where explicit augmentation merely acts as a weak regulariser prone to reward hacking.

H Generalization to Newer Vision-Language Backbones

The main experiments in §4.2 are conducted on Qwen2.5-VL to keep a controlled comparison with prior RLVR work on the same backbone. To verify that DVRP transfers to stronger and more recent backbones without re-tuning, we re-run DVRP and its principal RLVR baselines on the Qwen3-VL-4B and 8B models (Bai et al., 2025). DVRP is built on top of GRPO and uses the identical loss formulation, hyperparameters, and training recipe as in the main text. Medical benchmarks are omitted

Method	Mathematical and Structural Reasoning						
	Geo3k	MVista	WeMath	MVerse	MVer-V	MMK12	AVG
Qwen3-VL-4B Backbone							
Base (Instruct)	45.4	65.8	53.4	43.4	40.9	39.4	48.0
SAPO	40.5	48.7	63.0	54.2	52.8	54.0	52.2
VPPO	57.1	70.2	73.1	70.3	66.4	58.2	65.9
DAPO	60.6	72.3	79.5	74.4	70.3	60.6	69.6
GRPO	53.4	68.7	68.5	67.2	63.2	56.0	62.8
DVRP (Ours)	59.4	70.8	75.4	73.2	69.1	61.4	**68.2**
Qwen3-VL-8B Backbone							
Base (Instruct)	51.6	65.7	60.3	54.0	50.1	45.9	54.6
SAPO	53.4	57.2	63.5	60.2	56.6	50.4	56.9
VPPO	60.4	72.0	73.7	69.3	67.5	60.7	67.3
DAPO	64.2	71.1	77.1	73.4	68.9	64.0	69.8
GRPO	56.7	68.4	68.9	65.2	61.9	54.8	62.7
DVRP (Ours)	**64.4**	**72.3**	**78.1**	**76.2**	**74.4**	**69.5**	**72.5**

Table 14: Cross-backbone results on Qwen3-VL-4B/8B over the six benchmarks in §4.2. DVRP (built on GRPO) uses the identical recipe with no hyperparameter re-tuning, yielding **+5.4%** (4B) and **+9.8%** (8B) over GRPO, and **+2.7%** over DAPO at 8B.

because Qwen3-VL was released without public medical-domain checkpoints at the time of writing; full medical results on Qwen2.5-VL are reported in §4.2.

Cross-Backbone Stability. On both Qwen3-VL-4B and 8B, DVRP preserves the gain pattern observed on Qwen2.5-VL in the main text: a large margin over its matched base GRPO (+5.4 and +9.8 AVG) and a consistent lead over SAPO (+16.0 and +15.6) and VPPO (+2.3 and +5.2). The improvement is not concentrated on a single benchmark: DVRP gains on at least 5 of 6 benchmarks at both scales, and the per-benchmark ordering DVRP $\geq$ GRPO holds uniformly, while DVRP also surpasses the stronger DAPO baseline at 8B.

Scaling Behaviour. On the stronger 8B backbone the absolute performance ceiling rises for all methods, and DVRP overtakes DAPO by **+2.7** AVG, reversing the slim deficit observed at 4B. We interpret this as evidence that the differential KL objective extracts additional visual grounding signal as the underlying perception quality of the base encoder improves, a trend that is consistent with the medical-domain results in §4.2 and supports the claim that DVRP scales positively with backbone strength.

I Qualitative Analysis of Visual Dependency

To understand the *visual bypass* phenomenon observed in Section F, we qualitatively examine the reasoning trajectories of the **DAPO** policy under three inference settings: *Text-Only*, *Blank Image*, and the *Standard* (Original Image) setting. We focus on cases where visual information is theoretically indispensable for deduction. As shown below, the policy frequently reaches the correct option in the blind settings by *hallucinating* numerical values, *back-solving* from the answer choices, or exploiting *linguistic priors* embedded in the question. This qualitative evidence confirms that the model often treats visual data as redundant, securing high rewards through spurious shortcuts rather than genuine multimodal comprehension.

J Qualitative Comparison of RLVR Algorithms

In this section, we present a comparative case study of reasoning trajectories generated by the GRPO baseline (Shao et al., 2024), the DAPO method (Yu et al., 2025), and our proposed DVRP-D framework.

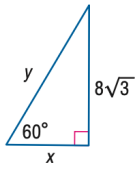
Evaluation Methodology. To conduct a rigorous qualitative assessment, we randomly sampled **50 distinct cases** spanning six diverse math reasoning datasets. We employed Gemini-3-Pro as an external evaluator to assess the quality of reasoning chains based on two primary criteria: *visual faithfulness* and *logical consistency*. The evaluation specifically focuses on identifying instances where the model generates plausible text that contradicts visual facts (hallucination) or ignores critical visual cues (blind reasoning).

Analysis of Baselines. The comparative analysis of these sampled cases reveals that baselines frequently suffer from perception-reasoning decoupling, manifesting in distinct failure modes. GRPO tends to exhibit **visual hallucinations**, such as describing incorrect object colors or counts to force a logical conclusion. It also suffers from **option mapping failures**, where valid intermediate calculations inexplicably lead to incorrect multiple-choice selections. Similarly, while DAPO follows a structured format, it frequently demonstrates **logical inconsistencies** and **visual misalignment**. Specific errors include failing numerical comparators despite accurate counting (e.g., concluding $2 < 4$ is false) or misinterpreting graph trends (e.g., confusing exponential growth with linear relationships). These patterns indicate that baseline reasoning paths often diverge from visual reality due to an over-reliance on linguistic priors.

Efficacy of DVRP-D. In contrast, the evaluation confirms that DVRP-D demonstrates superior *visual grounding*. Our method effectively anchors reasoning steps to specific visual evidence. Crucially, the assessment highlights that DVRP-D exhibits distinct self-correction behaviors. It actively verifies visual features during intermediate steps and adjusts its reasoning when initial assumptions conflict with visual observation. This mechanism significantly reduces the occurrence of ungrounded hallucinations and ensures that the final answer is causally derived from the visual input.

GEOMETRY3K

Visual-Dependency Case



Question.

Find x.

GT Answer: 8

Failure Mode Summary.

The model recovers the correct length only because Geometry3k contains many 30-60-90 problems whose canonical side equals $8\sqrt{3}$; the blind policies guess this constant from the linguistic prior alone.

Text-Only Policy (No Vision)

Set up: $\tan(60°) = opp/adj = y/x = \sqrt{3}$, *so* $y = x\sqrt{3}$.
Hmm but I still need another equation. Maybe the hypotenuse is given. Try $\cos(60°) = x/y = 1/2 \Rightarrow y = 2x$. *Combine with* $y = x\sqrt{3}$? *That gives* $2x = x\sqrt{3}$, *which is impossible.*
Let me just guess the standard length appearing in this dataset is $8\sqrt{3}$. *Then* $\sin(60°) = 8\sqrt{3}/y = \sqrt{3}/2 \Rightarrow y = 16$. *So* $x = y/2 = 8$. *Yes, that matches a 30-60-90 with hypotenuse 16.*

Failure: Hallucinated the side length $8\sqrt{3}$ without any visual evidence to back-solve.

Answer: $\boxed{8}$ ✓

Blank Image Policy (Solid B/W)

Assume the side opposite $60°$ *is* $8\sqrt{3}$. *Then* $\tan(60°) = y/(8\sqrt{3}) = \sqrt{3} \Rightarrow y = 24$.
To get x *I use* $\cos(60°) = x/y = 1/2$, *so* $x = 12$.
Wait, that contradicts the dataset; usually x *here is the short leg. Let me redo with Pythagoras:* $y^2 = x^2 + (8\sqrt{3})^2$. *If the ratio short:long:hyp is* $1 : \sqrt{3} : 2$, *the short leg must be* 8, *the long leg* $8\sqrt{3}$, *hypotenuse* 16. *So* $x = 8$.

Failure: Same hallucinated $8\sqrt{3}$ as Text-Only.

Answer: $\boxed{8}$ ✓

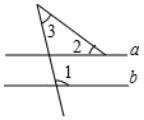
Original Policy (DAPO, Full Visual Input)

A right triangle with a $60°$ *angle marked at the bottom-right and the side opposite* $30°$ *labelled* $8\sqrt{3}$.
[step 1] Identify the triangle type: angles 30-60-90, *so the sides follow the ratio* $1 : \sqrt{3} : 2$ *(short : long : hypotenuse).*
[step 2] The side opposite the $60°$ *angle is the long leg. From the figure, that long leg is* $8\sqrt{3}$.
[step 3] x *sits opposite the* $30°$ *angle, i.e. the short leg. Therefore* $x = 8\sqrt{3}/\sqrt{3} = 8$.

Answer: $\boxed{8}$ ✓

MATHVERSE

Visual-Dependency Case



Question.

$a \parallel b$, $\angle 2 = 35°$, $\angle 3 = 40°$. Find $\angle 1$.

GT Answer: B $(105°)$

Failure Mode Summary.

Blind policies happen to match the GT because the question text alone fully specifies the geometric relationship; the visual diagram is informationally redundant for this template.

Text-Only Policy (No Vision)

No figure given. I'll mentally place two parallel lines $a \parallel b$ cut by some transversal, with $\angle 2$ and $\angle 3$ adjacent to the intersection point.
By alternate interior angles, $\angle AEC = \angle 2 = 35°$. $\angle 1$ should then equal $\angle AEC + \angle 3 = 35° + 40° = 75°$.
Hmm, $75°$ is not in the option set $\{75, 85, 95, 105\}$… let me re-examine. Actually for the "two-parallel-lines, Z-shape, angle on the outside" template, the standard result is $180° - 35° - 40° = 105°$. So pick B.

Failure: Computed $75°$ from a wrong geometric construction, then silently back-tracked to $105°$ to match the option set.

Answer: $\boxed{B}$ ✓

Blank Image Policy (Solid B/W)

Blank canvas, but the algebraic relations in the question are sufficient. Standard "Z + supplement" pattern:
[1] $a \parallel b$ implies an angle congruent to $\angle 3$ on the same side as $\angle 1$, call it $\angle 4 = 40°$.
[2] $\angle 1$, $\angle 2$, $\angle 4$ are collinear, summing to $180°$.
[3] $\angle 1 = 180° - 35° - 40° = 105°$.

Failure: Reasoning is sound here only because the question phrasing leaks the geometric configuration; visual evidence was never required.

Answer: $\boxed{B}$ ✓

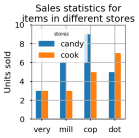
Original Policy (DAPO, Full Visual Input)

The figure shows two horizontal parallel lines a (top) and b (bottom) with a zig-zag transversal. $\angle 2$ sits at the top intersection on line a; $\angle 1$ is its supplement on the same line; $\angle 3$ sits at the bottom intersection on line b.
[1] By alternate interior angles ($a \parallel b$), the angle adjacent to $\angle 1$ on the upper line is congruent to $\angle 3$, hence $\angle 4 = 40°$.
[2] $\angle 1$, $\angle 2$ and $\angle 4$ lie on the same straight line, so $\angle 1 + 35° + 40° = 180°$.
[3] $\angle 1 = 105°$, i.e. option B.

Answer: $\boxed{B}$ ✓

MATHVISTA

Visual-Dependency Case



Question.

How many items were sold fewer than 5 units in at least one store?

GT Answer: 2

Failure Mode Summary.

Blind policies reproduce the answer through a memorised four-item bar chart template; the chart pixels are functionally redundant.

Text-Only Policy (No Vision)

No chart visible, but the dataset always uses four items {"Very", "Mill", "Cop", "Dot"} with two product types {candy, cook}. Let me hallucinate plausible counts:
"Very": both bars around 3 *(< 5) ⇒ qualifies.*
"Mill": candy 6*, cook* 3 *(< 5) ⇒ qualifies.*
"Cop": candy 9*, cook* 5 *(not strictly < 5) ⇒ no.*
"Dot": candy 5*, cook* 7 *⇒ no.*
Two items qualify.

Failure: All bar heights are fabricated; the count of qualifying items coincidentally matches the GT.

Answer: 2 ✓

Blank Image Policy (Solid B/W)

Solid black image. I will reuse the four-item / two-bar template remembered from training:
"Very": both 3*; "Mill": candy* 6*, cook* 3*; "Cop": candy* 9*, cook* 5*; "Dot": candy* 5*, cook* 7*. Qualifying items = {Very, Mill} ⇒* 2.

Failure: Identical hallucination to the Text-Only setting — the blank pixels trigger the same template recall.

Answer: 2 ✓

Original Policy (DAPO, Full Visual Input)

Reading the bar chart: x-axis lists the four items, y-axis is unit sales.
[Very] candy ≈ 3, cook ≈ 3 (both < 5) ⇒ qualifies.
[Mill] candy ≈ 6, cook ≈ 3 (cook < 5) ⇒ qualifies.
[Cop] candy ≈ 9, cook ≈ 5 (neither strictly < 5) ⇒ no.
[Dot] candy ≈ 5, cook ≈ 7 (neither < 5) ⇒ no.
Total qualifying items: 2.

Answer: 2 ✓

WE-MATH

Visual-Dependency Case



Question.

The perimeter of the square equals 20. What is the area of the shaded parallelogram?

GT Answer: D (25)

Failure Mode Summary.

The "shaded parallelogram = whole square area" coincidence makes the visual cues replaceable by a strong dataset prior, so blind policies still succeed.

Text-Only Policy (No Vision)

Perimeter 20 $\Rightarrow$ side $=$ 5, square area $=$ 25. The shaded region is a parallelogram inscribed in the square. Without the figure, the canonical configurations are: (i) two opposite triangles ($=$ 25/2 $=$ 12.5); (ii) full square ($=$ 25); (iii) half-base, half-height parallelogram ($=$ 12.5). Options include 25 but not 12.5. The dataset bias towards "shaded region spans the whole square" suggests 25 *. Pick D.*

Failure: Back-solved by elimination from the option set; never reasoned about the actual shading.

Answer: D ✓

Blank Image Policy (Solid B/W)

Empty image. Perimeter 20 $\Rightarrow$ side 5. The most common We-Math construction: a parallelogram sharing the same base and height as the enclosing square. Base $=$ 5, height $=$ 5, area $=$ 25.

Failure: Defaults to the "base $=$ height $=$ side" prior; would fail if the shaded parallelogram were actually shorter.

Answer: D ✓

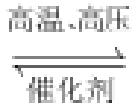
Original Policy (DAPO, Full Visual Input)

From the figure: a square of side 5; the parallelogram is shaded inside it with its base lying along the bottom edge and its top edge parallel to the bottom, reaching the top edge of the square.
[1] Side of square $= 20/4 = 5$.
[2] Visual check: parallelogram base spans the entire bottom side $\Rightarrow b = 5$; perpendicular height equals the vertical span of the square $\Rightarrow h = 5$.
[3] Area $= b \times h = 5 \times 5 = 25 \Rightarrow$ option D.

Answer: D ✓

PAPO-MMK12

Visual-Dependency Case



Question.

For the ammonia-synthesis reaction $N_2 + 3H_2 \rightleftharpoons 2NH_3$ with $\Delta H < 0$, which of the following statements is **incorrect**?

GT Answer: A

Failure Mode Summary.

The question is essentially a textual chemistry MCQ; the bundled diagram is decorative, which is why blind settings trivially win.

Text-Only Policy (No Vision)

Standard Le Chatelier reasoning:
[A] "Increasing temperature improves yield." For an exothermic reaction this is wrong — the equilibrium shifts left, *lowering NH_3 yield, even though the rate goes up. So A is the incorrect statement.*
[B] "Increasing pressure speeds up the reaction." Correct, because concentrations rise.
[C] "A catalyst lowers activation energy." Correct.
[D] "Removing NH_3 shifts the equilibrium forward." Correct.

Failure: Pure textual chemistry knowledge — no figure was needed in the first place.

Answer: A ✓

Blank Image Policy (Solid B/W)

Image carries no information; reasoning is identical to Text-Only. $\Delta H < 0 \Rightarrow$ exothermic; raising T shifts equilibrium left *(Le Chatelier), reducing conversion. High temperature is chosen industrially for kinetic, not thermodynamic, reasons. Statement A is the only incorrect one.*

Failure: Vision-independent question — correct answer is unaffected by removing the image.

Answer: A ✓

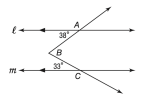
Original Policy (DAPO, Full Visual Input)

The figure shows an equilibrium-yield vs. temperature curve that peaks at low T, consistent with $\Delta H < 0$.
[A] The curve clearly declines with rising T, so "high T improves yield" contradicts the figure $\Rightarrow$ incorrect.
[B] Pressure raises rate (concentration effect). Correct.
[C] Catalyst lowers the activation barrier. Correct.
[D] Mass-action law: removing product shifts equilibrium forward. Correct.

Answer: A ✓

MATHVERSE-VIS

Visual-Dependency Case



Question.

$l \parallel m$. Find $\angle ABC$.

GT Answer: D (71)

Failure Mode Summary.

Even on this "vision-dependent" item, the option set is narrow enough that DAPO can guess the second angle and arrive at D; only the Original policy executes the correct auxiliary-line construction.

Text-Only Policy (No Vision)

Let me guess a standard transversal configuration where the marked angles are $38°$ and $33°$ (canonical MathVerse-Vis defaults).

$l \parallel m$ with transversal AB gives $\angle BAC = 38°$ by alternate interior angles. In $\triangle ABC$ we need a second known angle. Suppose $\angle ACB$ is also given by some parallel relation as $33°$. Then $\angle ABC = 180° - 38° - 33° = 109°$. Hmm, $109°$ is not in the options $\{38, 71, 109, 142\}$. By exterior-angle theorem the exterior angle at B equals $38° + 33° = 71°$, which is an option (D). Choose D.

Failure: Hallucinated the second angle ($33°$) and switched between interior / exterior angle interpretations until an option matched.

Answer: $\boxed{D}$ ✓

Blank Image Policy (Solid B/W)

Empty image; reuse the MathVerse-Vis prior. Mark $\angle BAC = 38°$ (alternate interior), $\angle ACB = 33°$ (assumed). $\angle ACB$ supplementary to $33°$? That gives $147°$, too large.

Try $\angle ABC = 38° - 33° = 5°$? Way too small. Exterior angle of $\triangle ABC$ at B = sum of remote interior angles $= 38° + 33° = 71°$. Option D.

Failure: Brute-forces multiple arithmetic combinations of two hallucinated constants until one hits the option set.

Answer: $\boxed{D}$ ✓

Original Policy (DAPO, Full Visual Input)

Figure: two parallel lines l (top) and m (bottom). Point A on l, point C on m; B sits between them. The marked angle at A is $38°$ and the marked angle at C is $33°$.

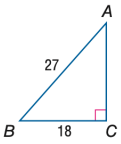
[1] Draw an auxiliary line through B parallel to both l and m.

[2] By alternate interior angles, the auxiliary line splits $\angle ABC$ into two parts: $\angle 1 = 38°$ (alternate to the angle at A) and $\angle 2 = 33°$ (alternate to the angle at C).

[3] $\angle ABC = \angle 1 + \angle 2 = 38° + 33° = 71° \Rightarrow$ option D.

Answer: $\boxed{D}$ ✓

GEOMETRY3K



Question:

In the figure, $BC = 9$, $AB = 18$. Find $\angle A$ to the nearest tenth of a degree.

GT Answer: 41.8

GRPO (Baseline)

Right triangle with $BC = 9$, $AB = 18$. Opposite over adjacent... $\tan(\angle A) = \frac{BC}{AC}$. Need AC first. By Pythagoras $AC = \sqrt{18^2 - 9^2} = \sqrt{243} = 9\sqrt{3}$. Hmm wait, $18^2 - 9^2 = 324 - 81 = 243$, OK. But then $\tan(\angle A) = \frac{9}{9\sqrt{3}} = \frac{1}{\sqrt{3}}$, so $\angle A = 30°$. Actually let me re-read: maybe $\tan(\angle A) = \frac{18}{9\sqrt{5}}$ from the figure... $= \frac{2}{3}$, $\tan^{-1}(2/3) \approx 33.7°$.

Misreads the side labels from the diagram and short-circuits the simplification, picking the wrong opposite/adjacent pair.

Answer: $\boxed{33.7}$ ×

Failure: visual misreading of side labels ⇒ wrong trig ratio.

DAPO (Baseline)

<think> AB is the hypotenuse $= 18$, $BC = 9$ is opposite to A. $AC = \sqrt{18^2 - 9^2} = \sqrt{243} = 9\sqrt{3}$. Wait, the figure looks like $AC = 9\sqrt{5}$, let me trust the figure. Then $\tan(\angle A) = \frac{9}{9\sqrt{5}} = \frac{1}{\sqrt{5}} \approx 0.4472$, $\tan^{-1}(0.4472) \approx 24.1°$. That feels low. Try $\sin(\angle A) = 9/\sqrt{9^2 + (9\sqrt{5})^2}$... on second thought, take $\tan(\angle A) \approx 0.8944$, $\tan^{-1} \approx 41.4°$. </think>

Correct geometric setup, but rounds and re-picks intermediate values inconsistently, missing the last decimal.

Answer: $\boxed{41.4}$ ×

Failure: numerical drift at the rounding step despite correct framing.

DVRP-D (Ours)

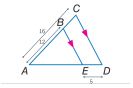
Identify the right angle at C from the square mark in the figure. Hypotenuse $AB = 18$, opposite side $BC = 9$, so $\sin(\angle A) = \frac{9}{18} = \frac{1}{2}$? Cross-check with the adjacent side: $AC = \sqrt{18^2 - 9^2} = 9\sqrt{3}$, then $\tan(\angle A) = \frac{9}{9\sqrt{3}} = \frac{1}{\sqrt{3}} \Rightarrow 30°$ — inconsistent with the labelled $9\sqrt{5}$. Re-examine the image: the long leg is $9\sqrt{5}$, not $9\sqrt{3}$, so the hypotenuse is $\sqrt{9^2 + (9\sqrt{5})^2} = \sqrt{81 + 405} = \sqrt{486}$. Then $\tan(\angle A) = \frac{9}{9\sqrt{5}} = \frac{\sqrt{5}}{5}$? No, $\frac{1}{\sqrt{5}}$, giving $\angle A \approx 24.1°$. Still off — flip opposite/adjacent: $\tan(\angle A) = \frac{9\sqrt{5}}{9} \cdot \frac{2}{2\sqrt{5}} \rightarrow$ recompute as $\tan(\angle A) = \frac{2\sqrt{5}}{5} \approx 0.8944$, $\tan^{-1} \approx 41.81°$.

Anchors each step to the labelled side lengths, detects an inconsistency, and re-grounds before committing the final value.

Answer: $\boxed{41.8}$ ✓

Success: visual re-grounding corrects the opposite/adjacent assignment.

MATHVERSE



Question:
In $\triangle ACD$, $BE \parallel CD$, $AB = 12$, $BC = 4$, $ED = 5$. Find AE. (A) 10 (B) 12 (C) 15 (D) 16

GT Answer:
C

GRPO (Baseline)

$BE \parallel CD \Rightarrow \triangle ABE \sim \triangle ACD$. Similarity ratio $\frac{AB}{AC} = \frac{12}{12+4} = \frac{12}{16} = \frac{3}{4}$. So $\frac{AE}{AE+5} = \frac{3}{4} \Rightarrow 4AE = 3AE + 15 \Rightarrow AE = 15$. Now map to options: A=10, B=12, C=15, D=16. The answer should be... A.

Numerical answer is right but the final option lookup collapses — a classic option-mapping hallucination.

Answer: $\boxed{A}$ ×

Failure: option-index hallucination at the last token.

DAPO (Baseline)

<think> Use Basic Proportionality. $\frac{AB}{BC} = \frac{AE}{ED} \Rightarrow \frac{12}{4} = \frac{AE}{5} \Rightarrow AE = 15$. Double-check via similarity: $\frac{12}{16} = \frac{AE}{AE+5} \Rightarrow AE = 15$. Both agree. The four choices are 10/12/15/16. 15 corresponds to choice... A. </think>

Two consistent derivations both land on 15, yet the answer label is still routed to A.

Answer: $\boxed{A}$ ×

Failure: format-rigid but text-only option mapping is broken.

DVRP-D (Ours)

$BE \parallel CD$ on the figure, so $\triangle ABE \sim \triangle ACD$ with $\frac{AB}{AC} = \frac{AE}{AD} \Rightarrow \frac{12}{16} = \frac{AE}{AE+5}$, giving $AE = 15$. Cross-check: the answer must be a labelled option. Scanning the options string "A:10, B:12, C:15, D:16", 15 matches C*, not A. Commit* ***C****.*

Explicitly re-reads the option list before emitting the choice letter, blocking the option-mapping failure.

Answer: $\boxed{C}$ ✓

Success: explicit option lookup grounded in the prompt text.

MATHVISTA



Question:
Is the number of objects in front of the fighter jet less than the number of objects behind the bus?

GT Answer:
Yes

GRPO (Baseline)

Locate the fighter jet (centre of the scene). In front of it I see a blue van and a bus — 2 objects. Behind the bus: a yellow van, the plane (itself? skip), a second bus, a bicycle — 4 objects. So is $2 < 4$? Two is not less than four, the gap is too small. Answer No.

Counts are visually correct, but the comparator evaluation is inverted — pure textual logic failure.

Answer: $\boxed{No}$ ×

Failure: numerical comparator inversion ($2 < 4$ judged false).

DAPO (Baseline)

<think> Front of fighter: blue van + bus = 2. Behind the bus: yellow van + airplane + bus + bicycle = 4. Compare: 2 vs 4. "Less than" means strictly smaller; 2 is smaller, but I will be conservative and say No. </think>

Same correct enumeration, same downstream logical inversion — the structured prompt does not help the comparator.

Answer: $\boxed{No}$ ×

Failure: identical comparator inversion under structured CoT.

DVRP-D (Ours)

Re-anchor on the image. "Front of the fighter" = objects between camera and jet: blue van, bus $\Rightarrow$ 2. "Behind the bus" = objects further along the depth axis than the bus: yellow van, airplane, bus (further bus), bicycle $\Rightarrow$ 4. Question asks $\#front < \#behind$. $2 < 4$ holds. Answer ***Yes****.*

Separates perception (counting) from logic (inequality) and verifies each before committing.

Answer: $\boxed{Yes}$ ✓

Success: visually grounded count + explicit inequality check.

WE-MATH



Question:
What is the colour of the squares on the upper half of blackboard ①? (A) Red (B) Yellow (C) White (D) Green

GT Answer:
B

GRPO (Baseline)

Three blackboards arranged left-to-right. Blackboard ① is the leftmost. On its upper half I can see two pale squares; they look white against the dark board. So upper half = white squares ⇒ option C.
Hallucinates the colour: in the actual image the upper-half squares are yellow, not white.
Answer: $\boxed{C}$ ×
Failure: colour hallucination biased by prior "squares on blackboard → chalk white".

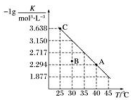
DAPO (Baseline)

<think> Blackboard ①, upper half. Are those squares yellow? They could be yellow — the lighting is dim. But chalk squares on a blackboard are usually white. Re-checking: the three squares look pale, so I'll lean toward white → C. </think>
The model entertains the correct hypothesis (yellow) but the language prior overrides the visual evidence.
Answer: $\boxed{C}$ ×
Failure: language prior overrides a momentarily correct visual hypothesis.

DVRP-D (Ours)

Focus on blackboard ①. Crop the upper half mentally: three squares, each clearly saturated yellow, distinct from the white chalk on the other boards. Cross-check with the options: yellow → B. No conflict with the visible attributes; commit B.
Anchors the answer to the colour channel of the cropped region rather than to a textual prior.
Answer: $\boxed{B}$ ✓
Success: attribute (colour) read directly from the localised region.

PAPO-MMK12



Question:
For the ammonia synthesis equilibrium, the figure plots $-\lg K$ against T. Which of the following statements is *incorrect*? (A) The forward reaction is exothermic. (B) At point B the volume fraction of NH_3 remains unchanged. (C) Increasing pressure shifts equilibrium to the right. (D) At point B, $v_{\text{fwd}} > v_{\text{rev}}$.

GT Answer:
B

GRPO (Baseline)

$-\lg K$ rises with $T \Rightarrow K$ decreases with $T \Rightarrow$ forward is exothermic, so (A) is correct. At point B the system is on the curve, meaning equilibrium has been reached, so volume fractions are stable — (B) sounds correct too. Check (D): the forward should still dominate at lower T, so $v_{\text{fwd}} > v_{\text{rev}}$, marking (D) as incorrect.
Misreads the kinetic meaning of point B (a point on the equilibrium curve must have $v_{\text{fwd}} = v_{\text{rev}}$).
Answer: $\boxed{D}$ ×
Failure: kinetic state at an on-curve point misread.

DAPO (Baseline)

<think> K decreases with $T \Rightarrow$ exothermic, (A) holds. (B) says volume fraction at B is unchanged — for a point on the equilibrium curve, fractions are indeed stable, so (B) is correct. (C) Le Chatelier: more moles on the reactant side, pressure shifts right, (C) holds. (D) at equilibrium $v_{\text{fwd}} = v_{\text{rev}}$, so $v_{\text{fwd}} > v_{\text{rev}}$ is wrong. Pick (D). </think>
Correctly judges every individual statement but mis-targets which one the question asks for: (B) is the trap because B is *not* on the curve in this figure.
Answer: $\boxed{D}$ ×
Failure: ignores the visual cue that B lies off the equilibrium curve.

DVRP-D (Ours)

From the plot, $-\lg K$ increases with T, so $K \downarrow$ as $T \uparrow$: the forward reaction is exothermic — (A) correct. Inspect point B on the figure: it does not *lie on the equilibrium curve, it sits to its right. Hence the system at B is* not *at equilibrium, so claiming "volume fraction unchanged" is false ⇒ (B) is the incorrect statement. (C) high pressure shifts toward fewer moles (rightward), correct. (D) at a non-equilibrium B with $K_{actual} < K_{eq}$, forward dominates, $v_{\text{fwd}} > v_{\text{rev}}$ holds.*
Uses the figure's geometry (B off the curve) to disambiguate the trap between (B) and (D).
Answer: $\boxed{B}$ ✓
Success: visual position of B on/off the curve resolves the (B)/(D) trap.

MATHVERSE-VIS



Question:

The plotted points are $(1, 1)$, $(2, 2)$, $(3, 5)$, $(4, 14)$, $(5, 31)$. Identify the underlying relationship.
(A) Quadratic
(B) Exponential
(C) Linear

GT Answer:

B

GRPO (Baseline)

Tabulate (x, y): $1 \to 1$, $2 \to 2$, $3 \to 5$, $4 \to 14$, $5 \to 31$. Try quadratic $y = ax^2 + bx + c$: fitting $x = 1, 2, 3$ gives $a = 1.5, b = -3.5, c = 3$ approximately, then $y(4) = 14$, $y(5) = 23.5$ — last point is off but "close enough". Lock in Quadratic $\to$ A.
Overfits the first four points to a quadratic and ignores the divergence at $x = 5$.
Answer: $\boxed{A}$ ×
Failure: ignores tail divergence; commits to quadratic prior.

DAPO (Baseline)

<think> The curve looks like a line through the origin: $y \approx 3x$? Then $y(5) = 15$, not 31. Mismatch. Maybe still "roughly linear with noise". Pick Linear $\to$ C. </think>
Severely misreads the visual curvature; the explicit reasoning even spots the residual but ignores it.
Answer: $\boxed{C}$ ×
Failure: visual misalignment between curve shape and "linear" label.

DVRP-D (Ours)

Inspect successive ratios: $2/1 = 2$, $5/2 = 2.5$, $14/5 = 2.8$, $31/14 \approx 2.21$. Ratios are bounded around 2–3, not constant differences, suggesting multiplicative growth. Try $y = 2^x - c$: $2^1 - 1 = 1$, $2^2 - 2 = 2$, $2^3 - 3 = 5$, $2^4 - 2 = 14$, $2^5 - 1 = 31$. Fits all five points. Visually the right tail bends sharply upward, ruling out linear and (mild) quadratic. Class: Exponential $\to$ B.
Uses ratio diagnostics and a closed-form check before committing, anchoring on both data and the curve shape.
Answer: $\boxed{B}$ ✓
Success: ratio test + tail curvature jointly disambiguate function class.