SIVA-RL: SENSITIVITY–INVARIANCE VISUAL ALIGNMENT FOR MULTIMODAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) optimizes vision-language models for task-level correctness, but correct answers do not guarantee visual grounding. Existing visual-intervention methods contrast policy behavior on clean and modified images, yet assign supervision by intervention type rather than observed effect. This embeds an assumption we show does not hold. In a paired audit, 36.3% of clean-correct examples are outcome-drop under one operator and outcome-stable under another. That rate is $6.1\times$ the disagreement any single per-sample dependence level admits. The required direction belongs to the sample–intervention pair and must be measured, not inferred from operator choice. We propose **SIVA-RL**, a **S**ensitivity–**I**nvariance **V**isual **A**lignment framework that adds sample-wise, outcome-conditioned supervision through a two-stage training recipe. A frozen audit policy converts the per-sample reward drop into soft routing weights: large-drop pairs receive a margin-bounded sensitivity objective, valid low-drop pairs a bounded consistency objective, and ambiguous pairs are down-weighted. Across seven matched backbone–scale configurations, SIVA-RL raises the nine-benchmark average by 1.78 to 9.50 points. These configurations span Qwen2.5-VL at 3B and 7B, Qwen3-VL Thinking at 2B and 8B, and Qwen3.5 at 4B. Gains are larger on the vision-dependent group in five of seven comparisons, with vision-dependent gains of up to 9.49 points.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves reasoning in vision-language models by optimizing task-level correctness, without densely annotated reasoning chains Shao et al. (2024); Guo et al. (2025); Huang et al. (2025b); Yao et al. (2025); Meng et al. (2025). A final-answer reward, however, indicates whether a prediction is correct, not how it was reached: a policy can accumulate high reward while relying on language priors, answer-frequency biases, or memorized templates instead of the visual evidence the task requires Goyal et al. (2017); Agrawal et al. (2018); Niu et al. (2021); Xia et al. (2025). The shortcut is especially consequential in counting, geometry, and diagram understanding, where ignoring the image produces systematically wrong answers.

To counteract it, recent multimodal RL methods use visual interventions as a training signal: PAPO Wang et al. (2025b) contrasts policy behavior on clean and corrupted images, and others assign visual credit at the token or trajectory level or inject distorted views to broaden exploration Xiao et al. (2025); Huang et al. (2025a); Wang et al. (2026c); Li et al. (2026); Liu et al. (2026); Li et al. (2025); Zhang et al. (2025b); Gao et al. (2026). A recurring design principle ties the intervention type to a fixed optimization target. Destructive interventions prescribe separation. Mild or structure-preserving ones prescribe consistency. The principle assumes that a sample's optimization target can be *inferred from the intervention's design* rather than measured from its effect on that sample. This assumption does not hold.

The dependence label is not a property of the sample alone; it is specific to the sample–intervention pair. We establish this through a paired outcome audit of 1,800 examples across nine benchmarks, using a fixed GRPO-trained 3B policy and three operators: black image, random mask, and distance-constrained PatchSwap (protocol in Appendix A). Among the 906 examples answered correctly on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

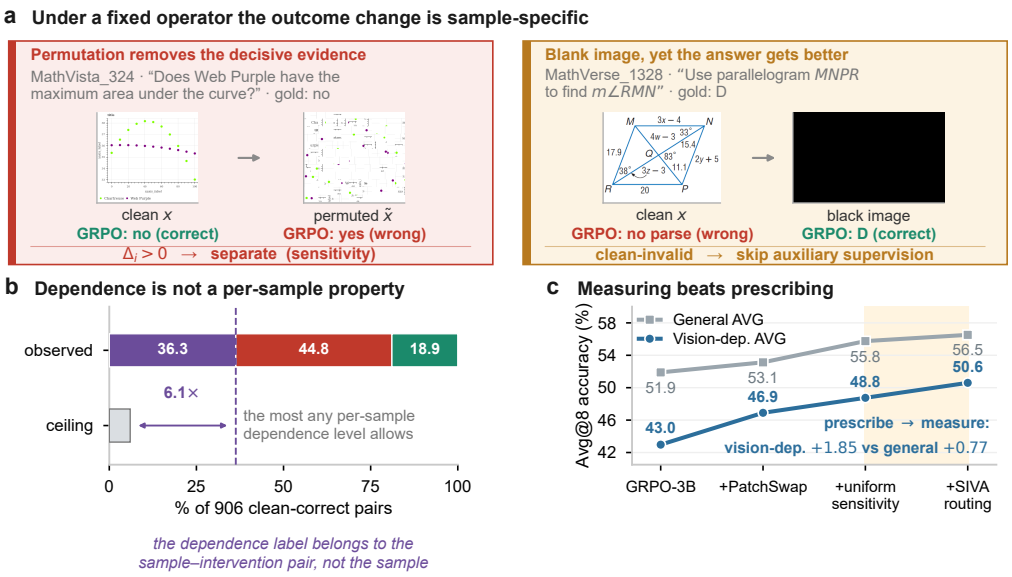


Figure 1: **The required supervision direction is specific to the sample–intervention pair. (a)** Two audit pairs (Appendix F): PatchSwap removes the decisive chart evidence and the answer flips ($\Delta_i > 0$, separate). The clean-validity gate excludes the right-hand pair because its clean prediction is invalid. **(b)** Joint agreement of the three operators on the 906 clean-correct examples: 44.8% drop under all three (red), 18.9% survive all three (teal), 36.3% disagree (purple). A single per-sample dependence level probed at three severities would nest the drop sets, capping disagreement at 6.0% (Appendix A). **(c)** Final step of the GRPO-3B ablation chain of Table 4: completing the routed objective adds $+1.85$ points on vision-dependent benchmarks against $+0.77$ on general ones.

their clean images, the operators individually break the outcome on 67.8%, 62.9%, and 61.8%. Their joint pattern is the more restrictive finding: only 44.8% of pairs lose the outcome under all three and 18.9% survive all three, while 36.3% disagree. The same example is outcome-drop under one operator and outcome-stable under another. If dependence were a per-sample quantity that the operators merely probe at different severities, their drop sets would be nested and disagreement could not exceed the spread of the marginals, $67.8 - 61.8 = 6.0\%$; the observed rate is $6.1\times$ that ceiling (Figure 1b). Among the 709 examples with parseable answers under all three views, the disagreement rate is 37.9%. Of the 614 examples a blank image breaks, the policy cannot answer without the image. Of those examples, 21.8% are untouched by PatchSwap. Across the broader audit, interventions can also improve failed clean predictions (Figure 1a, right). Operator identity alone does not determine the behavioral effect.

This leads to the core principle of the work: reliable supervision under visual interventions requires *measuring*, not prescribing, the effect of each intervention on each sample. Because the label does not transfer across operators, the measurement cannot be amortized into a reusable per-sample score. It has to be taken on the intervention the supervision is applied to. We use the observed reward drop to route each pair: large-drop pairs receive a margin-bounded sensitivity objective, valid low-drop pairs a consistency objective, and intermediate pairs attenuated weights. A clean-correctness gate anchors both branches to examples answered correctly on the clean image.

We introduce **SIVA-RL**, a **S**ensitivity–**I**nvariance **V**isual **A**lignment framework with three components: *localized intervention construction* via token-aligned within-image PatchSwap; *sample-wise outcome valuation* by a frozen audit policy that converts the reward drop into soft routing weights; and *dual-direction clean-anchored alignment*, which re-scores existing clean rollout responses under both views, enforcing a minimum policy divergence for high-drop pairs and a ceiling for valid low-drop ones. Our main contributions are:

- **Visual dependence is a property of the sample–intervention pair.** A paired outcome audit of 1,800 examples across nine benchmarks (Appendix A) shows that 36.3% of clean-correct examples are outcome-drop under one operator and outcome-stable under another,

$6.1\times$ the disagreement permitted by a single per-sample dependence level. Across the broader audit, interventions can also improve failed clean predictions. Intervention identity alone does not determine the behavioral effect. Operator-conditioned assignment therefore prescribes a separation target for a large minority of pairs whose outcome does not move.

- **Outcome-conditioned selective visual alignment.** We formulate visual-intervention-based multimodal RL as a selective supervision problem in which the measured reward drop sets both the direction and the strength of the auxiliary signal, and instantiate it as SIVA-RL.
- **Gains across RL backbones and model scales.** SIVA-RL raises the overall average across seven matched configurations spanning Qwen2.5-VL, Qwen3-VL Thinking, and Qwen3.5. Gains are larger on the vision-dependent benchmark group in five of seven configurations.

2 Related Work

Multimodal reinforcement learning with verifiable rewards. GRPO enables critic-free optimization through group-relative advantages; DAPO extends it with improved clipping, dynamic sampling, and entropy regularization for large-scale training Shao et al. (2024); Yu et al. (2026). Recent work brings RLVR to multimodal reasoning through visual cold-start data, rule-verifiable rewards, and group- or step-level policy optimization Huang et al. (2025b); Yao et al. (2025); Meng et al. (2025); Wang et al. (2026a). The reward in these methods is the verifiable answer on the clean input. A correct answer can therefore be independent of the image. None of these methods records whether the same prediction would change under a visual intervention. SIVA-RL augments GRPO- and DAPO-style training with outcome-conditioned intervention alignment, which routes each constructed pair by the observed reward drop.

Visual perception and intervention supervision. One line of work strengthens visual grounding through auxiliary perception rewards, clean–intervention policy contrast, or token- and trajectory-level visual credit assignment Xiao et al. (2025); Wang et al. (2025b); Huang et al. (2025a); Wang et al. (2026c); Li et al. (2026). PAPO Wang et al. (2025b) is the closest predecessor: it maximizes the divergence between the policy's behavior on clean and corrupted views, applied uniformly to every constructed pair. Another line introduces distorted, masked, mixed, or otherwise intervened inputs to improve visual exploration, sensitivity, and robustness Liu et al. (2026); Li et al. (2025); Zhang et al. (2025b); Gao et al. (2026). These establish that supervision beyond answer correctness benefits visual grounding. A subset of those approaches, most directly PAPO and counterfactual consistency methods, assigns a fixed optimization role to the constructed intervention. Other methods instead modify token-level credit or differential reasoning objectives. SIVA-RL targets the former design choice: when an intervention is used to prescribe a directional clean–intervention alignment objective, its effect should be measured at the sample–intervention level. SIVA-RL uses PAPO-style separation for Stage 1 warm-up and activates outcome-conditioned alignment in Stage 2 (Section 3).

3 Method

3.1 Overview and Base RL Backbone

We consider a multimodal reasoning dataset $\mathcal{D} = \{(q_i, x_i, a_i^\star)\}_{i=1}^{N}$, where q_i, x_i, and $a_i^\star$ denote the question, image, and target answer, respectively. For each clean input, the rollout policy samples a group of G responses,

$$y_{i,g} \sim \pi_{\theta_{\text{old}}}\big(\cdot \mid q_i, x_i\big), \qquad g = 1, \ldots, G. \tag{1}$$

Each response receives a verifiable reward $r_{i,g}$, from which the backbone computes a group-relative advantage $A_{i,g}$ and the actor update. We write the full GRPO- or DAPO-style backbone objective as $\mathcal{L}_{\text{base}}$ Shao et al. (2024); Yu et al. (2026); its advantage estimation, clipping, filtering, and regularization are detailed in the appendix.

Two-stage training. The reported recipe has two stages. Stage 1 is a PAPO-style Wang et al. (2025b) perception-separation warm-up that applies a *uniform* separation objective to every constructed pair. Stage 2 switches that term off ($\beta_{\text{prcp}} = 0$) and applies the outcome-conditioned

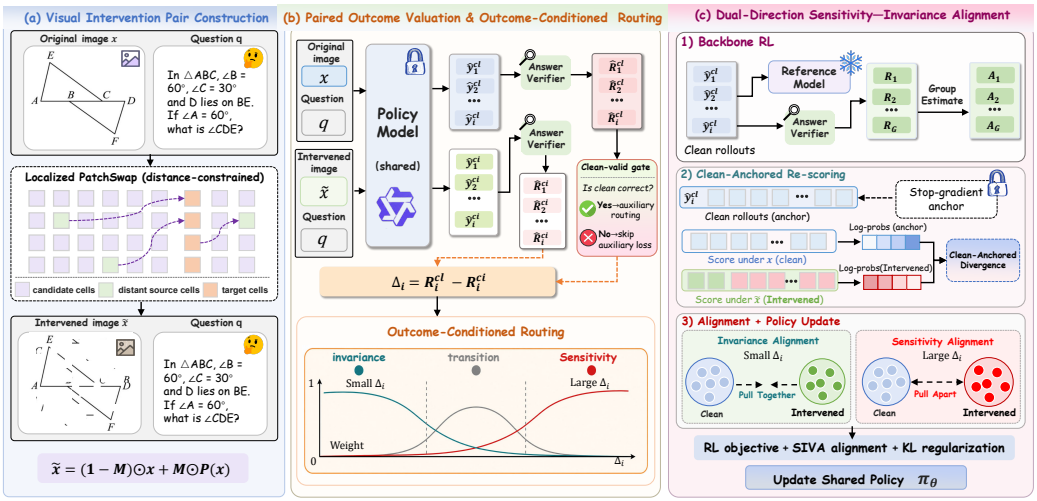


Figure 2: **Overview of SIVA-RL.** The upper branch represents the standard clean-image RLVR backbone. **(a)** Token-aligned, distance-constrained PatchSwap constructs a localized candidate intervention $\tilde{x}_i$. **(b)** A frozen audit policy performs paired decoding and converts the clean-to-intervention reward drop into sensitivity and invariance weights. **(c)** The clean rollout responses are re-scored under both visual conditions. Large-drop pairs are encouraged to exceed a sensitivity margin, while valid low-drop pairs are constrained within an invariance margin.

routing and dual-direction alignment below. Table 1 reports the complete two-stage recipe; Table 4 evaluates its components. Schedules and coefficients are in Appendix B.

SIVA-RL preserves the clean-image backbone update and augments it with an outcome-conditioned intervention-alignment objective with three components (Figure 2): **(a)** localized intervention construction, **(b)** outcome-conditioned soft routing, which decides per sample whether a candidate intervention reinforces sensitivity or encourages invariance, and **(c)** clean-anchored dual-direction alignment, which reuses existing clean-rollout responses and so adds no stochastic sampling beyond the paired audit needed for valuation.

3.2 Localized PatchSwap Construction

Given an image x_i, we partition its pixel space into a grid Ω_i aligned with the effective visual-token grid after patchification and spatial merging. Let $p_{i,u}$ denote the image cell indexed by $u \in \Omega_i$. We first sample candidate cells with ratio η and then select effective replacement targets with probability ρ:

$$C_i \sim \mathrm{SampleMask}(\Omega_i, \eta), \quad C_{i,u} \in \{0,1\}, \tag{2}$$

$$B_{i,u} \sim \mathrm{Bernoulli}(\rho), \tag{3}$$

$$T_{i,u} = C_{i,u} B_{i,u}. \tag{4}$$

Here $C_{i,u} = 1$ marks cell u as eligible for intervention; $T_{i,u} = 1$ marks it as an active replacement target. Eligible but unselected cells appear unmodified in the intervention.

For every target cell u, we form a valid within-image source set

$$\mathcal{S}_i(u) = \left\{ v \in \Omega_i : C_{i,v} = 0,\ \bar{d}_i(u,v) \geq d_{\min} \right\}, \qquad s_i(u) \sim \mathrm{Unif}\big(\mathcal{S}_i(u)\big), \tag{5}$$

where $\bar{d}_i(u,v) \in [0,1]$ is the normalized grid distance between u and v, and $d_{\min}$ excludes spatially adjacent sources. The intervention view is then assembled as

$$\tilde{p}_{i,u} = \begin{cases} p_{i,s_i(u)}, & T_{i,u} = 1, \\ p_{i,u}, & T_{i,u} = 0, \end{cases} \qquad \tilde{x}_i = \mathrm{Assemble}\big(\{\tilde{p}_{i,u}\}_{u \in \Omega_i}\big). \tag{6}$$

This disrupts localized evidence and spatial relationships while preserving the image's global statistics and style, giving interventions harder than global masking yet less destructive than full removal. Grid resolution, η, ρ, and $d_{\min}$ are specified in Appendix B.

3.3 OUTCOME-CONDITIONED SOFT ROUTING

The same PatchSwap operator can produce different behavioral effects across samples. SIVA-RL therefore assigns no fixed optimization role to constructed interventions. SIVA-RL instead evaluates each clean–intervention pair with a frozen audit policy $\pi_{\bar{\theta}}$, a no-gradient snapshot of the current policy. For compactness, let $x_i^{\mathrm{cl}} = x_i$ and $x_i^{\mathrm{ci}} = \tilde{x}_i$. Paired sampled audit decoding and answer evaluation are defined by

$$\hat{y}_{i,m}^{v} \sim \pi_{\bar{\theta}}(\cdot \mid q_i, x_i^{v}), \qquad R_i^{v} = \frac{1}{M}\sum_{m=1}^{M} r_{\mathrm{ans}}(\hat{y}_{i,m}^{v}, a_i^{\star}), \quad v \in \{\mathrm{cl}, \mathrm{ci}\}, \tag{7}$$

where r_{ans} is the answer-level verifier used for intervention valuation, and R_i^v is the empirical verifier score over the sampled audit rollouts for view v. The audit policy is fixed during the update, and no gradient is propagated through the decoded answers or their rewards.

We define the clean-anchored validity gate and observed reward drop as

$$b_i = \mathbb{I}[R_i^{\mathrm{cl}} \geq \tau_c]\, \Phi_i^{\mathrm{cl}} \Phi_i^{\mathrm{ci}}, \qquad \Delta_i = R_i^{\mathrm{cl}} - R_i^{\mathrm{ci}}, \tag{8}$$

where $\Phi_i^v \in \{0, 1\}$ indicates that the sampled audit outputs for view v contain a parseable response for the answer verifier. A clean–intervention pair is *valid* when $b_i = 1$. The clean-outcome gate prevents failed clean predictions from contributing spurious sensitivity or invariance signals. To obtain routing weights, we define

$$\mathrm{Ramp}(z; a, b) = \mathrm{clip}\left(\frac{z-a}{b-a}, 0, 1\right), \qquad a < b, \tag{9}$$

and compute

$$\begin{aligned} w_i^{\mathrm{sen}} &= b_i\ \mathrm{Ramp}(\Delta_i; \delta_{\mathrm{sen}}^{\mathrm{lo}}, \delta_{\mathrm{sen}}^{\mathrm{hi}}), \\ w_i^{\mathrm{inv}} &= b_i\ \left[1 - \mathrm{Ramp}(\Delta_i; \delta_{\mathrm{inv}}^{\mathrm{lo}}, \delta_{\mathrm{inv}}^{\mathrm{hi}})\right]. \end{aligned} \tag{10}$$

The sensitivity weight increases monotonically with the reward drop; the invariance weight is highest for valid pairs the intervention barely changes. The two ramps define mutually exclusive branches with $\delta_{\mathrm{inv}}^{\mathrm{hi}} = \delta_{\mathrm{sen}}^{\mathrm{lo}}$ (Appendix B). Pairs near the boundary receive small weights, which increase as the observed drop moves into the corresponding branch. Training uses these weights and the validity gate directly.

3.4 CLEAN-ANCHORED DUAL-DIRECTION ALIGNMENT

After routing, SIVA-RL reuses every response $y_{i,g}$ already sampled from the clean-image rollout. The same token sequence is teacher-forced and re-scored under the clean and intervention visual conditions using the trainable policy π_θ:

$$\ell_{i,g,t}^{v} = \log \pi_\theta(y_{i,g,t} \mid q_i, x_i^{v}, y_{i,g,<t}), \qquad v \in \{\mathrm{cl}, \mathrm{ci}\}. \tag{11}$$

Because routing weights are defined at the input-pair level, all G responses for example i share the same w_i^{sen} and w_i^{inv}.

We treat the clean-conditioned likelihood as a stop-gradient anchor. Let $\mathrm{sg}[\cdot]$ denote the stop-gradient operator. The clean-anchored token discrepancy and its response-level average are

$$\begin{aligned} z_{i,g,t} &= \ell_{i,g,t}^{\mathrm{ci}} - \mathrm{sg}[\ell_{i,g,t}^{\mathrm{cl}}], \qquad d_{i,g,t} = \exp(z_{i,g,t}) - z_{i,g,t} - 1, \\ D_{i,g} &= \frac{1}{|\mathcal{T}_{i,g}|}\sum_{t \in \mathcal{T}_{i,g}} d_{i,g,t}, \end{aligned} \tag{12}$$

where $\mathcal{T}_{i,g}$ is the valid response-token set. Here $d_{i,g,t}$ is a non-negative Monte Carlo surrogate for the token-level clean-to-intervention KL; because only clean-rollout tokens are evaluated, $D_{i,g}$ approximates the full-vocabulary KL rather than computing it. Two margin objectives follow:

$$\begin{aligned} \mathcal{L}_{\mathrm{sen}} &= \frac{\sum_{i,g} w_i^{\mathrm{sen}} [m_{\mathrm{sen}} - D_{i,g}]_+}{\sum_{i,g} w_i^{\mathrm{sen}} + \epsilon}, \\ \mathcal{L}_{\mathrm{inv}} &= \frac{\sum_{i,g} w_i^{\mathrm{inv}} [D_{i,g} - m_{\mathrm{inv}}]_+}{\sum_{i,g} w_i^{\mathrm{inv}} + \epsilon}, \end{aligned} \tag{13}$$

where $[z]_+ = \max(z, 0)$. For a high-drop pair $\mathcal{L}_{\text{sen}}$ enforces a minimum divergence m_{sen} between the clean and intervention policies; for a valid low-drop pair $\mathcal{L}_{\text{inv}}$ holds it below m_{inv}. *Dual-direction* thus means pushing apart versus pulling together under one clean-anchored measure, not a symmetric divergence objective. The complete objective is

$$\mathcal{L}_{\text{SIVA}} = \mathcal{L}_{\text{base}} + \lambda_{\text{sen}}\mathcal{L}_{\text{sen}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}, \tag{14}$$

where λ_{sen} and λ_{inv} control the two auxiliary branches. Setting either to zero recovers the corresponding single-branch variant. In Stage 2, $\mathcal{L}_{\text{base}}$ excludes $\mathcal{L}_{\text{prcp}}$ ($\beta_{\text{prcp}} = 0$); that term is Stage 1 only. Remaining backbone regularizers stay in $\mathcal{L}_{\text{base}}$. Routing thresholds, margins, and coefficients are in Appendix B.

4 Experiments

We evaluate SIVA-RL across multiple multimodal RL backbones and model scales on nine benchmarks spanning mathematical, logical, counting, and multi-disciplinary multimodal reasoning. The experiments ask whether SIVA-RL improves its matched RL backbone across scales (§4.2), how intervention construction and outcome-conditioned alignment each contribute (§4.3), and whether sample-level routing variation persists throughout training under a fixed PatchSwap operator (§4.4).

4.1 Experimental Setup

Models and RL backbones. We train Qwen2.5-VL-3B and Qwen2.5-VL-7B Bai et al. (2025b) with both GRPO and DAPO Shao et al. (2024); Yu et al. (2026), and Qwen3-VL Thinking at 2B and 8B Bai et al. (2025a) and Qwen3.5-4B Qwen Team (2026) with GRPO. All seven SIVA-RL runs train on ViRL39K Wang et al. (2026a) and share the hyperparameters in Appendix B. Each SIVA-RL run follows the two-stage recipe of Section 3: a 200-step PatchSwap warm-up and a 100-step outcome-conditioned stage. We report the final checkpoint. Matched baselines use the same data, optimizer, batch, and 300-step budget and do not include the SIVA-RL auxiliary losses.
Benchmarks and metrics. Following prior perception-aware multimodal RL protocols Wang et al. (2025b; 2026c), we split the suite into two groups. General multimodal reasoning: Geometry3K (Geo3K) Lu et al. (2021), MathVista Lu et al. (2024), We-Math Qiao et al. (2025), MMK12 Meng et al. (2025), and MathVerse Zhang et al. (2024). Vision-dependent: LogicVista Xiao et al. (2024), the counting subset of Super-CLEVR Li et al. (2023), MMMU-Pro Yue et al. (2025), and the vision-only subset of MathVerse, denoted MathVerse-V Zhang et al. (2024). We report Avg@8 accuracy (%); group and overall scores are benchmark-level macro averages.

External results. External numbers are taken from the original papers or official reports, with unavailable entries denoted "–".

4.2 Main Results

Improvements over matched RL backbones. The two-stage recipe improves all seven matched configurations in Table 1, by 1.78 to 9.50 absolute points (2.60% to 20.28% relative). On Qwen2.5-VL, GRPO gains 5.97 points at 3B and 4.12 at 7B. DAPO gains 3.58 points at 3B and 8.24 at 7B. In each of these four comparisons the vision-dependent group moves by more than the general group. GRPO-3B gains +7.63 against +4.63. DAPO-7B gains +8.79 against +7.81.

The same 300-step budget carries over to Qwen3-VL Thinking and Qwen3.5. Qwen3-VL-2B rises from 46.84 to 56.34 (+9.50; +20.28% relative). The vision-dependent average rises by 9.49 points, from 43.97 to 53.46, and the general average rises by 9.50, from 49.14 to 58.64. LogicVista moves from 29.84 to 44.16, MathVerse from 47.98 to 63.17, MathVerse$_V$ from 45.41 to 60.57. Qwen3-VL-8B rises from 68.46 to 70.24 (+1.78; +2.60% relative) and improves every benchmark. Its vision-dependent average rises by 1.92 points and its general average by 1.66. We-Math moves from 80.43 to 82.79, Counting from 88.00 to 92.31, and LogicVista from 56.80 to 58.25. Qwen3.5-4B rises from 69.51 to 75.65 (+6.14; +8.84% relative) and improves every benchmark. Geo3K moves from 71.32 to 77.87, MathVerse from 74.24 to 82.41, and MathVerse$_V$ from 70.06 to 78.79. Five of the seven comparisons place the larger gain on vision-dependent reasoning. Qwen3-VL-2B splits 9.50 and 9.49 across the general and vision-dependent groups, and Qwen3.5-4B gains more on the general group.

Table 1: **Main results on general multimodal reasoning, vision-dependent multimodal reasoning, and overall performance with Qwen2.5-VL, Qwen3-VL Thinking, and Qwen3.5 backbones.** $\Delta^{\%}_{\text{rel}}$ denotes the relative overall improvement over the matched RL baseline at the same scale.

Method	General Multimodal Reasoning						Vision-Dependent Multimodal Reasoning					Overall	
	Geo3k	MathVista	We-Math	MMK12	MathVerse	AVG	LogicVista	Counting	MMMU-Pro	MathVerse$_V$	AVG	AVG	$\Delta^{\%}_{\text{rel}}$
Qwen2.5-VL													
GRPO-3B	28.72	59.34	58.90	57.24	55.25	51.89	38.14	55.81	25.66	52.26	42.97	47.92	–
GRPO-3B + SIVA-RL	34.26	62.12	61.92	64.63	59.66	56.52	42.42	75.19	29.12	55.68	50.60	53.89	+12.46
GRPO-7B	40.18	65.48	68.12	72.26	66.51	62.51	45.62	73.94	35.17	61.71	54.11	58.78	–
GRPO-7B + SIVA-RL	44.41	69.24	69.53	76.03	69.29	65.70	45.83	88.75	37.43	65.64	59.42	62.90	+7.01
DAPO-3B	31.20	60.89	59.95	66.83	56.25	55.02	40.69	74.25	28.42	53.09	49.11	52.40	–
DAPO-3B + SIVA-RL	37.81	61.41	62.59	68.74	61.18	58.35	43.34	80.25	30.04	58.49	53.03	55.98	+6.83
DAPO-7B	35.92	61.91	58.51	75.93	55.64	57.58	37.05	90.05	29.02	51.04	51.79	55.01	–
DAPO-7B + SIVA-RL	42.95	64.09	70.50	79.68	69.73	65.39	47.23	92.25	36.94	65.91	60.58	63.25	+14.98
Qwen3-VL (Thinking)													
GRPO-2B	39.29	53.58	57.12	47.71	47.98	49.14	29.84	80.13	20.51	45.41	43.97	46.84	–
GRPO-2B + SIVA-RL	50.98	62.36	70.08	46.59	63.17	58.64	44.16	82.69	26.42	60.57	53.46	56.34	+20.28
GRPO-8B	64.50	67.63	80.43	65.56	77.78	71.18	56.80	88.00	41.25	74.18	65.06	68.46	–
GRPO-8B + SIVA-RL	65.50	68.99	82.79	68.78	78.14	72.84	58.25	92.31	42.78	74.58	66.98	70.24	+2.60
Qwen3.5													
GRPO-4B	71.32	70.88	77.32	68.45	74.24	72.44	57.80	90.69	44.81	70.06	65.84	69.51	–
GRPO-4B + SIVA-RL	77.87	76.61	83.49	75.69	82.41	79.22	62.30	94.69	49.00	78.79	71.20	75.65	+8.84

Table 2: **Comparison with reported results of proprietary, general-purpose, and math-oriented multimodal VLMs.** Bold and underlined denote the best and second-best results in each column, ranked across all rows rather than within our own block. Overall ranking includes only models with results on all six benchmarks; green subscripts indicate gains over the base model.

Method	Mathematical Multimodal Reasoning						
	Geo3k	MathVista	We-Math	MathVerse	MathVerse$_V$	MMK12	AVG
Proprietary Models							
GPT-4o OpenAI (2024a)	–	64.7	62.8	50.2	53.8	55.8	57.5
GPT-4o-mini OpenAI (2024b)	–	59.9	56.3	42.3	45.1	51.9	51.1
Gemini-2.0-Flash Google DeepMind (2024)	–	70.4	47.4	47.8	48.7	65.2	55.9
General-Purpose Multimodal VLMs							
Qwen2.5-VL-72B Bai et al. (2025b)	–	**74.2**	49.1	47.3	48.6	70.5	57.9
InternVL2.5-8B Chen et al. (2024)	–	64.9	44.9	37.0	40.2	46.8	46.8
LLaVA-OneVision-7B Li et al. (2024)	–	58.5	44.1	–	–	–	51.3
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)	–	69.6	61.5	–	–	–	65.6
LLaVA-Critic-R1-7B Wang et al. (2025a)	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. (2025)	36.4	59.1	45.3	57.6	56.4	60.6	52.6
MM-Eureka-7B-CPGD Meng et al. (2025)	37.6	64.2	64.3	63.7	59.2	64.7	59.0
R1-VL-7B Zhang et al. (2025a)	31.9	63.5	56.1	42.0	43.2	55.3	48.7
VLAA-Thinker-7B Chen et al. (2025)	24.2	67.4	65.9	47.9	52.0	63.2	53.4
VL-Rethinker-7B Wang et al. (2026a)	33.6	61.3	66.5	64.0	60.8	59.8	57.7
RACRO-7B-CRO-GRPO Gou et al. (2025)	41.4	61.7	68.9	65.7	61.7	70.5	61.7
ThinkLite-VL-7B Wang et al. (2026b)	34.4	68.9	63.5	49.5	46.0	56.2	53.1
RL-based Qwen2.5-VL Variants							
PAPO-3B Wang et al. (2025b)	31.0	61.4	60.1	57.1	54.0	57.4	53.5
VPPO-3B Huang et al. (2025a)	35.8	63.6	61.4	61.3	58.1	65.3	57.6
DVRP-3B Gao et al. (2026)	34.5	65.5	60.3	57.7	54.5	61.2	55.6
PAPO-7B Wang et al. (2025b)	40.3	69.5	66.8	68.4	65.0	72.5	63.7
DVRP-7B Gao et al. (2026)	<u>43.4</u>	<u>70.9</u>	67.8	68.9	65.3	74.1	65.1
Our Models							
Qwen2.5-VL-3B (base)	20.6	40.6	23.9	30.9	28.2	34.8	29.8
DAPO + SIVA-RL	37.8 (+17.2)	61.4 (+20.8)	62.6 (+38.7)	61.2 (+30.3)	58.5 (+30.3)	68.7 (+33.9)	58.4 (+28.6)
GRPO + SIVA-RL	34.3 (+13.7)	62.1 (+21.5)	61.9 (+38.0)	59.7 (+28.8)	55.7 (+27.5)	64.6 (+29.8)	56.4 (+26.6)
Qwen2.5-VL-7B (base)	33.8	55.9	41.8	45.6	36.9	43.7	43.0
DAPO + SIVA-RL	43.0 (+9.2)	64.1 (+8.2)	**70.5** (+28.7)	**69.7** (+24.1)	**65.9** (+29.0)	**79.7** (+36.0)	<u>65.5</u> (+22.5)
GRPO + SIVA-RL	**44.4** (+10.6)	69.2 (+13.3)	<u>69.5</u> (+27.7)	<u>69.3</u> (+23.7)	<u>65.6</u> (+28.7)	<u>76.0</u> (+32.3)	**65.7** (+22.7)

Comparison with existing multimodal VLMs. The 7B SIVA-RL models achieve the strongest full-suite averages among reported 7B systems and perform strongly on We-Math, MathVerse, MathVerse-V, and MMK12 (Table 2). The 3B variants remain competitive with much larger models.

Table 3: **Ablation of intervention construction and outcome-conditioned alignment on Qwen2.5-VL-3B with GRPO.** PatchSwap w/o dist. sets $d_{\min} = 0.00$; PatchSwap denotes the default distance-constrained construction. Δ_{base} denotes the relative change of Overall AVG over the GRPO-3B baseline.

Setting	General Multimodal Reasoning					Vision-Dependent Multimodal Reasoning				Overall	
	Geo3k	MathVista	We-Math	MMK12	MathVerse	LogicVista	Counting	MMMU-Pro	MathVerse$_V$	AVG	Δ_{base}
GRPO-3B baseline	28.72	59.34	58.90	57.24	55.25	38.14	55.81	25.66	52.26	47.92	–
Random Mask	30.95	61.38	60.09	57.39	57.14	38.67	62.56	27.11	53.95	49.92	+4.17%
Random Mask + SIVA-RL	31.47	60.40	60.01	59.42	58.20	38.93	68.62	28.11	54.95	51.12	+6.68%
PatchSwap w/o dist.	30.30	58.02	59.60	52.01	56.39	38.93	63.38	26.08	52.99	48.64	+1.50%
PatchSwap w/o dist. + SIVA-RL	33.96	**62.64**	61.28	60.38	59.00	42.20	**75.38**	28.30	55.66	53.20	+11.02%
PatchSwap	31.66	59.36	60.89	55.61	58.14	40.35	64.50	27.63	55.13	50.36	+5.09%
PatchSwap + SIVA-RL	**34.26**	62.12	**61.92**	**64.63**	**59.66**	**42.42**	75.19	**29.12**	**55.68**	**53.89**	**+12.46%**

Table 4: **Ablation of outcome-conditioned alignment on Qwen2.5-VL-3B with a GRPO backbone.** Complete results are reported on all nine benchmarks. Hard Sens. and Soft Sens. are alternative routers; their Δ_{step} is the relative change of Overall AVG over +PatchSwap. Subsequent rows report the increment over the previous row. The last row shuffles $(w_i^{\text{sen}}, w_i^{\text{inv}})$ within each batch.

Setting	General Multimodal Reasoning					Vision-Dependent Multimodal Reasoning				Overall	
	Geo3k	MathVista	We-Math	MMK12	MathVerse	LogicVista	Counting	MMMU-Pro	MathVerse$_V$	AVG	Δ_{step}
GRPO-3B baseline	28.72	59.34	58.90	57.24	55.25	38.14	55.81	25.66	52.26	47.92	–
+PatchSwap	31.66	59.36	60.89	55.61	58.14	40.35	64.50	27.63	55.13	50.36	+5.09%
+Hard Sens.	31.32	60.74	61.49	61.64	57.82	42.00	65.06	27.67	54.82	51.40	+2.07%
+Soft Sens.	32.30	**63.90**	59.15	62.58	56.46	39.49	71.56	27.64	53.31	51.82	+2.90%
+Soft Sens. + Warmup (w/o Inv.)	**34.48**	61.36	**62.68**	61.88	58.37	40.75	71.81	27.45	55.00	52.64	+1.58%
+SIVA-RL	34.26	62.12	61.92	**64.63**	**59.66**	**42.42**	**75.19**	**29.12**	**55.68**	**53.89**	**+2.37%**
+SIVA-RL (shuffled)	31.29	60.18	59.73	59.20	57.95	38.67	68.39	27.91	54.78	50.90	−5.55%

4.3 Ablation Studies

Intervention Construction. Table 3 disentangles intervention construction from outcome-conditioned alignment. Among construction-only variants, distance-constrained PatchSwap is best overall (50.36), ahead of random masking (49.92) and unconstrained PatchSwap (48.64): a minimum source distance avoids overly local replacements while preserving more visual context than masking. Routing then improves all three operators, to 53.89, 51.12, and 53.20 respectively, so construction and sample-wise valuation play complementary roles.

Outcome-Conditioned Alignment. Table 4 evaluates the alignment components while holding distance-constrained PatchSwap fixed. Hard Sens. and Soft Sens. keep only the sensitivity branch. Soft Sens. uses the ramp in Equation 10. Relative to PatchSwap alone (50.36), they increase the overall average to 51.40 and 51.82. Soft routing achieves the higher overall score, improving on the hard-routing variant by 0.42 points under the same PatchSwap construction.

The uniform-sensitivity warm-up raises performance to 52.64, so early optimization benefits from a stabilized alignment signal before routing. Adding invariance alignment for clean-anchored low-drop pairs completes the SIVA-RL objective and reaches 53.89 overall, a 1.85-point gain on the vision-dependent subset. Low-drop interventions are therefore not uninformative: they supply usable clean-anchored consistency constraints. Shuffling $(w_i^{\text{sen}}, w_i^{\text{inv}})$ within each batch, which keeps the marginal weight distribution and breaks the match to Δ_i, drops the overall average from 53.89 to 50.90, back to the PatchSwap-only level of 50.36. The per-sample correspondence is therefore load-bearing. The same assignment shows up inside the vision-dependent columns. Under PatchSwap alone, Counting and LogicVista score 64.50 and 40.35. Full SIVA-RL raises them to 75.19 and 42.42. The shuffled router returns 68.39 and 38.67. MMMU-Pro moves from 27.63 to 29.12, and MathVerse$_V$ from 55.13 to 55.68. Both return to 27.91 and 54.78 once the weights are shuffled. The measured drop, rather than a second margin term, carries these gains. The warm-up and the invariance term do not move the same columns. Soft sensitivity leaves Geo3k at 32.30 and We-Math at 59.15. The uniform-sensitivity warm-up raises those two scores to 34.48 and 62.68. Invariance then leaves Geo3k at 34.26 and We-Math at 61.92, while Counting rises from 71.81 to 75.19. Shuffling the same weights returns Counting to 68.39. MMK12 falls from 57.24 to 55.61 under PatchSwap alone, then reaches 64.63 with full SIVA-RL. The shuffled router leaves MMK12 at 59.20. MathVista is highest under soft sensitivity, at 63.90, and is 62.12 under full SIVA-RL.



Figure 3: **Training-time outcome-conditioned routing diagnostics.** We report the sensitivity routing rate, invariance routing rate, and mean paired answer-reward drop for the four backbone–scale configurations.

Hyperparameter Sensitivity. We vary the PatchSwap candidate-cell ratio η and the minimum source distance $d_{\min}$ one at a time on the GRPO-3B backbone. Performance stays within 2.1 points across the full η sweep and 1.5 points across the $d_{\min}$ sweep, and both profiles have an interior optimum, since a too-small $d_{\min}$ barely perturbs the evidence while a too-large one drifts toward arbitrary corruption. Every setting stays above the GRPO-3B baseline of 47.92: the weakest, η=0.2, still reaches 51.78, and removing the distance constraint entirely ($d_{\min}$=0.00) reaches 53.20 against the tuned 53.89. The construction knobs therefore govern the *supply* of informative clean–intervention pairs while routing decides how each pair is used, yielding improvements across the tested perturbation settings. Full per-benchmark results are in Appendix C.

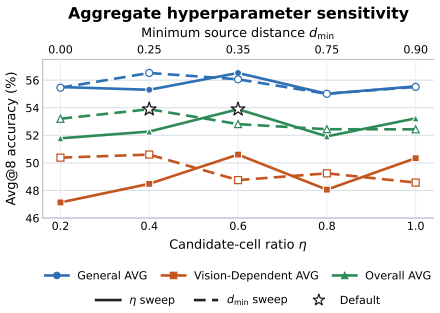


Figure 4: **Hyperparameter sensitivity.** Macro Avg@8 under varying η and $d_{\min}$.

4.4 Mechanistic Analysis

We track routing rates and the mean paired reward drop on the four Qwen2.5-VL runs. The question is whether sample-level routing variation persists past the 200-step warm-up. In Figure 3, neither the sensitivity rate nor the invariance rate collapses to a single regime. The mean paired answer-reward drop stays moderate rather than vanishing. The same PatchSwap operator therefore keeps supplying both branches throughout Stage 2.

The branches also separate on the clean-to-intervention divergence D (Figure 6). Sensitivity-routed pairs hold D in a 0.05–0.15 band, around the margin $m_{\text{sen}} = 0.05$. Invariance-routed pairs are pulled toward m_{inv}. DAPO-3B enters Stage 2 near $D = 0.23$ and reaches the 0.03–0.05 band within about fifteen steps. After that transient, the sensitivity band stays about two to three times the invariance band of 0.02–0.06. Both margin hinges stay active on most routed pairs (Table 11).

5 Conclusion

SIVA-RL assigns visual supervision by the observed outcome change of each sample–intervention pair. A paired audit finds that 36.3% of clean-correct examples are outcome-drop under one operator and outcome-stable under another. This rate is 6.1× the disagreement any single per-sample dependence level admits. Operator-conditioned supervision therefore assigns a separation target to a large minority of pairs whose outcome does not move. SIVA-RL measures that reward drop and routes each pair toward a margin-bounded sensitivity objective or a bounded consistency objective. Across seven matched configurations spanning Qwen2.5-VL, Qwen3-VL Thinking, and Qwen3.5, the nine-benchmark average rises by 1.78–9.50 points. The vision-dependent group gains more in five of seven configurations. The same budget reaches Qwen3-VL-2B Thinking (+9.50), Qwen3-VL-8B (+1.78), and Qwen3.5-4B (+6.14). Qwen3.5-4B improves every benchmark in the suite. Both Qwen3-VL Thinking runs stay above their matched baselines, and low-drop pairs are kept as consistency targets.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) to assist with manuscript drafting and proofreading. We have not used generative AI tools to generate experimental results, figures, tables, or citations. We reviewed all AI-assisted text for accuracy and take full responsibility for the final content of this paper, including all claims.

ETHICS STATEMENT

This work trains and evaluates on publicly released datasets and introduces no new data collection involving human subjects. Appendix E uses the public VQA-RAD, SLAKE, and PathVQA benchmarks, which are distributed in de-identified form by their original authors; we use them under their released licenses for research evaluation only, and we did not attempt any re-identification. The medical evaluation measures benchmark accuracy; the models have not been clinically validated and are not intended for diagnostic use. The training corpora inherit whatever demographic and content biases are present in their sources, and our method does not mitigate them.

REPRODUCIBILITY STATEMENT

Our experimental implementation is based on publicly available RL training frameworks. Complete hyperparameters are provided in Appendix B (Table 7), including backbone-specific RL configurations, intervention-construction parameters, routing thresholds, and loss coefficients. The training data (ViRL39K) and all evaluation benchmarks are publicly available. The algorithm pseudocode in Appendix B.4 provides a step-by-step description sufficient for independent re-implementation, and the paired outcome audit that motivates the method is specified separately in Appendix A so that it can be reproduced without the training pipeline. We will release the training and evaluation code together with the trained model weights for every reported configuration, to support independent verification of the reported results.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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

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

Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, and Yu-Gang Jiang. Thinking with deltas: Incentivizing reinforcement learning via differential visual reasoning policy. *arXiv preprint arXiv:2601.06801*, 2026.

Google DeepMind. Introducing Gemini 2.0: Our new AI model for the agentic era. Google DeepMind blog, December 2024. URL https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025a.

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

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

Qiming Li, Tianlun Li, Xiaolong Cheng, Hangyu Li, Ruiyan Gong, Kangning Niu, Kaitao Jiang, and Mu Xu. Prpo: Perception-reinforced policy optimization via token-level dynamic advantage reshaping. *arXiv preprint arXiv:2606.08708*, 2026.

Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Linghe Kong, Lichao Sun, and Weiran Huang. Vision matters: Simple visual perturbations can boost multimodal math reasoning. *arXiv e-prints*, pp. arXiv–2506, 2025.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14963–14973, 2023.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *Advances in Neural Information Processing Systems*, 38:2923–2957, 2026.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

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

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12700–12710, 2021.

OpenAI. Hello GPT-4o. OpenAI product announcement, May 2024a. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI product announcement, July 2024b. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891, 2026a.

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

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *Advances in Neural Information Processing Systems*, 38: 118818–118850, 2026b.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning. *arXiv preprint arXiv:2604.09349*, 2026c.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025b.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

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

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2376–2385, 2025.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1859–1869, 2025a.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025b.

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

APPENDIX CONTENTS

A Paired Outcome Audit Protocol **15**

A.1 Design . . . 15
A.2 Outcome Definitions . . . 15
A.3 Results . . . 16
A.4 Joint Agreement Across Operators . . . 16
A.5 Relation to the Training-Time Routing Signal . . . 17

B Implementation Details **17**

B.1 Training Recipe . . . 17
B.2 Training Budget . . . 19
B.3 Validation Data and Hyperparameter Selection . . . 20
B.4 Algorithm Pseudocode of SIVA-RL . . . 20
B.5 Backbone Instantiations and Final Objectives . . . 20

C Full Hyperparameter Sensitivity Results **21**

C.1 Sweep Setup and Full Results . . . 21
C.2 Findings . . . 22
C.2.1 Both knobs have an interior optimum . . . 22
C.2.2 Construction matters most where vision matters most . . . 23
C.2.3 Gains persist across construction settings . . . 23
C.3 Discussion and Practical Recommendations . . . 23

D Training-Dynamics Diagnostics **24**

D.1 Reward–Accuracy Dynamics . . . 24
D.2 Routing-Branch Separation . . . 24
D.3 Magnitude of the Auxiliary Objectives . . . 25
D.4 Routing Control . . . 25

E Auxiliary Medical-Domain Accuracy **26**

E.1 Setup and Results . . . 26
E.2 Evaluation Scope . . . 26

F Qualitative Case Studies **26**

F.1 Visual Recovery . . . 26
F.2 Misleading Black-Image Correctness . . . 29
F.3 PatchSwap-Only Instability (Chart) . . . 30
F.4 Local Geometric Stability . . . 31

F.5 Residual Failure: Dense Multi-Instance Counting 32

A Paired Outcome Audit Protocol

The audit reported in Section 1 and Figure 1b is the empirical basis for the claim that a fixed intervention operator does not determine a sample's supervision direction. This appendix specifies how it was run, so that the statistic can be reproduced independently of the training pipeline.

A.1 Design

The audit is a paired, per-sample comparison. For each example we decode an answer on the clean image and on each of three intervention views of the same image, score all four answers with the same verifier, and record the transition. The audit uses greedy decoding on a fixed policy; training uses the sampled valuation defined in Section 3.3.

Audited policy. The audited policy is our GRPO-trained Qwen2.5-VL-3B baseline at optimization step 100. It produces all four answers for each example and achieves a strict-format greedy clean accuracy of $906/1800 = 50.3\%$ on the audit sample.

Sampling frame. We audit 1,800 examples, 200 from each of the nine evaluation benchmarks of Section 4.1: Geo3K, MathVista, We-Math, MMK12, MathVerse, LogicVista, the Super-CLEVR counting subset, MMMU-Pro, and MathVerse$_V$. For each benchmark we sample uniformly without replacement from the same evaluation file that Section 4 scores, using seed 42 with a deterministic per-benchmark offset. The resulting manifest is fixed once and reused for the clean view and all three intervention views, so the four conditions are compared on identical examples.

Decoding and verification. Answers are produced by a single greedy decode (temperature 0, top-p 1, top-k 1) with at most 2,048 new tokens and a 9,216-token context cutoff, under the Qwen2-VL chat template with the system message "You are a helpful assistant." The prompt is the benchmark question, which already carries the instruction to reason inside `<think>...</think>` and to place the final answer in `\boxed{}`. Answers are scored by a boxed-answer math verifier and thresholded to a binary correct/incorrect outcome. When the parser returns no boxed span, the example is retained and scored incorrect rather than discarded; this is stricter than the training-time validity gate, which excludes unparseable pairs through Φ_i^v (Equation 8), all such outputs remain in the audit denominator.

Intervention operators. Three operators are applied to each image. *Black image* replaces the entire image with a constant black frame. *Random mask* zeroes randomly chosen cells on an adaptive grid with side $\max(16, \lfloor \min(H, W)/12 \rfloor)$ pixels until at least 60% of the image pixels are covered. *PatchSwap* is the distance-constrained construction of Section 3.2. The audit uses the training parameters $\eta = 0.6$, $\rho = 0.75$, and $d_{\min} = 0.25$ (Appendix B).

The audit measures outcome heterogeneity under these three operators. The PatchSwap rates are the rates of the intervention used in training.

A.2 Outcome Definitions

Let $c_i \in \{0, 1\}$ be the binary clean outcome and c_i^{op} the binary outcome under operator op. Restricted to the clean-correct subset $\{i : c_i = 1\}$, we call a pair *outcome-drop* when $c_i^{\mathrm{op}} = 0$ and *outcome-stable* when $c_i^{\mathrm{op}} = 1$. The two categories are complementary by construction, so each operator is summarized by a single rate. We restrict to the clean-correct subset for the same reason the training-time validity gate b_i does: when the policy already fails on the clean image, neither a drop nor a stable outcome licenses a supervision direction.

A.3 RESULTS

Of the 1,800 audited examples, 906 are answered correctly on their clean image. Within that subset, the outcome-drop and outcome-stable rates are $67.8\%/32.2\%$ for the black image, $62.9\%/37.1\%$ for the random mask, and $61.8\%/38.2\%$ for PatchSwap. Every operator therefore produces both categories, and in aggregate the outcome-stable share never falls below roughly one third.

Table 5: **Per-benchmark breakdown of the paired outcome audit.** Outcome-drop rates are computed within each benchmark's clean-correct subset. PatchSwap is the distance-constrained operator of Section 3.2, with $\eta = 0.6$, $\rho = 0.75$, and $d_{\min} = 0.25$.

Benchmark	Audited	Clean-correct	Outcome-drop rate (%)		
			Black	Rnd. mask	PatchSwap
Geo3K	200	68	75.0	75.0	80.9
MathVista	200	117	70.9	56.4	60.7
We-Math	200	125	56.8	51.2	48.0
MMK12	200	113	34.5	35.4	25.7
MathVerse	200	112	46.4	49.1	43.8
LogicVista	200	81	74.1	65.4	60.5
Counting	200	132	100.0	87.9	94.7
MMMU-Pro	200	65	86.2	87.7	90.8
MathVerse$_V$	200	93	75.3	73.1	67.7
Total	1,800	906	67.8	62.9	61.8

The aggregate rate hides a wide spread, and part of the heterogeneity is across benchmarks. Under PatchSwap the outcome-stable share ranges from 5.3% on the counting subset to 74.3% on MMK12. The aggregate rate therefore combines within-benchmark and across-benchmark variation.

Within individual benchmarks, in seven of the nine benchmarks the outcome-stable share lies between 19.1% and 74.3%, so both supervision directions occur in quantity *within* a single benchmark under a single operator; counting and MMMU-Pro are the two near-homogeneous exceptions, at 94.7% and 90.8% drop. Benchmark identity is not available as a routing signal during training, since the training corpus is a single mixed corpus rather than the evaluation suite. The practical claim is therefore the within-benchmark one: a per-sample measurement is needed because neither the operator nor any task label available at training time determines the direction.

A.4 JOINT AGREEMENT ACROSS OPERATORS

The three marginal rates above are individually compatible with a reading that would leave operator-conditioned supervision intact: that visual dependence is a scalar per-sample quantity v_i, that each operator op has a severity threshold θ_{op}, and that the outcome drops exactly when $v_i > \theta_{\mathrm{op}}$. Under that model the marginal rates differ only because the operators are differently severe, and a single offline audit with one operator would suffice to label every sample. Because all three operators were applied to the same 906 clean-correct examples, we can test this reading directly.

The three-way agreement classes are as follows, with Wilson 95% intervals. 406 examples (44.8%; $[41.6, 48.1]$) lose the outcome under all three operators and 171 (18.9%; $[16.5, 21.6]$) retain it under all three, so 63.7% agree. The remaining 329 (36.3%; $[33.2, 39.5]$) *disagree*: the same example is outcome-drop under at least one operator and outcome-stable under at least one other. Table 6 gives all eight patterns.

The disagreement rate exceeds what a per-sample dependence level permits. Nested drop sets pin the disagreement rate to the spread of the marginals: every example that drops under the mildest operator must drop under the other two, so the disagreeing set is exactly the difference between the largest and smallest marginal, $67.8 - 61.8 = 6.0\%$. This is not a point prediction but an upper bound, attained only when the nesting is perfect; any weaker severity ordering gives less. The observed 36.3% is $6.1\times$ that bound. The same conclusion follows combinatorially and without

Table 6: **All eight outcome-agreement patterns on the 906 clean-correct examples.** D denotes outcome-drop and S outcome-stable. † marks the four patterns that no severity-ordered per-sample dependence level can produce, since such a model forces the drop sets to be nested as {PatchSwap} ⊆ {mask} ⊆ {black}.

Black	Rnd. mask	PatchSwap	Count	% of 906	Class
D	D	D	406	44.8	all-drop
D	D	S	78	8.6	mixed
D	S	D	74	8.2	mixed†
D	S	S	56	6.2	mixed
S	D	D	45	5.0	mixed†
S	D	S	41	4.5	mixed†
S	S	D	35	3.9	mixed†
S	S	S	171	18.9	all-stable

reference to rates: the four patterns marked † in Table 6 are individually impossible under nesting, and they account for $195/906 = 21.5\%$ of the clean-correct subset.

Under conditional independence given the observed marginals, the disagreement rate would be 69.1%. The observed 36.3% sits well below that, so the operators are strongly positively associated and operator identity plainly carries information about the likely direction. The finding is bracketed on both sides: 6.0% (single per-sample level) $< 36.3\%$ (observed) $< 69.1\%$ (independence). Operator identity is informative but insufficient, and the dependence label is a property of the sample–intervention pair rather than of either factor alone.

Parse-valid subset. Restricting to the 709 examples whose clean and all three intervention responses parse successfully, the disagreement rate is $269/709 = 37.9\%$ ($[34.4, 41.6]$), marginally higher rather than lower, and the all-stable share rises to 24.1%. The disagreement persists among parse-valid outputs.

The label does not transfer between operators. The most direct consequence for the method is visible in the pairwise contingency between the most and least destructive operators. Cross-tabulating the black image against PatchSwap gives 480 examples that drop under both, 212 stable under both, 80 that only PatchSwap breaks, and 134 that only the black image breaks. Of the 614 examples that a blank image breaks, the policy cannot answer without the image. Of these examples, 134, or 21.8% ($[18.7, 25.3]$), are unaffected by PatchSwap; on the parse-valid subset the figure is $114/443 = 25.7\%$. A dependence label transferred from a cheap full-image audit would therefore mislabel more than a fifth of the pairs it marks as dependent, which is why SIVA-RL values the specific constructed view online rather than reusing a cached per-sample score.

A.5 Relation to the Training-Time Routing Signal

The audit uses greedy binary outcomes. Training computes $\Delta_i = R_i^{\text{cl}} - R_i^{\text{ci}}$ from M sampled decodes per view. With $M = 1$, the observed drop is near-discrete; increasing M averages more audit samples. On GRPO-3B, $M = 4$ and $M = 8$ raise the nine-benchmark average by 0.33 and 0.57 points (Table 9).

B Implementation Details

B.1 Training Recipe

The main text focuses on the SIVA-RL mechanism. This appendix reports the executable training recipe. The SIVA-RL rows in Table 1 share the hyperparameters below. Qwen2.5-VL-3B/7B, Qwen3-VL Thinking 2B/8B, and Qwen3.5-4B use the same optimizer, batch, PatchSwap construction, routing thresholds, and loss coefficients. Medical-domain rows are the only exception, and they are marked in the tables. SIVA-RL runs use AdamW-bf16, gradient checkpointing, same-image PatchSwap intervention views, and the reward function's answer-accuracy component as r_{ans}. The

Table 7: Core training and rollout hyperparameters for SIVA-RL.

Hyperparameter	Value
General Training	
Backbone models	Qwen2.5-VL-3B/7B; Qwen3-VL Thinking 2B/8B; Qwen3.5-4B
Optimizer / learning rate / weight decay	AdamW-bf16 / 10^{-6} / 10^{-2}
Maximum prompt / response length	4096 / 2048
Gradient checkpointing	enabled
Reward signal	answer-accuracy scalar with format and parse validity checks
Rollout and Evaluation	
Rollouts per prompt	$G = 5$ for training rollouts
Rollout temperature / top-p	1.0 / 0.99
Intervention valuation decoding	$M = 1$ audit decode per view for the reported Table 1 recipe; temperature 0.7, top-p 0.95
Validation decoding multiplicity	GRPO: avg@8; DAPO: avg@8
Two-Stage Schedule	
Math stage 1	ViRL39K train; 4 GPUs; 200 steps; PatchSwap warm-up without SIVA routing; all seven configurations
Math stage 2	ViRL39K train; 4 GPUs; 100 steps; dual-direction SIVA alignment, $\beta_{\text{prcp}} = 0$; final checkpoint; all seven configurations
Medical stage 1	VQA-RAD SLAKE PathVQA train split; 4 GPUs; 200 steps; PatchSwap warm-up without SIVA routing
Medical stage 2	VQA-RAD SLAKE PathVQA train split; 4 GPUs; 100 steps; dual-direction SIVA alignment, $\beta_{\text{prcp}} = 0$; final checkpoint
Batching	
Math global / rollout / validation batch	128 / 256 / 512
Medical global / rollout / validation batch	64 / 384 / 256
Micro batch	4 for update and 8 for experience computation
Backbone-Specific RL Configuration	
GRPO-backbone	GRPO advantage estimator; reference-KL branch enabled; online filtering off; default actor clipping
DAPO-backbone	DAPO advantage estimator; reference-KL branch disabled; online filtering on with accuracy filter $[0.01, 0.99]$; clip low 0.2, clip high 0.28
KL penalty type / coefficient	low-variance KL; $\beta_{\text{ref}} = 10^{-2}$ for the GRPO backbone, branch disabled ($\beta_{\text{ref}} = 0$) for DAPO
Actor clipping $(\epsilon_{\text{low}}, \epsilon_{\text{high}}, \epsilon_{\text{dual}})$	GRPO: $(0.2, 0.3, 3.0)$; DAPO: $(0.2, 0.28, 3.0)$
Perception-separation coefficient β_{prcp}	10^{-2} in Stage 1; 0 in Stage 2; identical schedule for both backbones
Entropy coefficients β_{ori} / β_{aug}	0.05 / 0.05

answer reward is a scalar returned by the evaluator, not a strictly binary label. Training uses a two-stage schedule. Stage 1 is a 200-step PatchSwap perception-separation warm-up that treats all constructed intervention pairs as sensitivity-style perturbations without outcome routing or invariance alignment. Stage 2 is a 100-step outcome-conditioned stage with paired outcome valuation and dual-direction routing, and with the inherited perception-separation term switched off ($\beta_{\text{prcp}} = 0$). Matched baselines use the same optimizer, batch, and 300-step budget and do not include the SIVA-RL auxiliary losses. We report the final checkpoint.

Table 8: PatchSwap and outcome-routing hyperparameters for SIVA-RL.

Hyperparameter	Value
PatchSwap Intervention Views	
Visual cell size	28 pixels, corresponding to a 14-pixel vision patch with spatial merge size 2
Patch source	same-image non-candidate cells satisfying $d_{\min}$; if this source set is empty, or no source satisfies the distance constraint, we fall back to same-image cells excluding the target cell
Math candidate-cell ratio / replacement probability	0.6 / 0.75
Medical candidate-cell ratio / replacement probability	0.4 / 0.75
Outcome Valuation and Alignment	
Clean correctness threshold τ_c	0.75
Sensitivity reward-drop ramp	$[\delta^{\text{lo}}_{\text{sen}}, \delta^{\text{hi}}_{\text{sen}}] = [0.05, 0.5]$
Invariance reward-drop ramp	$[\delta^{\text{lo}}_{\text{inv}}, \delta^{\text{hi}}_{\text{inv}}] = [0.0, 0.05]$
Abstention interval	none: the ramps meet at $\Delta_i = 0.05$, so every valid pair is routed with attenuated weight (Section 3.3)
Sensitivity coefficient / margin	$\lambda_{\text{sen}} = 0.005$, $m_{\text{sen}} = 0.05$
Invariance coefficient / margin	$\lambda_{\text{inv}} = 0.001$, $m_{\text{inv}} = 0.01$
Audit samples per view M	1 for Table 1; 4 and 8 in the GRPO-3B sweep of Appendix B.2
Stage-2 perception-separation term	disabled: $\beta_{\text{prcp}} = 0$

B.2 Training Budget

The training budgets comprise optimization steps, audit generation, paired re-scoring, and the rollout batch schedule.

Optimization steps and reported checkpoints. Every configuration in Table 1 uses a 200-step warm-up followed by 100 routed steps. Matched baselines use the same 300-step budget. We report the final checkpoint. Table 4 reports component comparisons on GRPO-3B.

Generation budget per prompt. The backbone draws $G = 5$ clean rollouts per prompt. The Table 1 recipe decodes $M = 1$ audit rollout under each view. Paired re-scoring (Equation 11) adds two teacher-forced forward passes per sampled response.

Audit-sample count M. Table 9 varies only the frozen audit count on GRPO-3B, holding $G = 5$ and the rest of the Table 1 recipe fixed. $M = 1$ is the reported setting. Raising M from 1 to 4 increases the nine-benchmark average by 0.33 points. Raising it to 8 increases the average by 0.57 points.

Table 9: **Audit-sample count M on GRPO-3B.** $M = 1$ is the Table 1 recipe. Δ is versus $M = 1$.

	Geo3k	MathVista	We-Math	MMK12	MathVerse	LogicVista	Counting	MMMU-Pro	MathVerse$_V$	AVG	gens	Δ
$M = 1$	34.26	62.12	61.92	64.63	59.66	42.42	75.19	29.12	55.68	**53.89**	7	–
$M = 4$	34.03	63.06	62.59	64.10	59.40	40.44	79.00	28.32	57.02	54.22	13	+0.33
$M = 8$	34.86	62.28	63.20	63.86	60.11	42.00	77.81	29.01	57.03	54.46	21	+0.57

Batch schedule. The rollout batch is 256 in both stages, for every configuration in Table 1. The global batch is 128. Remaining batch settings are in Table 7.

Algorithm 1 SIVA-RL training step

Require: Current policy π_θ, old policy $\pi_{\theta_{\text{old}}}$, audit policy $\pi_{\bar{\theta}}$, batch $\mathcal{B} = \{(q_i, x_i, a_i^\star)\}_{i=1}^{B}$
Require: Group size G, audit sample count M, PatchSwap parameters $(\eta, \rho, d_{\min})$, clean threshold τ_c, reward-drop intervals, margins $(m_{\text{sen}}, m_{\text{inv}})$, coefficients $(\lambda_{\text{sen}}, \lambda_{\text{inv}})$, and $\beta_{\text{prcp}} = 0$

Phase 1: Clean grouped rollout and base RL statistics
for each prompt (q_i, x_i) in $\mathcal{B}$ **do**
  Sample clean responses $y_{i,g} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q_i, x_i)$ for $g = 1, \ldots, G$
  Compute outcome rewards $R_{i,g}$ and grouped response advantages $A_{i,g}$
end for
Phase 2: Localized intervention construction
for each image x_i **do**
  Sample candidate cells C_i and effective targets T_i on the visual-cell grid
  For each target cell, first sample a same-image non-candidate source cell with distance at least $d_{\min}$; if no valid source exists, fall back to all same-image cells except the target
  Construct PatchSwap intervention view $\tilde{x}_i$
end for
Phase 3: Outcome-conditioned valuation
for each pair $(q_i, x_i, \tilde{x}_i)$ **do**
  Set $x_i^{\text{cl}} = x_i$ and $x_i^{\text{ci}} = \tilde{x}_i$
  for $v \in \{\text{cl}, \text{ci}\}$ **do**
    Sample $\hat{y}_{i,m}^v \sim \pi_{\bar{\theta}}(\cdot \mid q_i, x_i^v)$ for $m = 1, \ldots, M$ without gradient updates
    Compute $R_i^v = \frac{1}{M} \sum_{m=1}^{M} r_{\text{ans}}(\hat{y}_{i,m}^v, a_i^\star)$
    Compute parse-validity indicator Φ_i^v
  end for
  Compute validity gate $b_i = \mathbb{I}[R_i^{\text{cl}} \geq \tau_c]\Phi_i^{\text{cl}}\Phi_i^{\text{ci}}$ and drop $\Delta_i = R_i^{\text{cl}} - R_i^{\text{ci}}$
  Convert Δ_i into soft routing weights w_i^{sen} and w_i^{inv}
end for
Phase 4: Intervention-view re-scoring and alignment
for each sampled response $y_{i,g}$ **do**
  Re-score the same token sequence under x_i^{cl} and x_i^{ci}
  Compute sampled divergence surrogate $D_{i,g}$
  Accumulate $w_i^{\text{sen}}[m_{\text{sen}} - D_{i,g}]_+$ and $w_i^{\text{inv}}[D_{i,g} - m_{\text{inv}}]_+$
end for
Phase 5: Policy update
Update θ using $\mathcal{L}_{\text{base}} + \lambda_{\text{sen}}\mathcal{L}_{\text{sen}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}$ with $\beta_{\text{prcp}} = 0$ and the remaining backbone regularizers enabled

B.3 Validation Data and Hyperparameter Selection

Periodic held-out evaluations use the 2,000-question MMK12 set, which is also included in the reported Avg@8 benchmark suite. These examples contribute no gradient updates. Reported numbers use the final Stage-2 checkpoint. MMK12 scores are not used for checkpoint selection.

The construction defaults $\eta = 0.6$ and $d_{\min} = 0.25$ were fixed before the one-factor-at-a-time sweeps. Appendix C reports those sweeps on the nine evaluation benchmarks as robustness measurements.

B.4 Algorithm Pseudocode of SIVA-RL

Algorithm 1 summarizes one SIVA-RL training step. SIVA-RL differs from a standard grouped-rollout RL update in one way: it separates the clean rollout used for policy optimization from the paired audit used for intervention outcome valuation. The audit averages answer rewards over M sampled rollouts and determines soft routing weights. The auxiliary loss then re-scores the already sampled clean responses under the clean and PatchSwap visual conditions.

B.5 Backbone Instantiations and Final Objectives

The clean-rollout importance ratio is

$$\rho_{i,g,t}(\theta) = \frac{\pi_\theta(y_{i,g,t} \mid q_i, x_i, y_{i,g,<t})}{\pi_{\theta_{\text{old}}}(y_{i,g,t} \mid q_i, x_i, y_{i,g,<t})}. \tag{15}$$

Let

$$\bar{\rho}_{i,g,t}(\theta) = \exp\Big(\mathrm{clip}\big(\log \rho_{i,g,t}(\theta), \log(1-\epsilon_{\mathrm{low}}), \log(1+\epsilon_{\mathrm{high}})\big)\Big) \tag{16}$$

be the clipped ratio used by the shared actor-loss routine in our code. With response-level advantage $A_{i,g}$ broadcast to all tokens of response $y_{i,g}$, the per-token clipped actor loss can be written as

$$\ell^{\mathrm{PG}}_{i,g,t} = \begin{cases} \max(-A_{i,g}\rho_{i,g,t}, -A_{i,g}\bar{\rho}_{i,g,t}), & A_{i,g} \geq 0, \\ \min(\max(-A_{i,g}\rho_{i,g,t}, -A_{i,g}\bar{\rho}_{i,g,t}), -A_{i,g}\epsilon_{\mathrm{dual}}), & A_{i,g} < 0. \end{cases} \tag{17}$$

The clean-image policy-gradient term is then

$$\mathcal{L}_{\mathrm{PG}}(\theta) = \mathrm{Avg}_{i,g,t}\big[\ell^{\mathrm{PG}}_{i,g,t}\big]. \tag{18}$$

Here ϵ_{low}, ϵ_{high}, and ϵ_{dual} denote the clipping hyperparameters used by the actor-loss routine.

The GRPO and DAPO backbones share this clipped actor loss and differ mainly in the advantage estimator. We therefore instantiate

$$\mathcal{L}^{\mathrm{G}}_{\mathrm{base}}(\theta) = \mathcal{L}_{\mathrm{PG}}(\theta; A^{\mathrm{grpo}}) + \beta^{\mathrm{G}}_{\mathrm{ref}}\mathcal{L}_{\text{KL-ref}}(\theta) + \beta^{\mathrm{G}}_{\mathrm{prcp}}\mathcal{L}_{\mathrm{prcp}}(\theta) + \beta^{\mathrm{G}}_{\mathrm{ori}}\mathcal{L}^{\mathrm{ori}}_{\mathrm{ent}}(\theta) + \beta^{\mathrm{G}}_{\mathrm{aug}}\mathcal{L}^{\mathrm{aug}}_{\mathrm{ent}}(\theta), \tag{19}$$

and

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{base}}(\theta) = \mathcal{L}_{\mathrm{PG}}(\theta; A^{\mathrm{dapo}}) + \beta^{\mathrm{D}}_{\mathrm{ref}}\mathcal{L}_{\text{KL-ref}}(\theta) + \beta^{\mathrm{D}}_{\mathrm{prcp}}\mathcal{L}_{\mathrm{prcp}}(\theta) + \beta^{\mathrm{D}}_{\mathrm{ori}}\mathcal{L}^{\mathrm{ori}}_{\mathrm{ent}}(\theta) + \beta^{\mathrm{D}}_{\mathrm{aug}}\mathcal{L}^{\mathrm{aug}}_{\mathrm{ent}}(\theta), \tag{20}$$

where $\mathcal{L}_{\mathrm{prcp}}$ is the PAPO-style perception-separation branch Wang et al. (2025b) computed on the intervention-view condition, and disabled terms are omitted by setting their coefficients to zero according to the configuration. During the Stage-1 warm-up the SIVA valuation loss and outcome-conditioned routing are switched off entirely ($\lambda_{\mathrm{sen}} = \lambda_{\mathrm{inv}} = 0$), and PatchSwap views act as sensitivity-style perturbations through $\mathcal{L}_{\mathrm{prcp}}$ alone ($\beta_{\mathrm{prcp}} = 10^{-2}$). The PAPO branch applies uniform separation during Stage 1. Table 4 shows routing is more effective once that separation has been established.

With SIVA enabled in Stage 2, $\beta_{\mathrm{prcp}} = 0$ so that $\mathcal{L}_{\mathrm{prcp}}$ does not oppose $\mathcal{L}_{\mathrm{inv}}$ on low-drop pairs, and the final objectives become

$$\mathcal{L}^{\mathrm{G}}_{\mathrm{SIVA}}(\theta) = \mathcal{L}_{\mathrm{PG}}(\theta; A^{\mathrm{grpo}}) + \beta^{\mathrm{G}}_{\mathrm{ref}}\mathcal{L}_{\text{KL-ref}}(\theta) + \beta^{\mathrm{G}}_{\mathrm{ori}}\mathcal{L}^{\mathrm{ori}}_{\mathrm{ent}}(\theta) + \beta^{\mathrm{G}}_{\mathrm{aug}}\mathcal{L}^{\mathrm{aug}}_{\mathrm{ent}}(\theta) + \lambda_{\mathrm{sen}}\mathcal{L}_{\mathrm{sen}}(\theta) + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}}(\theta), \tag{21}$$

and

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{SIVA}}(\theta) = \mathcal{L}_{\mathrm{PG}}(\theta; A^{\mathrm{dapo}}) + \beta^{\mathrm{D}}_{\mathrm{ref}}\mathcal{L}_{\text{KL-ref}}(\theta) + \beta^{\mathrm{D}}_{\mathrm{ori}}\mathcal{L}^{\mathrm{ori}}_{\mathrm{ent}}(\theta) + \beta^{\mathrm{D}}_{\mathrm{aug}}\mathcal{L}^{\mathrm{aug}}_{\mathrm{ent}}(\theta) + \lambda_{\mathrm{sen}}\mathcal{L}_{\mathrm{sen}}(\theta) + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}}(\theta). \tag{22}$$

The SIVA-specific alignment term is therefore shared across the two backbone families. Both reuse the same clean rollout responses, both compute the same outcome-conditioned sensitivity and low-drop consistency losses, and both differ only through the backbone configuration summarized in Tables 7 and 8. Throughout the appendix, we keep the same notation as in the main text: λ_{sen} and λ_{inv} denote the sensitivity and invariance coefficients, m_{sen} and m_{inv} denote the corresponding margins, and $\mathcal{L}_{\mathrm{prcp}}$ denotes the PAPO perception branch used during Stage 1.

C Full Hyperparameter Sensitivity Results

C.1 Sweep Setup and Full Results

Table 10 reports the complete per-benchmark sweeps over the candidate-cell ratio η and the minimum source distance $d_{\min}$, the two knobs of the localized PatchSwap construction. The protocol and fixed defaults are specified in Appendix B.3.

All sweeps use the GRPO-3B backbone and the two-stage SIVA-RL schedule of Appendix B. We isolate each construction hyperparameter with a one-factor-at-a-time protocol. When we vary the candidate-cell ratio η, the minimum source distance is fixed at its default $d_{\min} = 0.25$. When we vary $d_{\min}$, the candidate-cell ratio is fixed at $\eta = 0.6$. All remaining settings match the main configuration: the replacement probability ρ, the grid resolution, the routing thresholds, the margins, the audit count $M = 1$, and the loss coefficients are held at the values in Tables 7 and 8. A separate

sweep of M is reported in Appendix B.2. Each cell of Table 10 is therefore a full training run scored under the same Avg@8 protocol as the main results. The numbers are directly comparable to the GRPO-3B baseline of 47.92.

We sweep η and $d_{\min}$ because they control the two independent degrees of freedom of a localized PatchSwap view: how much of the image is eligible for replacement, and how far the replacement content is drawn from. The candidate-cell ratio η sets the spatial extent of the intervention. A small η perturbs a few cells, while a large η exposes most of the grid to replacement. The minimum source distance $d_{\min}$ sets the semantic distance of the pasted content. A small $d_{\min}$ keeps the swap local and visually plausible, while a large $d_{\min}$ imports distant, unrelated structure. The replacement probability ρ and the grid resolution interact with both knobs, but they change the strength of the intervention rather than its spatial or semantic character. We therefore hold ρ and the grid fixed and report their setting only through the main configuration.

Table 10: **Full per-benchmark hyperparameter ablations on Qwen2.5-VL-3B.** We vary one hyperparameter at a time while keeping the others fixed. All results are Avg@8 accuracies (%).

Benchmark	Candidate-cell-ratio sweep η					Minimum source-distance sweep $d_{\min}$				
	0.2	**0.4**	**0.6**	**0.8**	**1.0**	**0.00**	**0.25**	**0.35**	**0.75**	**0.90**
General Reasoning										
Geo3k	32.57	33.13	**34.26**	32.30	33.69	33.96	**34.26**	34.01	32.82	33.88
MathVista	62.90	62.42	62.12	**63.52**	61.29	62.64	62.12	**63.05**	61.12	62.71
We-Math	**63.00**	62.28	61.92	60.94	62.69	61.28	61.92	**62.05**	60.83	61.82
MMK12	60.56	59.77	**64.63**	60.67	59.97	60.38	**64.63**	61.52	60.52	59.55
MathVerse	58.47	58.90	59.66	57.63	**60.09**	59.00	59.66	59.68	**59.74**	59.58
General AVG	55.50	55.30	**56.52**	55.01	55.55	55.45	**56.52**	56.06	55.01	55.51
Vision-Dependent Reasoning										
LogicVista	41.08	**42.67**	42.42	40.80	41.58	42.20	**42.42**	41.95	42.03	40.97
Counting	64.25	67.12	**75.19**	68.19	75.12	**75.38**	75.19	67.69	71.12	68.94
MMMU-Pro	27.75	28.27	**29.12**	28.59	28.06	28.30	**29.12**	28.27	27.64	28.25
MathVerse$_V$	55.44	55.86	55.68	54.66	**56.61**	55.66	55.68	**57.10**	56.16	56.14
Vision AVG	47.13	48.48	**50.60**	48.06	50.34	50.38	**50.60**	48.75	49.24	48.57
Overall AVG	51.78	52.27	**53.89**	51.92	53.23	53.20	**53.89**	52.81	52.44	52.43

Read column by column, Table 10 shows that no single benchmark drives the aggregate trends. The overall average stays within a 2.1-point band across the entire η sweep and within a 1.5-point band across the $d_{\min}$ sweep. The per-benchmark optima are also dispersed: the best η for Geo3k, MathVista, and MathVerse differ from one another, and the same holds for $d_{\min}$. The defaults $\eta = 0.6$, $d_{\min} = 0.25$ are therefore a robust compromise across benchmarks rather than a setting that suits any one of them. The overall average stays flat, while the vision-dependent group moves more under the same knobs (Appendix C.2).

C.2 Findings

Three trends stand out.

C.2.1 Both knobs have an interior optimum

Neither hyperparameter improves monotonically. The candidate-cell ratio η peaks at the default $\eta = 0.6$ (53.89 overall), drops to 51.92 at $\eta = 0.8$, and then partially recovers to 53.23 at $\eta = 1.0$. The minimum source distance $d_{\min}$ shows a cleaner interior peak: accuracy is highest at $d_{\min} = 0.25$ (53.89) and falls steadily as sources are forced farther away (52.81, 52.44, and 52.43 at 0.35, 0.75, and 0.90). Both profiles follow from the construction rationale in Section 3.2: a useful intervention must be hard enough to remove local visual evidence yet mild enough to keep the image in-distribution. A too-small $d_{\min}$ copies a near-duplicate neighbor and barely perturbs the evidence. A too-large $d_{\min}$ pastes an unrelated distant patch and drifts toward arbitrary corruption. The intermediate setting balances these two failure modes. This balanced regime is exactly what SIVA-RL is designed to construct.

C.2.2 Construction matters most where vision matters most

Sensitivity to both knobs concentrates in the vision-dependent group. Across the η sweep, the vision-dependent average moves by 3.5 points (47.13 to 50.60), whereas the general average moves by only 1.5 points (55.01 to 56.52). The counting subset is the extreme case and swings by roughly 11 points (64.25 to 75.19) as η varies. General reasoning can lean on textual structure and language priors, so it stays largely insensitive to how the visual intervention is built. Vision-dependent reasoning has no such fallback, so the construction directly governs the available learning signal. This contrast corroborates the central premise of the paper. The visual intervention becomes consequential when the task genuinely depends on visual evidence. The pattern mirrors our main results, where SIVA-RL gains are largest on the vision-dependent subset.

C.2.3 Gains persist across construction settings

Every configuration in Table 10 stays well above the GRPO-3B baseline of 47.92. Even the weakest setting (51.78 at $\eta = 0.2$) improves the overall average by 3.9 points. Two configurations are especially telling. The $\eta = 1.0$ stress test, which falls back to the relaxed source rule of Appendix B, still reaches 53.23. Removing the distance constraint entirely ($d_{\min} = 0.00$) still reaches 53.20, close to the tuned 53.89. The gains persist across these construction settings. The construction hyperparameters mainly control the *supply* of informative clean–intervention pairs. The outcome-conditioned routing then decides *how* each pair is used.

C.3 Discussion and Practical Recommendations

For practitioners, the sweep yields a simple recipe. Start from our defaults $\eta = 0.6$ and $d_{\min} = 0.25$, which the sweep finds to sit in a flat region rather than at a sharp peak, and treat both as coarse knobs rather than quantities that require fine search. The flat interior of both profiles means that any $\eta \in [0.4, 0.6]$ and any $d_{\min} \in [0.25, 0.35]$ lands within roughly one point of the best overall average. A run that must minimize tuning cost can fix both knobs at their defaults and still capture most of the gain. This tolerance matters in multimodal RL, where a single configuration is expensive and a dense grid search over construction hyperparameters is rarely affordable.

Construction hyperparameters set the supply and difficulty of intervention pairs; the router of Section 3.3 assigns their alignment weights. The router reads the observed reward drop of each pair and assigns it to the sensitivity branch or the invariance branch, with intermediate pairs attenuated. The router absorbs much of the variation that a fixed operator would otherwise propagate into training. This absorption is why even the unconstrained and stress-test settings stay close to the tuned optimum. Construction quality raises the ceiling of the available signal. Routing determines how much of that ceiling the policy reaches.

D TRAINING-DYNAMICS DIAGNOSTICS

We provide additional training-dynamics diagnostics for the math-domain runs. The plots show reward and divergence trajectories across both training stages.

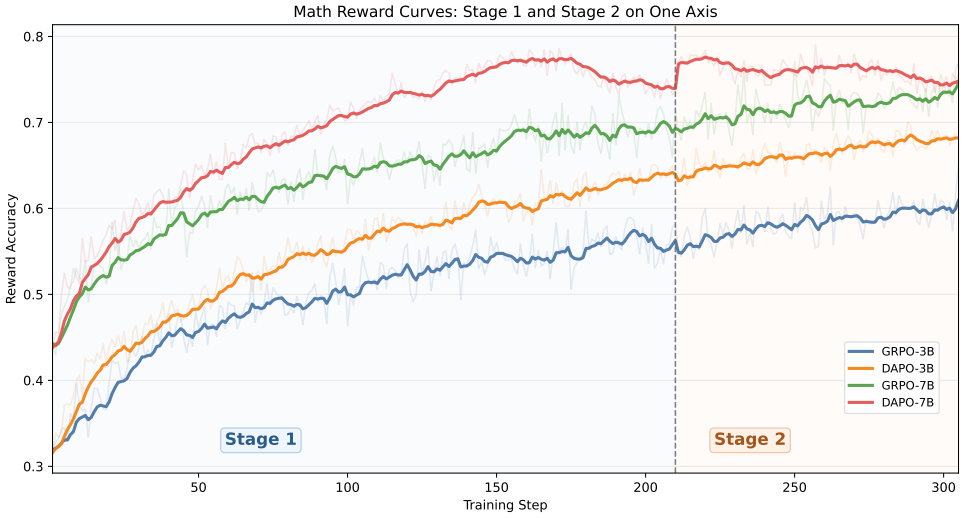


Figure 5: **Math-domain reward-accuracy dynamics across the two-stage schedule.** The dashed vertical line marks the transition from the 200-step PatchSwap/PAPO-style warm-up to the 100-step SIVA-RL stage, with curves shown for GRPO-3B, DAPO-3B, GRPO-7B, and DAPO-7B.

D.1 REWARD–ACCURACY DYNAMICS

The routing stage reignites a plateauing warm-up. Figure 5 preserves the expected scale and backbone ordering throughout training: DAPO-7B leads, followed by GRPO-7B, DAPO-3B, and GRPO-3B. During the 200-step warm-up, reward accuracy rises quickly and then flattens as the uniform PatchSwap perturbation saturates. At the transition to Stage 2 (dashed line), every curve resumes a clear upward slope, and the larger models show a visible step up at the switch. The continued improvement during the SIVA-RL stage indicates that outcome-conditioned routing supplies a learning signal that the uniform warm-up has already exhausted. This dynamic matches the stepwise ablation in Section 4.3: adding routing and invariance alignment on top of the warm-up raises the overall average from 52.64 to 53.89.

The backbone ordering is stable from the first steps to the last. The two 7B models start near 0.44 reward accuracy, while the two 3B models start near 0.32. This 0.10–0.13 scale gap persists through both stages and never closes. Within each scale, the DAPO backbone tracks above its GRPO counterpart for most of training, so the four curves stay ordered as DAPO-7B, GRPO-7B, DAPO-3B, and GRPO-3B without crossing.

The warm-up does more than flatten for the strongest model. DAPO-7B peaks near 0.77 around step 165 and then drifts down to roughly 0.74 by the Stage-1/Stage-2 boundary, a mild regression under prolonged uniform perturbation. The switch to outcome-conditioned routing reverses this drift. DAPO-7B and GRPO-7B both jump discretely at the dashed line, and GRPO-7B then climbs from about 0.69 to 0.74 across Stage 2. The 3B models gain more gradually, while DAPO-7B holds near its post-jump level. Routing therefore helps most where the warm-up had stalled or regressed, not uniformly across configurations.

D.2 ROUTING-BRANCH SEPARATION

Both routing branches stay active and separated. Figure 6 tracks the mean clean-to-intervention divergence D for the two routes during Stage 2. The sensitivity branch holds D at a positive, bounded level for all configurations, consistent with the sensitivity margin m_{sen} that pushes high-

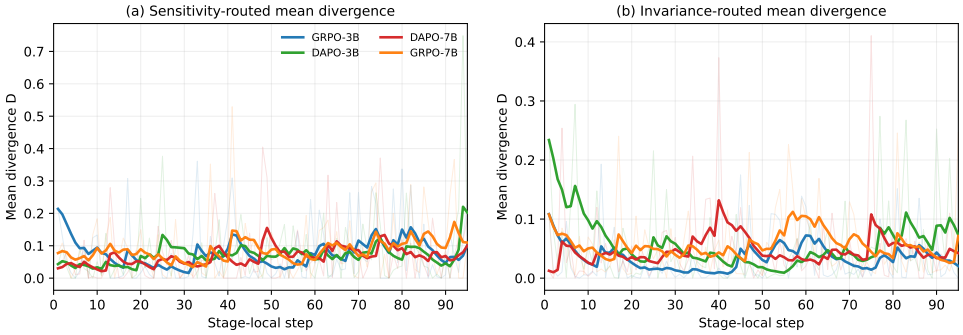


Figure 6: **Intervention-view divergence dynamics during the math-domain SIVA-RL stage.** The two panels report mean divergence D for sensitivity-routed and invariance-routed pairs, respectively. Each panel reports how the re-scored clean responses behave under PatchSwap interventions across backbone and model scales.

drop pairs apart. The invariance branch starts higher in the first few steps and is then driven down toward the small invariance margin m_{inv}, where it stays for the rest of training. The two routes never collapse into a single regime, and sensitivity divergence remains above invariance divergence throughout. This separation is the training-time signature of the dual-direction objective. The separation confirms the mechanistic claim in Section 4.4: the same PatchSwap operator keeps feeding both answer-changing and low-drop pairs rather than degenerating into a uniform perturbation loss.

The two panels share a structure but sit at different scales; note the different y-axis ranges. In the sensitivity panel, the smoothed curves stay in a roughly 0.05–0.15 band for all four configurations. They fluctuate around the sensitivity margin rather than collapsing below it. In the invariance panel, the transient at the start is the clearest signal. DAPO-3B enters Stage 2 near $D = 0.23$ and is pulled down within about fifteen steps to a 0.03–0.05 band, where the other configurations also settle. This early descent is the invariance objective acting on low-drop pairs and tightening clean-to-intervention consistency toward m_{inv}.

After the transient, the two routes stay separated by a clear gap. Sensitivity divergence holds around 0.05–0.15 while invariance divergence stays around 0.02–0.06, so the sensitivity band sits roughly two to three times higher throughout Stage 2. Late in training the sensitivity curves drift upward, with GRPO-7B and DAPO-3B reaching about 0.2 near the final steps. This upward drift shows that PatchSwap keeps producing strongly answer-changing pairs rather than running out of them. The faint unsmoothed traces spike as high as 0.4–0.5 in both panels. These bursts are individual high-divergence batches that the soft routing weights absorb without destabilizing the smoothed trend.

D.3 Magnitude of the Auxiliary Objectives

Because $D_{i,g} \geq 0$, the sensitivity hinge satisfies $\mathcal{L}_{\text{sen}} \leq m_{\text{sen}}$, so with $\lambda_{\text{sen}} = 0.005$ and $m_{\text{sen}} = 0.05$ the sensitivity term contributes at most 2.5×10^{-4} to the total objective. With $\lambda_{\text{inv}} = 0.001$ and the invariance-route divergence in the observed 0.02–0.06 band, the invariance term contributes on the order of 10^{-5}. Both are well below the reference-KL coefficient of 10^{-2}.

Measured as $1 -$ margin-satisfied rate on bidirectional diagnostic runs, the sensitivity hinge is active on 72.1%, 63.3%, 63.1%, and 67.0% of sensitivity-routed pairs for GRPO-3B, DAPO-3B, GRPO-7B, and DAPO-7B, and the invariance hinge on 66.6%, 88.0%, 72.1%, and 75.9% of invariance-routed pairs. Both hinges are active on most routed pairs.

These fractions quantify hinge activity in the bidirectional diagnostic runs.

D.4 Routing Control

The shuffled-routing control in Table 4 preserves the marginal weight distribution and breaks the per-sample match to Δ_i. That run drops from 53.89 to 50.90. The per-sample correspondence is therefore load-bearing.

Table 11: **Hinge-active fractions of the SIVA-RL auxiliary branches during Stage 2.** The active fraction is the share of routed pairs whose hinge is non-zero, i.e. the share on which the branch actually supplies gradient, computed as $1 -$ margin-satisfied rate. Values are Stage-2 means read from bidirectional diagnostic runs.

Configuration	Sensitivity hinge active	Invariance hinge active
GRPO-3B	72.1%	66.6%
DAPO-3B	63.3%	88.0%
GRPO-7B	63.1%	72.1%
DAPO-7B	67.0%	75.9%

E Auxiliary Medical-Domain Accuracy

E.1 Setup and Results

We evaluate the full training recipe on public medical VQA benchmarks and compare it with pre-trained instruct checkpoints. A matched medical RL baseline is unavailable. The VQA-RAD, SLAKE, and PathVQA base-model entries are our own Avg@8 evaluations of the official Qwen2.5-VL-3B and Qwen2.5-VL-7B checkpoints.

E.2 Evaluation Scope

This evaluation measures answer accuracy; visual evidence use was not separately assessed.

F Qualitative Case Studies

We present qualitative cases to clarify what the quantitative gains in the main paper mean at the sample level. The five categories are:

1. Visual recovery
2. Misleading black-image correctness
3. PatchSwap-only instability
4. Local geometric stability
5. Residual failure

This structure matters because SIVA-RL does not assume that every visual intervention induces the same training signal. SIVA-RL instead interprets a perturbation through the paired outcome change: large drops provide sensitivity supervision, low-drop pairs provide invariance supervision, and persistent failures reveal limitations outside the current training signal.

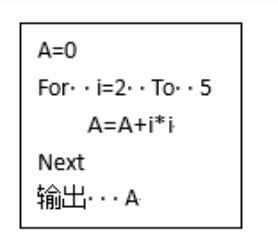
On the diagnostic benchmark probes, random mask, PatchSwap, and black-image interventions all produce mixtures of outcome-drop and outcome-stable behavior. The perturbation operator alone therefore cannot determine whether a paired example should be separated or kept consistent. The examples below illustrate this heterogeneity and show one failure mode that SIVA-RL does not eliminate.

F.1 Visual Recovery

Table 12: **Auxiliary medical VQA accuracy (Avg@8).** Deltas are relative to the pretrained instruct checkpoints.

Method	In-Domain Datasets			Out-of-Domain Datasets	Overall
	VQA-RAD	SLAKE	PathVQA	MedXpertQA	AVG
Proprietary Models					
Gemini-2.0-flash-lite	59.4	73.1	64.9	–	–
GPT-4.1-Nano	61.8	73.1	70.6	–	–
GPT-4o	63.9	71.6	75.9	–	–
General-Purpose Multimodal VLMs					
Qwen-VL-Chat	47.0	56.0	55.1	–	–
Yi-VL-34B	53.0	58.9	47.3	–	–
LLaVA-v1.6-7B	52.6	57.9	47.9	–	–
LLaVA-v1.6-13B	55.8	58.9	51.9	–	–
LLaVA-v1.6-34B	58.6	67.3	59.1	–	–
LLaVA-v1.5-LLaMA3-8B	54.2	59.4	54.1	–	–
Medical Multimodal VLMs					
Med-Flamingo	45.4	43.5	54.7	22.1	41.4
RadFM	50.6	34.6	38.7	<u>23.4</u>	36.8
LLaVA-Med-7B	51.4	48.6	56.8	20.8	44.4
LLaVA_Med-LLaMA3-8B	60.2	61.2	54.5	–	–
PubMedVision-8B	63.8	74.5	59.9	–	–
HuatuoGPT-Vision-34B	68.1	76.9	63.5	22.1	57.7
MedVLThinker-7B	63.7	67.8	65.2	20.9	54.4
CAPO-7B	**78.5**	<u>79.1</u>	68.9	–	–
Medical Agentic Systems					
MedAgents	65.6	67.9	63.2	–	–
MDAgents	66.8	68.2	65.4	–	–
AFlow	67.3	68.9	66.4	–	–
MMedAgent-RL-7B	<u>71.5</u>	76.2	72.3	–	–
Our Models					
Qwen2.5-VL-3B (base)	54.6	57.9	49.2	20.7	45.6
DAPO + SIVA-RL	62.7 (+8.1)	68.1 (+10.2)	66.5 (+17.3)	23.0 (+2.3)	55.1 (+9.5)
GRPO + SIVA-RL	63.0 (+8.4)	68.8 (+10.9)	65.2 (+16.0)	**24.5** (+3.8)	55.4 (+9.8)
Qwen2.5-VL-7B (base)	62.1	62.6	54.9	20.1	49.9
GRPO + SIVA-RL	67.1 (+5.0)	**79.9** (+17.3)	**80.7** (+25.8)	22.3 (+2.2)	**62.5** (+12.6)
DAPO + SIVA-RL	67.4 (+5.3)	72.7 (+10.1)	<u>77.4</u> (+22.5)	22.3 (+2.2)	<u>60.0</u> (+10.1)

Case 1: Visual Recovery



clean image

Task

Benchmark. MMK12, example 343. **Gold.** D.
Question. After running the visual program, what is the output value?

GRPO Baseline: wrong visual reading

Prediction. B on the clean image.
Saved response excerpt. The model reads the update as "$A = A + i^2 \cdot i$" and then accumulates an incorrect sequence.

SIVA-RL: recovers the rendered loop

Prediction. D on the clean image.
Saved response excerpt. The model follows the loop from $i = 2$ to $i = 5$ with "$A = A + i \cdot i$", producing the correct option.

Interpretation

Outcome pattern.

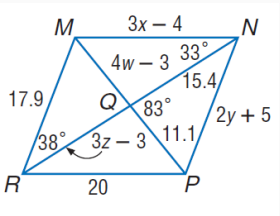
	Clean	Black	PatchSwap
GRPO	B	C	A
SIVA-RL	D	C	D

This case illustrates visual recovery rather than memorized answer correction. The black-image probe remains incorrect for SIVA-RL, so the correct clean prediction cannot be attributed to a text-only prior. The PatchSwap result stays correct because the local replacement does not destroy the program structure needed to read the loop. In SIVA-RL terms, this behavior is exactly the goal: preserve the response under low-impact local perturbations while still relying on the visible program when the image is available.

Figure 7: **Visual recovery case.** SIVA-RL corrects a baseline OCR/program-reading error by using the rendered loop, while the black-image response remains incorrect.

F.2 Misleading Black-Image Correctness

Case 2: Misleading Black-Image Correctness



clean image



black-image probe

Task

Benchmark. MathVerse, example 1328. **Gold.** D.
Question. Use parallelogram $MNPR$ to find $m\angle RMN$. Choices: A:33, B:38, C:71, D:109.

GRPO Baseline: black-correct but clean-failed

Predictions. Clean: no parse; black: D.
Saved response excerpt. On the clean diagram, it applies generic parallelogram facts but does not produce a valid final answer; on the black image, it guesses the correct option from text alone.

SIVA-RL: correct only with visible geometry

Predictions. Clean: D; black: B.
Saved response excerpt. The clean response uses adjacent/supplementary angle constraints in the visible parallelogram and selects D.

Interpretation

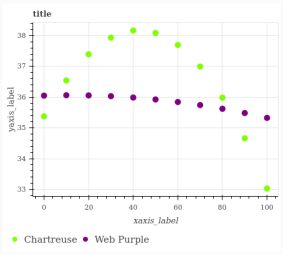
This case shows why black-image correctness is not reliable evidence of invariance. The baseline becomes correct only after the diagram is removed. This behavior is more consistent with answer-prior guessing than with robust visual reasoning. The clean-validity gate excludes this pair because its clean prediction is invalid.

Figure 8: **Misleading black-image correctness.** The baseline's correct black-image answer is not evidence of robust visual reasoning; it masks a clean-image failure.

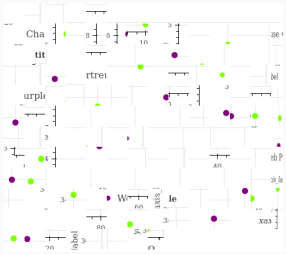
F.3 PatchSwap-Only Instability (Chart)

This is the MathVista example in Figure 1a, under the distance-constrained PatchSwap used in the audit and in training.

Case 3: PatchSwap Diagnostic (Chart)



clean image



training PatchSwap

Task

Benchmark. MathVista, example 324. **Gold.** no.
Question. Does Web Purple have the maximum area under the curve?

GRPO Baseline: PatchSwap-only failure

Predictions. Clean: no; PatchSwap: yes.
Saved response excerpt. The clean response compares Chartreuse and Web Purple correctly, but the PatchSwap response concludes that Web Purple is higher.

SIVA-RL: stable under local replacement

Predictions. Clean: no; PatchSwap: no.
Saved response excerpt. The response continues to compare the curve values and area under the localized intervention, keeping the correct answer.

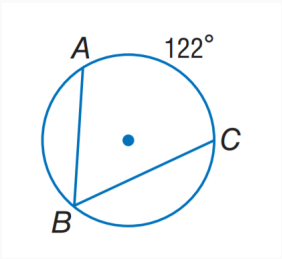
Interpretation

Both models are correct on the clean image and on the black probe. Only PatchSwap exposes the baseline instability. Figure 1a records the same flip. The final model keeps the correct answer under that PatchSwap.

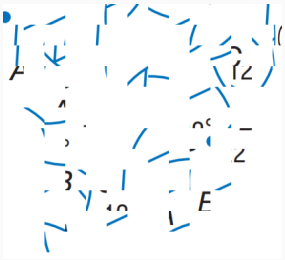
Figure 9: **PatchSwap diagnostic case.** The same MathVista_324 example as Figure 1a, under distance-constrained PatchSwap. A localized within-image replacement reveals a baseline stability gap that the black-image probe does not reveal.

F.4 Local Geometric Stability

Case 4: PatchSwap Diagnostic (Geometry)



clean image



PatchSwap intervention

Task

Benchmark. MathVerse, example 1331. **Gold.** C.
Question. Find $m\angle B$. Choices: A:29, B:58, C:61, D:122.

GRPO Baseline: local geometry break

Predictions. Clean: C; PatchSwap: B.
Saved response excerpt. The clean response uses the inscribed-angle relation, but the PatchSwap response treats the figure as a generic triangle and switches to B.

SIVA-RL: consistent final decision

Predictions. Clean: C; PatchSwap: C.
Saved response excerpt. The model keeps the correct final option under the localized intervention instead of changing to the triangle-based answer.

Interpretation

This additional geometry case mirrors the chart case in a symbolic diagram. Clean accuracy alone does not reveal whether the model has learned a stable geometric relation or merely followed a fragile visual cue. PatchSwap exposes the baseline's dependence on local diagram fragments: the answer changes from the correct arc/angle relation to a generic triangle answer. The final model exhibits improved stability on this case.

Figure 10: **Additional PatchSwap case on geometry.** The localized intervention uncovers a baseline error that is absent on the clean image.

F.5 Residual Failure: Dense Multi-Instance Counting

Case 5: Residual Failure: Dense Multi-Instance Counting



clean image



PatchSwap intervention



black-image probe

Task

Benchmark. Super-CLEVR counting, example 62. **Gold.** 8. **Question.** How many different items are there in the image?

SIVA-RL Clean: near miss

Prediction. 9 on the clean image.
Saved response excerpt. The model enumerates vehicle-like objects but double-counts or over-separates visually similar instances, producing one extra item.

Perturbed Views: unstable counts

Predictions. PatchSwap: 15; black: 0; random mask: 20.
Saved response excerpt. Under PatchSwap and random mask, the model falls back to a generic item-counting strategy and over-counts fragmented or repeated visual evidence.

Interpretation

In this dense counting scene, the model predicts nine objects against a gold count of eight. The clean prediction is already wrong, and all three perturbed views remain wrong, so the example provides little reliable positive supervision for either sensitivity or invariance. Such persistent-error cases motivate stronger object-level counting supervision or region-level credit assignment beyond the current answer-level paired-outcome signal.

Figure 11: **Residual failure on dense multi-instance counting.** Even after SIVA-RL, the model can over-count similar objects on the clean image and become more unstable under local or global visual perturbations.