MULTIMODAL RLVR WITH ROUTED DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) improves multimodal reasoning, yet a shared outcome-derived advantage cannot distinguish errors in reading an image from errors in reasoning over it. These functions recur within the same response, making both the location and the source of expert feedback consequential. We introduce **RLRD** (**RL**VR with **R**outed **D**istillation), which allocates token-level feedback from frozen perception and reasoning experts with different post-training backgrounds. Image-perturbation sensitivity and predictive entropy serve as routing proxies, selecting positions along each on-policy response for the corresponding expert. The routed teacher–student log-probability gaps enter an existing sign-preserving advantage reweighting rule, adjusting local update magnitudes while retaining the direction set by the verifier reward. Across six vision–language benchmarks and Qwen2.5-VL-3B/7B students, RLRD achieves the highest averages in both perception-oriented and reasoning-oriented groups among the evaluated methods, improving over the strongest non-RLRD baseline in each group by 1.39–2.02 percentage points. Controlled ablations show gains over random routing at the same teacher-token budget and over swapped or averaged expert assignments. Together, these results support selecting both feedback positions and expert sources to improve learning from interleaved multimodal responses.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has emerged as an effective post-training recipe for multimodal large language models (MLLMs). By reducing a generated answer to a scalar reward through an automatic verifier, methods such as GRPO improve visual reasoning without dense step-level supervision (Shao et al., 2024b; Yu et al., 2025). However, this trajectory-level reward supervision introduces a fundamental credit-assignment bottleneck. A multimodal reasoning trajectory is not a homogeneous sequence: some tokens serve a *perception* role, reading visual evidence, while others serve a *reasoning* role, performing logical branching (Huang et al., 2025; Lu et al., 2026b) (Figure 1). When the final answer is wrong, an outcome reward cannot tell whether the failure originated in misperception or in invalid reasoning (Jiao et al., 2026; Lu et al., 2026b). The same scalar signal is therefore broadcast across functionally distinct token positions, providing little guidance about *where* the trajectory should be corrected.

This credit-assignment problem is harder because perception and reasoning do not occupy cleanly separated stages; instead, they recur and interact throughout generation (Lu et al., 2026b; Yin et al., 2026). Within a single trajectory, an MLLM may read a number from a diagram, compare it against a textual constraint, revisit a spatial relation, and only then derive the final answer. The functions of individual tokens and the feedback useful at those positions can therefore vary within a response and across examples. A response-level reward, or even a fixed phase-level decomposition, cannot expose these local dependencies. This motivates allocating expert feedback to selected positions *within* naturally generated multimodal responses according to student-side routing signals.

Recent approaches make complementary progress toward this goal. Token-selective RL methods localize important positions using uncertainty, disagreement, or visual sensitivity, yet still optimize them primarily with outcome-derived policy-gradient signals (Wang et al., 2025b; Huang et al., 2025). Token-aware on-policy distillation (OPD) variants instead select informative positions from student uncertainty or teacher–student disagreement, restrict distillation to critical spans, or arbitrate among heterogeneous teachers at the token level (Xu et al., 2026; Wang et al., 2026a; Yin et al.,

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.

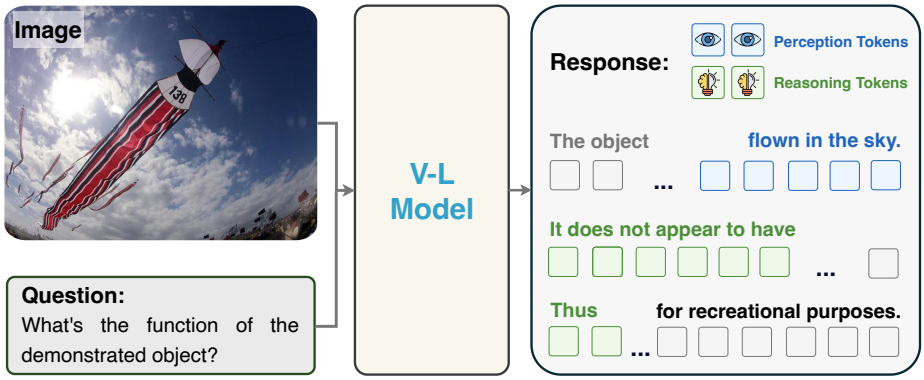


Figure 1: Token roles in a multimodal response. Perception tokens (blue) read visual evidence from the image; reasoning tokens (green) perform logical branching. The two roles are interleaved within a single trajectory rather than partitioned into consecutive stages.

2026). Process-supervised approaches such as StepGRPO also enrich outcome rewards with intermediate reasoning criteria (Zhang et al., 2025). RLSD already uses teacher–student probability gaps to construct sign-preserving token weights for the GRPO advantage (Yang et al., 2026). Our question concerns the allocation of this feedback: *which positions* should receive it and *which expert* should score each selected token when perception and reasoning are interleaved.

We propose **RLRD** (**RL**VR with **R**outed **D**istillation), which couples two routing signals with frozen experts obtained through different post-training objectives and curricula. For each on-policy trajectory, RLRD ranks tokens by *visual sensitivity*, measured through image perturbation, and *predictive uncertainty*, measured by old-policy token entropy. It routes the top-ranked tokens along each dimension to the corresponding frozen perception or reasoning expert. At each routed position, the selected teacher–student log-probability gap enters the clipped, sign-aligned advantage reweighting used by RLSD (Yang et al., 2026). RLRD changes the source and allocation of teacher feedback within that update: two external experts replace a single privileged-information self-teacher, and student-side scores determine the assignment.

We evaluate RLRD with Qwen2.5-VL-3B/7B-Instruct students (Bai et al., 2025) on vision-language benchmarks covering both multimodal perception and complex reasoning. Across both model scales, RLRD improves over outcome-only GRPO, with dual-expert routing providing further gains over single-expert variants. Removing either routing branch and its expert feedback primarily harms its corresponding benchmark group; random routing also underperforms RLRD under the same token budget. These results support the usefulness of the routing configuration for downstream accuracy. The expert-assignment ablation shows a larger penalty for swapping experts than the gain from matched assignment over averaging their logits.

Our main contributions are as follows:

- We formulate **token-level feedback allocation** in multimodal RLVR as the joint selection of positions and expert sources, motivated by interleaved perception and reasoning roles that a shared outcome-derived advantage cannot distinguish.
- We propose **RLRD**, which uses **visual sensitivity and predictive uncertainty** to route tokens to frozen perception and reasoning experts. Their likelihood gaps enter an existing sign-preserving GRPO reweighting rule, refining update magnitudes while retaining the verifier-derived direction.
- Experiments on **six multimodal benchmarks and two student scales** yield the highest averages in both benchmark groups among the evaluated methods. Controlled routing and expert-assignment ablations assess the contribution of selecting feedback positions and sources.

2 Related Work

2.1 Multimodal RLVR and On-Policy Distillation

RLVR learns from automatically verifiable outcomes (Shao et al., 2024b; Yu et al., 2025; Liu et al., 2026c; Stojanovski et al., 2026), including multimodal final-answer verification (Wan et al., 2026; Shao et al., 2024a). On-policy distillation (OPD) supervises student-visited states (Agarwal et al., 2024; Xu et al., 2025). Variants select tokens or spans, or arbitrate among teachers (Xu et al., 2026; Wang et al., 2026a; Yin et al., 2026; Liu et al., 2026b; Yuan et al., 2026); other work routes samples between GRPO and self-distillation by correctness (Li et al., 2026a). PTD-PO distills token distributions from a reference model conditioned on privileged spatial and intermediate reasoning hints into failed rollouts (Xiang et al., 2026).

RLSD (Yang et al., 2026) uses outcome rewards to determine update direction and a privileged-information self-teacher to set its magnitude through clipped, sign-preserving advantage reweighting. RLRD adopts this rule and changes feedback allocation: two frozen external experts score selected tokens under the student's context, with student-side routing determining feedback positions and sources.

2.2 Multimodal Token-Level Credit Assignment

Multimodal reasoning combines visual grounding and inference (Xiong et al., 2025; Tong et al., 2024; Jiao et al., 2026; Li et al., 2026b). R1-VL's StepGRPO adds process rewards through key-step matching and reasoning-structure criteria (Zhang et al., 2025). Perception-aware methods identify image-dependent positions through perturbations or masking (Wang et al., 2025b; Huang et al., 2025; Xiao et al., 2025; Lu et al., 2026a; Ye et al., 2026); other methods use uncertainty or disagreement (Lu et al., 2026b; Miao et al., 2026). Visual relevance also guides token pruning and reinforcement fine-tuning (Jiang et al., 2025; Liu et al., 2025b; Hu et al., 2026). DUPL combines image-perturbation divergence and policy entropy to shape advantages across original and perturbed branches (Liu et al., 2025a); MSSR uses entropy-based advantage shaping for single-rollout RLVR (Liu et al., 2026a). RLRD instead uses these types of student-side signals to allocate feedback from external experts.

Predictive uncertainty and downstream influence are distinct. APPO combines entropy with a continuation-based score for branching in tool-using agents (Wang et al., 2026b). RLRD's ablations assess entropy as an allocation proxy alongside image sensitivity; they do not establish that it identifies causally decisive tokens.

3 Preliminaries

Multimodal RLVR. In multimodal RLVR, a prompt $x = (I, q)$ consists of an image I and a query q, with reference answer $y^\star$. The student policy samples a response trajectory and induces a context at each generation step:

$$\tau = (u_1, \dots, u_T) \sim \pi_\theta(\cdot \mid x), \qquad h_t = (x, u_{<t}), \tag{1}$$

where $m_t \in \{0, 1\}$ marks valid, non-padding response positions. A verifier $\mathrm{V}(\cdot)$ extracts the predicted answer $\hat{y} = \mathrm{V}(\tau)$ and returns the scalar outcome reward $R(\tau) = r(\hat{y}, y^\star)$. RLVR maximizes $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$. GRPO samples a group of n trajectories $\{\tau_i\}_{i=1}^n$ for the same prompt and assigns every valid token in trajectory τ_i the group-normalized advantage

$$\widehat{A}_t^{(i)} = \frac{R(\tau_i) - \bar{R}_g}{\sigma_{R,g} + \delta}, \quad t : m_t^{(i)} = 1, \tag{2}$$

where $\bar{R}_g$ and $\sigma_{R,g}$ are the group reward mean and standard deviation, and δ is a small constant. With the probability ratio

$$r_t(\theta) = \frac{\pi_\theta(u_t \mid h_t)}{\pi_{\theta_{\mathrm{old}}}(u_t \mid h_t)}, \tag{3}$$

the actor is optimized over valid response positions using the clipped surrogate

$$\mathcal{J}_{\mathrm{GRPO}} = \mathbb{E} \sum_{t:m_t=1} \min\Big(r_t(\theta)\widehat{A}_t, \mathrm{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\widehat{A}_t\Big). \tag{4}$$

The expectation averages over sampled responses, while the inner sum includes all valid tokens in each response. RLRD uses the same reduction. Because $\widehat{A}_t^{(i)}$ is constant across all valid positions of a trajectory, the update conveys whether the response succeeded but not which token caused its success or failure.

Token-role heterogeneity. Tokens in a multimodal response can serve perception, reasoning, or other answer-forming functions. These functions are interleaved, and a token may involve both perception and reasoning. RLRD allocates expert feedback using visual sensitivity and predictive uncertainty through routing masks $\mathcal{M}_v, \mathcal{M}_r$ (Section 4.1). Its operational assignment $\rho_t \in \{v, r, \varnothing\}$ selects the perception expert, the reasoning expert, or no expert, respectively. Each routed token receives feedback from one expert even when its functions overlap.

On-policy distillation. On-policy distillation (OPD) complements outcome supervision by evaluating a teacher π_T on prefixes visited by the current student policy. Given a student-sampled context $h_t = (x, u_{<t})$, let

$$p_t(\cdot) = \pi_\theta(\cdot \mid h_t), \qquad q_t(\cdot) = \pi_T(\cdot \mid h_t) \tag{5}$$

be their next-token distributions over vocabulary $\mathcal{V}$. The sequence-level reverse-KL objective decomposes into token-level terms on student-visited prefixes:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(p_t \| q_t)\right]. \tag{6}$$

Vanilla response-wide OPD supplies dense local guidance, but applying the same teacher objective at every response position treats all student-visited states as equally informative and ignores their distinct functional roles. A lightweight alternative retains only the teacher–student gap on the sampled token,

$$g_t^T = \log \pi_T(u_t \mid h_t) - \log \pi_\theta(u_t \mid h_t), \tag{7}$$

which is the realized log-density difference at u_t. It requires retaining one sampled-token teacher log-probability per scored position rather than the full teacher distribution. In Section 4.2, RLRD instantiates this quantity with role-specific gaps Δ_t^v and Δ_t^r, selects them according to the assigned routes, and incorporates the selected correction into the GRPO advantage. Thus, RLVR provides the global outcome signal, while routed expert feedback supplies local likelihood information at selected positions.

4 METHOD

As illustrated in Figure 2, **RLRD** has three stages: (i) *interleaved token routing* selects positions using visual sensitivity and predictive uncertainty; (ii) *routed expert feedback* obtains a sampled-token likelihood gap from the assigned frozen expert; and (iii) *routed advantage reweighting* integrates that gap into GRPO. Sections 4.1–4.3 describe these stages.

4.1 INTERLEAVED TOKEN ROUTING

Rather than partitioning the response into a vision prefix and a reasoning suffix, RLRD defines role-specific token sets over the shared index set $\{1, \ldots, T\}$. Let $m_t \in \{0, 1\}$ denote the valid response mask, where $m_t = 1$ for non-padding response tokens that participate in the actor loss and $m_t = 0$ otherwise.

Perception score. We construct a perturbed image $\widetilde{I} = \mathcal{A}(I)$ by random patch blackening: the image is partitioned into patches of size $P \times P$ pixels, and each patch is independently zeroed with probability p_b. The old policy is then re-scored on the perturbed context $\widetilde{h}_t = (\widetilde{I}, q, u_{<t})$ for the *same* sampled token u_t:

$$\ell_t = \log \pi_{\theta_{\mathrm{old}}}(u_t \mid h_t), \qquad \widetilde{\ell}_t = \log \pi_{\theta_{\mathrm{old}}}(u_t \mid \widetilde{h}_t). \tag{8}$$

Let $\bar{\delta}_t = \widetilde{\ell}_t - \ell_t$ be the raw log-likelihood shift and $\delta_t = \mathrm{clip}(\bar{\delta}_t, -\delta_{\max}, \delta_{\max})$ its clipped value. We use a clipped form of a non-negative KL estimator as the visual-sensitivity score:

$$d_t^v = \mathrm{clip}\big(e^{\delta_t} - \delta_t - 1,\, 0,\, d_{\max}\big), \tag{9}$$

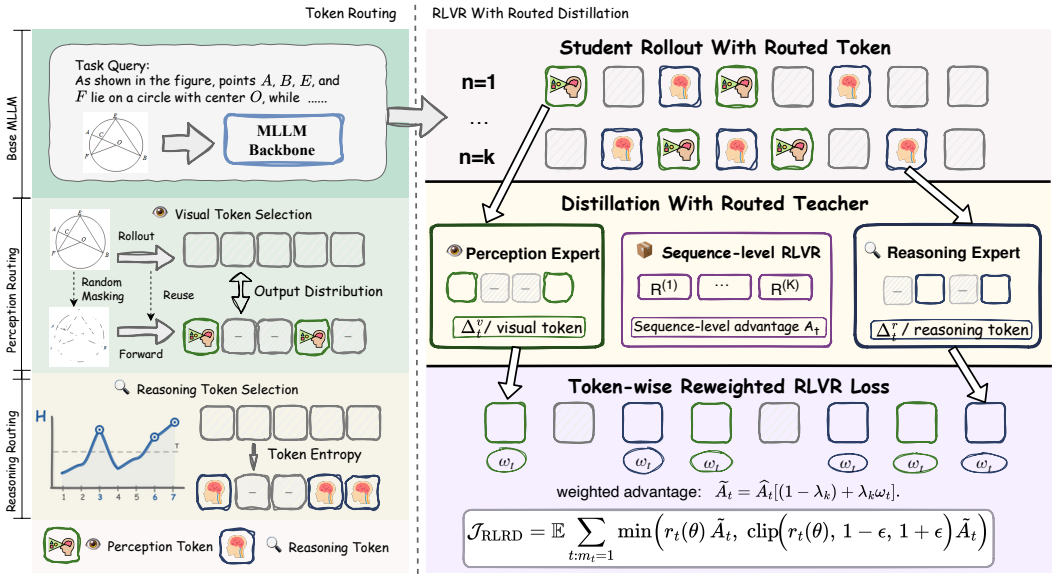


Figure 2: **RLRD overview.** The old policy ranks student-rollout tokens by visual sensitivity d_t^v and reasoning uncertainty s_t^r, forming budgeted perception and reasoning routes with perception priority on overlap. Assigned experts provide sampled-token gaps at routed positions, yielding clipped, sign-aligned corrections to the GRPO advantage. Unrouted positions retain the original GRPO update.

The underlying unclipped term $e^{\bar{\delta}_t} - \bar{\delta}_t - 1$ is non-negative and, when u_t is sampled from the unperturbed old policy, its expectation equals the KL divergence from the original to the perturbed next-token distribution. We clip the log-likelihood shift and the resulting score for numerical stability. Large d_t^v indicates sensitivity to the image perturbation and serves as a proxy for allocating perception-expert feedback. The Appendix derives the unclipped KL identity and describes the practical clipped surrogate.

Reasoning score. For reasoning routing, we use the old-policy token entropy

$$s_t^r = H(\pi_{\theta_{\text{old}}}(\cdot \mid h_t)), \tag{10}$$

which measures uncertainty about the next token. We use it to allocate reasoning-expert feedback. High entropy can also reflect lexical or formatting alternatives, and low entropy can accompany an incorrect decision (Wang et al., 2026b); this routing rule therefore does not identify causal reasoning importance by definition.

Rank-based token selection. Given per-response budgets $p_v, p_r \in (0, 1]$ and the valid-token count $T_m = \sum_t m_t$, we set $K_v = \lceil p_v T_m \rceil$ and $K_r = \lceil p_r T_m \rceil$, and retain the highest-scoring positions within each sampled response:

$$\mathcal{M}_v = \text{TopK}(\{d_t^v\}_{m_t=1},\, K_v), \qquad \mathcal{M}_r = \text{TopK}(\{s_t^r\}_{m_t=1},\, K_r). \tag{11}$$

By construction, $|\mathcal{M}_v| = K_v$ and $|\mathcal{M}_r| = K_r$, so their fractions match p_v and p_r up to ceiling-rounding error. Selection depends only on within-response ranks and is therefore insensitive to score scale. The interleaved masks may overlap; we assign shared positions to perception:

$$\rho_t = \begin{cases} v, & t \in \mathcal{M}_v, \\ r, & t \in \mathcal{M}_r \setminus \mathcal{M}_v, \\ \varnothing, & \text{otherwise}, \end{cases} \tag{12}$$

Let $O = |\mathcal{M}_v \cap \mathcal{M}_r|$. The final perception and reasoning routes contain K_v and $K_r - O$ positions, while their union covers $(K_v + K_r - O)/T_m$ of the response. Perception priority is a fixed design choice, supported here by its higher reported accuracy than reasoning priority at the same budgets (Table 11 in the Appendix). A token may involve both roles even though the router assigns one expert. The label ρ_t records the selected expert, with $\varnothing$ marking an unrouted position.

4.2 Routed Expert Feedback

The routing label ρ_t determines which frozen expert evaluates each selected token: the perception expert π_{ϕ_v} is used when $\rho_t = v$, and the reasoning expert π_{ϕ_r} when $\rho_t = r$. Both experts score the student-sampled token u_t under the visited context h_t. For a routed position with role $s = \rho_t \in \{v, r\}$, we define the sampled-token gap

$$\Delta_t^s = \log \pi_{\phi_s}(u_t \mid h_t) - \log \pi_\theta(u_t \mid h_t). \tag{13}$$

A positive gap means that the expert assigns higher probability to the sampled token than the actor. We recompute the gap using the current θ at each actor update and detach it when constructing the advantage. Detachment blocks gradients through the correction but leaves its numerical value dependent on the current actor. Rollout, routing, and the PPO denominator use θ_{old}. The gap measures relative likelihood, not token correctness. Role weights $\alpha_v, \alpha_r \in [0, 1]$ scale the selected gap:

$$\Delta_t = \begin{cases} \alpha_v \, \Delta_t^v, & \rho_t = v, \\ \alpha_r \, \Delta_t^r, & \rho_t = r, \\ 0, & \rho_t = \varnothing. \end{cases} \tag{14}$$

Each routed position therefore receives exactly one teacher signal. At an unrouted position, $\Delta_t = 0$, so no teacher correction is applied and the token retains the original GRPO update.

4.3 Routed Advantage Reweighting

RLRD inserts the routed gap Δ_t from Eq. (14) into the sign-preserving advantage reweighting of RLSD (Yang et al., 2026). Let $\widehat{A}_t$ be the GRPO advantage. The clipped correction factor is

$$\omega_t = \text{clip}\Big(e^{\,\text{sgn}(\widehat{A}_t)\,\Delta_t}, 1 - \epsilon_w, 1 + \epsilon_w\Big), \tag{15}$$

where $\epsilon_w \in (0, 1)$ bounds the per-token correction. We anneal its contribution across rollout iterations as

$$\lambda_k = \lambda_0 \cdot \max\big(0,\, 1 - (k{-}1)/K_{\text{dec}}\big), \tag{16}$$

where $\lambda_0 \in [0, 1]$, K_{dec} is the decay horizon in rollout iterations, and k indexes rollout iterations from $k{=}1$. Each iteration collects one rollout batch, and λ_k stays fixed across all actor updates on that batch. Thus, $\lambda_1{=}\lambda_0$ and $\lambda_k{=}0$ for $k \geq K_{\text{dec}}{+}1$. The reweighted advantage is

$$\widetilde{A}_t = \widehat{A}_t\big[(1 - \lambda_k) + \lambda_k \, \omega_t\big]. \tag{17}$$

Teacher preference ($\Delta_t > 0$) strengthens positive updates and weakens negative ones; $\Delta_t < 0$ has the opposite effect. The multiplier lies in $[1 - \lambda_k\epsilon_w,\, 1 + \lambda_k\epsilon_w]$, so the advantage sign is preserved and $|\widetilde{A}_t - \widehat{A}_t| \leq \lambda_k\epsilon_w|\widehat{A}_t|$. Interpolation thus limits the deviation from GRPO: $\lambda_k = 1$ gives direct reweighting, and $\lambda_k = 0$ recovers the original advantage. This bound concerns update magnitudes and does not establish superiority over additive corrections. For unrouted positions, $\Delta_t = 0$ gives $\widetilde{A}_t = \widehat{A}_t$; for $k \geq K_{\text{dec}} + 1$, all positions recover the original advantage. Appendix B.4 reports the clipping and decay sensitivity analysis.

The routed gap is detached when constructing the reweighted advantage. The actor then follows the standard GRPO clipped objective (with $r_t(\theta)$ as defined in the Preliminaries):

$$\mathcal{J}_{\text{RLRD}} = \mathbb{E} \sum_{t:m_t=1} \min\Big(r_t(\theta)\widetilde{A}_t, \text{clip}\big(r_t(\theta), 1 - \epsilon, 1 + \epsilon\big)\widetilde{A}_t\Big). \tag{18}$$

This introduces teacher guidance through the GRPO advantage rather than through an auxiliary distillation loss applied uniformly across response positions. The correction rescales outcome-derived advantages without providing an independent token-correctness reward. Appendix D illustrates the four sign combinations.

Training pipeline. Algorithm 1 details one rollout iteration. Under the default schedule, routed correction is active for the first 50 of 284 rollout iterations; subsequent iterations continue GRPO from the policy obtained through this early expert guidance (Appendix A.5).

Table 1: Main results (ViRL39K; Avg@8 accuracy, %). Post-trained entries average three training runs (seeds {42, 43, 44}); group averages include standard deviations. Base reports accuracy only. Full statistics appear in Appendix F.

Method	Perception-oriented				Reasoning-oriented			
	RWQA	MMStar	HalluBench	Avg.	MathVista	LogicVista	MathVerse	Avg.
Qwen2.5-VL-3B								
Base	59.70	52.80	60.04	57.51	57.80	37.28	30.84	41.97
GRPO	60.44	55.73	62.17	59.45 ± 0.43	63.70	39.06	36.29	46.35 ± 0.49
DAPO	59.39	56.58	64.13	60.03 ± 0.46	60.40	39.17	34.80	44.79 ± 0.48
VPPO	61.85	58.50	64.17	61.51 ± 0.38	64.60	43.07	37.53	48.40 ± 0.41
OPD	59.84	56.93	63.92	60.23 ± 0.36	65.70	40.79	35.17	47.22 ± 0.40
ToR	61.96	57.26	64.20	61.14 ± 0.34	65.80	43.04	38.50	49.11 ± 0.39
RLSD	60.94	57.49	64.17	60.87 ± 0.32	66.10	42.79	37.92	48.94 ± 0.36
RLRD-P	62.08	56.87	65.74	61.56 ± 0.29	67.20	42.19	37.15	48.85 ± 0.34
RLRD-R	59.22	57.73	63.12	60.02 ± 0.33	67.50	44.68	38.14	50.11 ± 0.32
RLRD-Dual	**63.40**	**58.93**	**66.37**	$\mathbf{62.90} \pm 0.29$	**67.90**	**45.98**	**39.50**	$\mathbf{51.13} \pm 0.30$
Qwen2.5-VL-7B								
Base	67.52	63.13	65.50	65.38	68.00	44.49	41.11	51.20
GRPO	68.60	64.06	68.80	67.15 ± 0.39	70.40	46.08	44.28	53.59 ± 0.45
DAPO	67.60	61.72	69.03	66.12 ± 0.44	69.10	45.76	42.58	52.48 ± 0.47
VPPO	69.08	65.18	69.40	67.89 ± 0.35	72.30	43.44	42.12	52.62 ± 0.40
OPD	68.48	64.22	67.12	66.61 ± 0.34	71.20	44.12	42.68	52.67 ± 0.39
ToR	68.70	65.42	70.60	68.24 ± 0.31	72.60	46.22	44.30	54.37 ± 0.35
RLSD	68.14	64.88	69.44	67.49 ± 0.30	72.10	45.74	43.94	53.93 ± 0.34
RLRD-P	69.74	65.76	70.60	68.70 ± 0.27	72.50	45.92	44.23	54.22 ± 0.32
RLRD-R	69.75	64.78	69.52	68.02 ± 0.28	73.30	46.84	45.34	55.16 ± 0.31
RLRD-Dual	**71.28**	**67.42**	**71.46**	$\mathbf{70.05} \pm 0.24$	**74.20**	**48.14**	**46.36**	$\mathbf{56.23} \pm 0.28$

5 Experiments

5.1 Experimental Setup

Benchmarks. We use two benchmark groups: *perception-oriented* **MMStar** (Chen et al., 2024), **RealWorldQA**, and **HallusionBench** (Guan et al., 2023); and *reasoning-oriented* **MathVista** (Lu et al., 2023), **LogicVista** (Xiao et al., 2024), and **MathVerse** (Zhang et al., 2024). P-Avg and R-Avg denote their respective means. Both groups require visual understanding and reasoning; the grouping reflects task emphasis. Appendix A.1 details the evaluation.

Baselines. We compare outcome-based **GRPO** (Shao et al., 2024b) and **DAPO** (Yu et al., 2025), token-aware **VPPO** (Huang et al., 2025) and **ToR** (Lu et al., 2026b), and on-policy distillation (**OPD**) (Agarwal et al., 2024). **RLRD-P/R** retain RLRD-Dual's routing and reweighting but use only the perception/reasoning expert on both routes. We also compare **RLSD** (Yang et al., 2026). Appendix A.4 provides implementation details.

Implementation Details. We use Qwen2.5-VL-3B/7B-Instruct students (Bai et al., 2025). Both frozen experts use the 7B backbone with different post-training recipes (Appendix A.2). Students train on ViRL39K (Wang et al., 2025a) under a shared training and validation protocol. Avg@8 averages accuracy over eight samples per example at temperature 1.0, not majority voting or Pass@8. Post-trained main results summarize three independent training runs (seeds 42, 43, and 44); ablations retain single-run estimates. Appendix A details training and evaluation settings.

5.2 Main Results

Table 1 compares the 3B and 7B students. Ablations examine routing branches and token selection (**Abl.1**), feedback coverage and expert assignment (**Abl.2**), and feedback objectives (**Abl.3**). Student-update budgets match, but RLRD adds two post-trained experts and an image-perturbation pass; these comparisons do not establish equal-compute superiority (Appendix A.5).

Dual-expert routing improves both evaluation groups. RLRD-Dual achieves the highest averages in both groups for both students in Table 1. Relative to the strongest non-RLRD baseline in each group, the gains are 1.39 and 2.02 points for the 3B student, and 1.81 and 1.86 points for the 7B student.

Among the single-expert students, RLRD-P has the higher P-Avg and RLRD-R the higher R-Avg at both scales. RLRD-Dual exceeds both on every benchmark.

[Abl.1] Effect of routing branches. On Qwen2.5-VL-3B, each branch removal disables its expert feedback and expands the remaining route to preserve the total routed-token count (Figure 3). Removing perception reduces P-Avg by 4.42 points; removing reasoning reduces R-Avg by 3.37. These effects concern complete branches.

Random routing retains both experts and matches each response's routed-token counts per role. Its roughly 3-point deficit on both groups supports score-based token selection at the same budget.

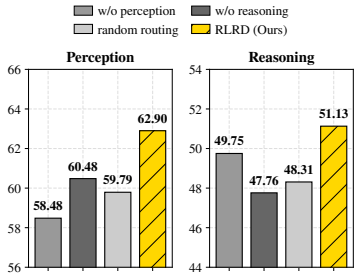


Figure 3: Routing ablations at a fixed teacher-token budget (Qwen2.5-VL-3B; Avg@8 accuracy, %).

[Abl.2] Feedback coverage and expert assignment. For the 3B student (Table 2), RLRD-Avg restricts Dense-Avg's averaged-expert feedback to routed positions, gaining 1.38 P-Avg and 1.27 R-Avg points under the same reweighting. With routing rules and budgets fixed, expert assignment in RLRD-Dual adds 0.78 and 0.42 points over RLRD-Avg. Swapping experts incurs larger drops of 3.56 and 4.01 points relative to RLRD-Dual. These point estimates support selective feedback and expert assignment.

Table 2: Expert-assignment and feedback-coverage ablations (Qwen2.5-VL-3B; Avg@8, %). Avg. denotes averaged expert logits.

Method	Assignment	P-Avg	R-Avg
RLRD-P	Perception $\rightarrow \{v, r\}$	61.56	48.85
RLRD-R	Reasoning $\rightarrow \{v, r\}$	60.02	50.11
Dense-Avg	Avg. $\rightarrow$ all valid tokens	60.74	49.44
RLRD-Avg	Avg. $\rightarrow \{v, r\}$	62.12	50.71
RLRD-Swap	Reasoning $\rightarrow v$, Perception $\rightarrow r$	59.34	47.12
RLRD-Dual	Perception $\rightarrow v$, Reasoning $\rightarrow r$	**62.90**	**51.13**

[Abl.3] Comparison of feedback objectives. Routed OPD retains the same GRPO outcome objective and adds an auxiliary routed KL loss using the same frozen perception and reasoning experts and routing configuration (Appendix A.4). For Qwen2.5-VL-3B, RLRD exceeds this baseline by 2.11 P-Avg and 3.02 R-Avg points (Table 3), favoring advantage reweighting over the auxiliary KL formulation in this setting. Routed OPD also exceeds uniform single-expert OPD by 0.56 and 0.89 points, respectively. Removing sign alignment or jointly removing clipping and annealing also lowers the point estimates; the latter comparison does not separate the two stabilization components.

5.3 Detailed Analysis

We examine routing-role localization and expert likelihood agreement.

[Ana.1] Routing-role localization. We evaluate both routing scores on 400 human-annotated correct responses randomly sampled from the 3B student's test-set outputs, with equal weight per response (Appendix C.1). Figure 4 reports higher AUPRC for visual sensitivity on the perception target (0.912 vs. 0.440) and entropy on the reasoning target (0.824 vs. 0.345). These comparisons support the scores' alignment with the annotated roles on this sample. The downstream routing ablations separately assess their training utility.

[Ana.2] Expert likelihood agreement. On 400 model-annotated correct base-student responses, the perception expert has lower NLL on perception tokens (1.108 vs. 1.611), while the reasoning expert has lower NLL on reasoning tokens (1.247 vs. 1.684; Figure 5). This crossover under shared prefixes supports role-matched likelihood agreement with sampled tokens. Appendix C.2 details the protocol and cell-wise intervals.

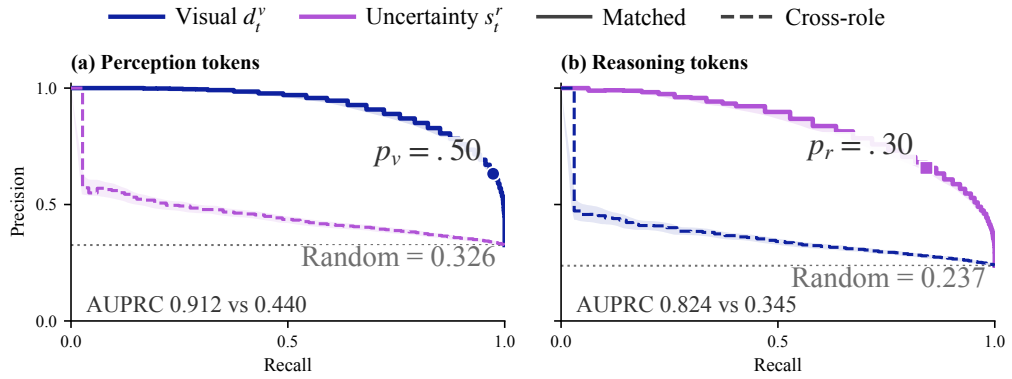


Figure 4: **Routing-role precision–recall curves.** Results on 400 human-annotated correct 3B responses, averaged over responses. Shading shows 95% bootstrap intervals; dotted lines and markers indicate label prevalence and default routing budgets.

Table 3: Feedback-objective comparison (Qwen2.5-VL-3B; ViRL39K; Avg@8 accuracy, %). Routed OPD adds KL to the same GRPO objective, using the same perception and reasoning experts and routing rule as RLRD.

Method	Objective / variant	P-Avg	R-Avg
GRPO	GRPO	59.45	46.35
OPD	Uniform KL	60.23	47.22
Routed OPD	GRPO + routed KL	60.79	48.11
RLRD w/o Sign	Unsigned correction	61.15	50.19
RLRD w/o Stabil.	No clip or annealing	60.65	50.27
RLRD-Dual	Full routed correction	**62.90**	**51.13**

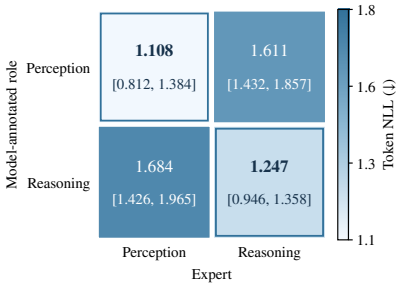


Figure 5: **Expert likelihood by token role.** Prompt-macro NLL on 400 model-annotated responses, with bootstrap 95% intervals. Outlines mark lower NLL.

6 Conclusion

We introduced RLRD, a framework for allocating expert feedback within multimodal RLVR responses. Its central design choice couples the selection of token positions with the selection of feedback sources: image sensitivity and predictive uncertainty route positions to frozen perception and reasoning experts. Their likelihood gaps enter an existing sign-preserving advantage reweighting rule, allowing expert feedback to refine local update magnitudes while the verifier reward determines the update direction. The default schedule applies this feedback during early rollout iterations, then continues outcome-based training from the resulting policy. Comparing guidance at different stages under matched cumulative strength would clarify how its timing affects learning.

Across six benchmarks and two student scales, RLRD achieves the strongest averages in both benchmark groups among the evaluated methods. Routing ablations support selective feedback under a fixed teacher-token budget, while assignment ablations show a larger cost of swapping experts than the benefit of matching them over logit averaging. These findings locate the empirical contribution in how feedback is allocated. Extending this allocation to mixed-role positions and more than two experts is a natural next step beyond the evaluated configuration.

AI USE STATEMENT

Generative AI tools were used for language polishing and assistance with figure preparation. All technical content, experiments, and conclusions were developed and verified by the authors.

ETHICS STATEMENT

This work uses publicly available vision–language models and benchmarks. Human annotation in the routing analysis assigns roles to model-generated responses to benchmark questions. The separate likelihood analysis uses model-generated labels and validity judgments, as described in Appendix C.2. The resulting models may inherit errors and biases from their base models, training data, and experts. Improvements in benchmark accuracy and expert likelihood agreement do not establish factual reliability or safety for use in high-stakes applications.

REPRODUCIBILITY STATEMENT

Section 4 specifies the routing scores, expert assignment, and optimization objective; Algorithm 1 gives the training procedure. Appendix A documents data splits, expert checkpoints, training settings, and evaluation protocols. Appendices B and C describe algorithmic details and the routing and likelihood analyses, including the paired-bootstrap procedure. Appendix E records software versions and checkpoint selection.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

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

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

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025. URL https://arxiv.org/abs/2506.04207v1.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

Ruina Hu, Chen Wang, Lai Wei, Jionghao Bai, Bin Yu, Weiran Huang, Kai Wang, and Yue Wang. Attend to evidence: Evidence-anchored spatial attention supervision for multimodal rlvr. *arXiv preprint arXiv:2605.30912*, 2026.

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*, 2025. URL https://arxiv.org/abs/2510.09285.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 4, pp. 4075–4083, 2025.

Zhengbo Jiao, Shaobo Wang, Zifan Zhang, Wei Wang, Bing Zhao, Hu Wei, and Linfeng Zhang. Credit where it is due: Cross-modality connectivity drives precise reinforcement learning for mllm reasoning. *arXiv preprint arXiv:2602.11455*, 2026.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026a.

Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*, 2026b.

Rui Liu, Dian Yu, Tong Zheng, Runpeng Dai, Zongxia Li, Wenhao Yu, Zhenwen Liang, Linfeng Song, Haitao Mi, Pratap Tokekar, and Dong Yu. Dual-uncertainty guided policy learning for multimodal reasoning. *arXiv preprint arXiv:2510.01444*, 2025a. URL https://arxiv.org/abs/2510.01444.

Rui Liu, Dian Yu, Lei Ke, Haolin Liu, Yujun Zhou, Zhenwen Liang, Haitao Mi, Pratap Tokekar, and Dong Yu. Stable and efficient single-rollout RL for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12018, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Stable_and_Efficient_Single-Rollout_RL_for_Multimodal_Reasoning_CVPR_2026_paper.html.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026b.

Xiaoyuan Liu, Tian Liang, Zhiwei He, Jiahao Xu, Wenxuan Wang, Pinjia He, Zhaopeng Tu, Haitao Mi, and Dong Yu. Trust, but verify: A self-verification approach to reinforcement learning with verifiable rewards. *Advances in Neural Information Processing Systems*, 38:130475–130501, 2026c.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025b.

Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Mingzhu Chen, Jiancan Wu, Kuien Liu, and Xiang Wang. Beyond where to look: Trajectory-guided reinforcement learning for multimodal rlvr. *arXiv preprint arXiv:2603.26126*, 2026a.

Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Guoyin Wang, Jiancan Wu, Xiang Wang, and Xiangnan He. Bridging perception and reasoning: Token reweighting for rlvr in multimodal llms. *arXiv preprint arXiv:2603.25077*, 2026b.

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. *arXiv preprint arXiv:2310.02255*, 2023.

Ziqi Miao, Haonan Jia, Lijun Li, Chen Qian, Yuan Xiong, Wenting Yan, and Jing Shao. Seeing with you: Perception-reasoning coevolution for multimodal reasoning. *arXiv preprint arXiv:2603.28618*, 2026.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, 37:8612–8642, 2024a.

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*, 2024b.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards. *Advances in Neural Information Processing Systems*, 38, 2026.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *Advances in Neural Information Processing Systems*, 38:153676–153713, 2026.

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. *arXiv preprint arXiv:2504.08837*, 2025a.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. Trace: Distilling where it matters via token-routed self on-policy alignment. *arXiv preprint arXiv:2605.10194*, 2026a.

Xucong Wang, Ziyu Ma, Yong Wang, Yuxiang Ji, Shidong Yang, Guanhua Chen, Pengkun Wang, and Xiangxiang Chu. APPO: Agentic procedural policy optimization. *arXiv preprint arXiv:2606.12384*, 2026b. URL https://arxiv.org/abs/2606.12384.

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.

Shizhe Xiang, Ke An, Wenlong Yu, Yue Liu, Jian Luan, Pei Fu, and Qilong Wang. Teaching the way, not the answer: Privileged tutoring distillation for multimodal policy optimization. *arXiv preprint arXiv:2606.07000*, 2026. URL https://arxiv.org/abs/2606.07000.

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.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13618–13628, 2025.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026.

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *arXiv preprint arXiv:2604.01840*, 2026.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-opd: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *arXiv preprint arXiv:2607.02592*, 2026.

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, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

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, 2025. URL https://arxiv.org/abs/2503.12937.

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

APPENDIX CONTENTS

A Experimental and Implementation Details **15**

A.1 Training and Evaluation Data 15

A.2 Expert Checkpoints and Training Data 15

A.3 Prompt and Answer Contract 16

A.4 Common Training Configuration 16

A.5 Hardware Configuration and Relative Cost 19

A.6 Evaluation Metrics and Decoding 19

B Algorithmic Details **20**

B.1 Visual-Score Identity and Clipped Surrogate 20

B.2 Routing Budgets, Overlap, and Sensitivity 21

B.3 Perturbation Sensitivity 21

B.4 Routed Advantage Reweighting 21

C Routing and Expert Analyses **22**

C.1 Routing-Role Localization 22

C.2 Expert Likelihood by Token Role 22

D Sign-Preserving Correction: Mathematical Illustration **23**

E Reproducibility Notes **24**

F Complete Main Results **25**

A Experimental and Implementation Details

This supplement provides the implementation, evaluation, and analysis details needed to reproduce and interpret the results in the main paper. Unless stated otherwise, all student post-training methods use ViRL39K for one epoch and share the rollout group size, global prompt batch, optimizer family, learning-rate schedule, and prompt/response length limits. We report *Avg@8*, the arithmetic mean accuracy over eight stochastic samples per example. Whenever included as an auxiliary diagnostic, *Pass@8* instead measures whether at least one of the eight samples is correct.

A.1 Training and Evaluation Data

ViRL39K contains 38,870 examples before prompt-length filtering. Each example contains an image, a question, a reference answer, and an instruction requesting a tagged reasoning trace followed by a boxed final answer. Filtering removes 2,518 over-length examples and retains 36,352 examples. With a global prompt batch of 128, the common one-epoch protocol therefore contains 284 rollout iterations.

Table 4 summarizes the six benchmarks reported in the main-results table of the paper. The perception-oriented group contains RealWorldQA, MMStar, and HalluBench (HallusionBench), while the reasoning-oriented group contains MathVista, LogicVista, and MathVerse. These groups emphasize different tasks but both involve visual understanding and reasoning. In the main-results table, HalluBench denotes an answer-level Yes/No accuracy proxy rather than the official composite score, and MathVerse denotes the 788-example Vision-Only `testmini` subset. HalluBench evaluation uses the 951 image-containing examples from the `all` split; text-only examples are excluded. LogicVista is stored locally under the `train` split name, but its 448 examples constitute the complete public evaluation set used here. For student post-training, Geo3K test contains 601 examples and is used only for online validation, checkpoint selection, and hyperparameter tuning; it does not contribute to the six-benchmark aggregate. None of the six final evaluation sets is used for post-training updates. We do not perform a separate near-duplicate contamination audit.

Table 4: Benchmarks used in the main results. P/R denote perception-/reasoning-oriented groups. Geo3K test is used for online student validation and is excluded from the six-benchmark aggregate. HalluBench is reported using an answer-level binary-accuracy proxy rather than the official multi-component composite score.

Benchmark	Group	Split	Size	Scoring rule
RealWorldQA	P	test	765	normalized exact answer match in the main table; relaxed match tracked separately
MMStar	P	val	1,500	option-letter exact match
HalluBench	P	all (image subset)	951	answer-level boxed Yes/No accuracy proxy (not the official composite score)
MathVista	R	testmini	1,000	strict boxed-answer equivalence
LogicVista	R	train	448	option-letter exact match on the complete released set
MathVerse	R	Vision-Only testmini	788	strict boxed-answer equivalence on the Vision-Only version

A.2 Expert Checkpoints and Training Data

Both frozen experts derive from Qwen2.5-VL-7B-Instruct. The names *perception expert* and *reasoning expert* denote their post-training emphasis and assigned feedback role. Both retain multimodal capabilities; the names do not imply exclusive abilities or a likelihood advantage on every token of the assigned role.

Perception expert. The perception expert is trained on a dataset combining Geo3K and ViRL39K (Wang et al., 2025a).

Reasoning expert. The reasoning expert uses the released final 7B checkpoint from the staged training procedure of Chen et al. (2025).[1] Training begins with supervised cold start on text reasoning traces from GRAMMAR, followed by multimodal RL and then text-only RL. The

[1] `https://huggingface.co/csfufu/Revisual-R1-final`

authors' released data packages are `csfufu/Grammer_dataset` (listed as 47.3K examples), `csfufu/mmrl` (30,937 examples), and `csfufu/textrl` (32,461 examples), respectively.[2] The GRAMMAR source pool combines mathematical and general reasoning data, including GeoQA, Geometry3K, TabMWP, DeepMath, OpenR1, and NuminaMath. Both RL stages use GRPO with prioritized advantage distillation, which filters and prioritizes sequences by their advantage magnitude. Training combines answer-correctness and efficient-length rewards; the final text-only stage freezes the vision tower (Chen et al., 2025). This curriculum provides the reasoning-oriented training background used for expert assignment.

Both experts are trained before RLRD and remain frozen during student training. The common settings below apply to student post-training.

A.3 Prompt and Answer Contract

For all six benchmarks, we place the benchmark-specific question after the image placeholder and append the same answer-format instruction:

> You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think> </think>` tags. The final answer MUST BE put in `\boxed{}`.

HalluBench additionally requests a Yes/No answer. A format-compliant output has the form

> `<think>` reasoning trace `</think> \boxed{answer}`.

The strict evaluator first extracts the boxed answer. The relaxed evaluator additionally accepts a normalized candidate from an explicit final-answer phrase or the final non-empty line. Unless stated otherwise, the main table uses strict matching for MathVista, MathVerse, and HalluBench, normalized exact match for RealWorldQA, and option-letter exact match for MMStar and LogicVista. Relaxed accuracy is retained only as a secondary diagnostic for datasets that admit string normalization.

A.4 Common Training Configuration

Table 5 lists the settings shared by every student post-training method. The students are Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct. RLRD uses a perception expert and a reasoning expert, both based on Qwen2.5-VL-7B-Instruct.

Table 5: Common settings shared across the main post-training runs.

Item	**Value**
Students	Qwen2.5-VL-3B / Qwen2.5-VL-7B
Experts	perception expert, reasoning expert
Training data	ViRL39K, one epoch
Raw / retained examples / rollout iterations	38,870 / 36,352 / 284
Global prompt batch	128
Rollouts per prompt	8
PPO mini-batch	64 prompt-equivalents (512 responses)
Prompt / response limit	1,024 / 2,048 tokens
Optimizer	AdamW
Learning rate	1×10^{-6}
Betas / weight decay	$(0.9, 0.999)$ / 0.01
Gradient clipping	1.0
Learning-rate schedule	constant, no warmup
Validation set	Geo3K test (601 examples)
Validation protocol	before training and every 5 rollout iterations thereafter, plus the final checkpoint

[2]Dataset links and release sizes: `https://github.com/CSfufu/Revisual-R1`. These describe the released packages, not the stage-specific training-subset sizes.

The launch configuration uses one actor epoch per rollout batch. With $128 \times 8 = 1{,}024$ responses per batch, this gives two optimizer updates of $64 \times 8 = 512$ responses each. GRPO and RLRD use PPO probability-ratio clipping with $\epsilon = 0.2$, distinct from the teacher-weight clipping parameter ϵ_w. Training rollouts use stochastic sampling with temperature 1.0, top-$p = 1.0$, and no top-k cutoff.

Table 6 isolates the method-specific differences. All methods are tuned with the same validation protocol. Routed methods use $(p_v, p_r) = (0.5, 0.3)$ unless the corresponding sensitivity study states otherwise.

Among the routed expert-assignment variants, the scoring rules, token budgets, and overlap policy are fixed. Each variant recomputes its routing masks under its own old policy on its on-policy responses.

Table 6: Method-specific differences under the shared training protocol. Data, policy updates, rollout group size, global batch size, optimizer, learning-rate schedule, and response-length limit are held fixed.

Method	Teacher usage	Method-specific setting
GRPO	None	outcome-only RLVR baseline with no teacher feedback
DAPO	None	decoupled clipping / normalization under the same data and update budget
VPPO	None	self-routed visually aware token emphasis under the same rollout and update budget
ToR	None	self-routed uncertainty emphasis under the same rollout and update budget
OPD	Validation-selected single expert over all response positions	uniform KL-style on-policy distillation; the reported main-table number uses the stronger of the single-teacher runs using the perception expert or the reasoning expert under the common validation protocol
RLSD	Answer-conditioned self-teacher on all valid response tokens	self-distillation using the ground-truth final answer; same sign-preserving reweighting, clipping, and annealing as RLRD
RLRD-P	perception expert on $\{v, r\}$	the single perception expert is used on both routed roles, matching the teacher-assignment ablation in the main paper
RLRD-R	reasoning expert on $\{v, r\}$	the single reasoning expert is used on both routed roles, matching the teacher-assignment ablation in the main paper
RLRD-Dual	perception expert on $\mathcal{M}_v$, reasoning expert on $\mathcal{M}_r \setminus \mathcal{M}_v$	matched dual-teacher routed advantage correction

Tables 5–7 summarize the training settings. RLRD applies the role weight directly to the selected expert–student gap and sets the gap to zero at unrouted positions. The gap is recomputed from the current actor and detached at each actor-loss evaluation; rollout, routing, and the PPO denominator use the frozen old policy. Thus, unrouted tokens retain their original GRPO advantage.

RLSD baseline. RLSD (Yang et al., 2026) uses an answer-conditioned self-teacher. The student generates responses from the image and question; the teacher additionally receives the ground-truth final answer to score the same sampled tokens. Its detached teacher–student log-probability gap reweights all valid response positions using RLRD's sign-preserving rule, correction clip $\epsilon_w = 0.2$, and annealing schedule with $\lambda_0 = 0.5$ and $K_{\text{dec}} = 50$ rollout iterations.

Dense-Avg control. Dense-Avg and RLRD-Avg use the same frozen perception and reasoning experts. We take the arithmetic mean of their vocabulary logits, apply log-softmax, and extract the student-sampled token's log-probability. The gap to the current actor is detached and enters the same GRPO advantage reweighting, clipping, and annealing. Dense-Avg applies feedback to all valid response tokens; RLRD-Avg restricts it to $\mathcal{M}_v \cup \mathcal{M}_r$, leaving other advantages unchanged. This comparison varies feedback coverage and token count; the random-routing ablation separately matches the teacher-token budget.

For a batch of N responses, let f_{it} denote the token-level clipped surrogate and m_{it} its valid-token mask. Both GRPO and RLRD use

$$\text{Reduce}(f, m) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t} m_{it} f_{it}. \tag{19}$$

This averages response-level sums, so each valid token enters with the same batch coefficient and longer responses contribute more terms. Responses are not individually length-normalized. RLRD

Table 7: Configurations for the teacher and objective ablations. When a method admits multiple teacher choices but the main paper reports one result, we select the configuration with the highest validation accuracy under the shared protocol.

Variant	Teacher stream	Coefficient / weights	Reported configuration	Selection note
OPD	one validation-selected expert on all valid response tokens	KL coefficient $= 1.0$	validation-selected single-teacher OPD baseline; per backbone we report the stronger of the runs using the perception expert or the reasoning expert under the shared Geo3K protocol	selected by highest Geo3K strict accuracy
Routed OPD	perception on $\mathcal{M}_v$, reasoning on $\mathcal{M}_r \setminus \mathcal{M}_v$	KL coefficient $= 1.0$	same GRPO outcome objective plus auxiliary routed KL; same routing rule and budgets $(p_v, p_r) = (0.5, 0.3)$ as RLRD-Dual	selected by highest Geo3K strict accuracy
RLRD-Dual	perception expert on $\mathcal{M}_v$, reasoning expert on $\mathcal{M}_r \setminus \mathcal{M}_v$	reported default: $\alpha_v = \alpha_r = 1.0$; $\epsilon_w = 0.2$; $\lambda_0 = 0.5$; $K_{\text{dec}} = 50$	full routed correction	selected by highest Geo3K strict accuracy
RLRD-Avg	averaged teacher logits on both routed roles	equal-weight average	arithmetic mean of the two teacher logits before sampled-token log-prob extraction	all other settings fixed
RLRD-Swap	reasoning expert on $\mathcal{M}_v$, perception expert on $\mathcal{M}_r \setminus \mathcal{M}_v$	same as RLRD-Dual	expert assignment swapped under the same routing rule and budgets	all other settings fixed
RLRD w/o Sign	matched dual teachers	same as RLRD-Dual except no sign alignment	replaces $\operatorname{sgn}(\widehat{A}_t)\Delta_t$ with Δ_t in the exponent	all other settings fixed
RLRD w/o Stabil.	matched dual teachers	no clipping and constant annealing weight	removes $[1-\epsilon_w, 1+\epsilon_w]$ clipping and sets $\lambda_k \equiv \lambda_0 = 0.5$ for all rollout iterations	all other settings fixed
w/o perception	routed ablation	same per-response final routed-token union budget as default	sets the perception route empty and expands reasoning selection to the default union size on each response	used only in Abl. 1
w/o reasoning	routed ablation	same per-response final routed-token union budget as default	sets the reasoning route empty and expands perception selection to the default union size on each response	used only in Abl. 1
random routing	routed ablation	same per-response final routed-token union budget and role counts as default	samples routed positions uniformly from valid response tokens while matching the default per-response perception/reasoning counts after overlap resolution	used only in Abl. 1

changes the advantage inside f_{it} while retaining this reduction; length differences can therefore affect the relative contribution of responses across methods.

Routed OPD retains the same GRPO outcome objective and adds an auxiliary routed KL loss:

$$\mathcal{J}_{\text{Routed OPD}} = \mathcal{J}_{\text{GRPO}} - \beta_{\text{KL}} \mathcal{L}_{\text{routed KL}}, \qquad \beta_{\text{KL}} = 1.0. \tag{20}$$

Here $\mathcal{L}_{\text{routed KL}}$ applies the KL penalty from the student to the assigned expert at selected positions: the perception expert scores $\mathcal{M}_v$, and the reasoning expert scores $\mathcal{M}_r \setminus \mathcal{M}_v$. The expert pair,

routing rule, and budgets match RLRD-Dual. Routed OPD uses the original outcome-derived advantages in its GRPO term, whereas RLRD incorporates the routed expert feedback through advantage reweighting.

A.5 Hardware Configuration and Relative Cost

For a normalized resource comparison, we map every 3B and 7B run to one 8×H100-80GB NVLink node. Single-teacher methods allocate four GPUs to the student actor and four to the teacher server. Dual-teacher RLRD allocates four GPUs to the student actor and two GPUs to each expert. All teacher parameters remain frozen.

Table 8 describes a normalized deployment. RLRD and Routed OPD each add one perturbed-image old-policy pass and two expert scoring streams. In the available implementation, each expert scores complete student responses; routing selects which scores enter the loss. The selected token fraction therefore does not give the fraction of teacher forward FLOPs. This accounting covers student post-training and excludes expert pre-training.

Table 8: Relative resource allocation for ViRL39K post-training on one 8×H100-80GB NVLink node. The table specifies a normalized deployment and its additional forward passes.

Method family	**Allocation and additional computation**
GRPO / DAPO	8 actor GPUs; no external-teacher forward passes (baseline cost).
VPPO / ToR	8 actor GPUs; additional self-routing statistics.
OPD	4 actor and 4 teacher GPUs; single-expert log-probability scoring while distillation is enabled.
Routed OPD	4 actor and 4 teacher-equivalent GPUs; one perturbed-image old-policy pass plus two expert scoring streams.
RLRD-Dual	4 actor, 2 perception expert, and 2 reasoning expert GPUs; one perturbed-image old-policy pass plus two expert scoring streams.

Table 9: Normalized accounting granularity for the reported ViRL39K runs. This table makes explicit the rollout-iteration budget and device partition behind the qualitative cost comparison in Table 8.

Accounting item	**Value**
Total rollout iterations per one-epoch ViRL39K run	284
Rollout iterations with active routed correction	50
Rollout iterations with zero correction coefficient	234
Fraction of rollout iterations with nonzero correction	$50/284 = 17.6\%$
Pre-overlap routed budget	$\lceil 0.5T_m \rceil + \lceil 0.3T_m \rceil$
Exact routed-token union	$\lceil 0.5T_m \rceil + \lceil 0.3T_m \rceil - O$
Overlap rule	perception priority
Single-teacher deployment	4 actor GPUs + 4 teacher GPUs, TP$=1$
Dual-teacher deployment	4 actor GPUs + 2 perception expert GPUs + 2 reasoning expert GPUs, TP$=1$ per teacher pool

The default annealing schedule applies a nonzero correction during the first 50 of 284 rollout iterations. A zero coefficient alone does not eliminate the forward passes; teacher-free tail iterations require explicitly disabling both teacher scoring and routing-only computation. Under the default budgets, the routed-token union is $|\mathcal{M}_v \cup \mathcal{M}_r| = \lceil 0.5T_m \rceil + \lceil 0.3T_m \rceil - O$, where $O = |\mathcal{M}_v \cap \mathcal{M}_r|$. This count describes loss selection rather than teacher forward computation.

A.6 Evaluation Metrics and Decoding

Let $c_{ik} \in \{0, 1\}$ indicate whether stochastic sample k for problem i is correct. For N problems and eight samples per problem, Avg@8 is

$$\text{Avg@8} = \frac{1}{8N} \sum_{i=1}^{N} \sum_{k=1}^{8} c_{ik}, \tag{21}$$

which estimates expected single-sample accuracy under the evaluation decoder. For completeness, oracle Pass@8 is

$$\text{Pass@8} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\max_{1 \le k \le 8} c_{ik} = 1 \right]. \tag{22}$$

Every main-table row uses temperature 1.0, top-$p = 0.95$, top-$k = -1$, a maximum generation length of 1,024, a maximum model length of 8,192, and eight independent samples per example. Prompts, answer extraction, and scoring are identical across methods. We first average correctness over the eight samples for each example and then average over the dataset, which is algebraically equivalent to Eq. (21).

We select checkpoints using strict answer accuracy on Geo3K. Ties are resolved first by validation format accuracy and then by the later checkpoint. The six final benchmarks are evaluated only after selection and never influence model or hyperparameter choice.

B Algorithmic Details

Algorithm 1 gives one complete RLRD rollout iteration. The following subsections detail its routing scores, masks, and advantage corrections.

Algorithm 1 One RLRD rollout iteration.

Require: student π_θ; teachers $\pi_{\phi_v}, \pi_{\phi_r}$; batch $\mathcal{B}$; rollout iteration k.
Freeze rollout policy $\theta_{\text{old}} \leftarrow \theta$.
Initialize routed rollout batch $\mathcal{D} \leftarrow \varnothing$.
for $(x, y^\star) \in \mathcal{B}$ **do**
  Sample $\{\tau_i\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$; get $\widehat{A}_t$.
  Compute routing scores s_t^r, d_t^v under θ_{old}.
  Form $\mathcal{M}_v, \mathcal{M}_r$; assign ρ_t. {Eq. (11), (12)}
  Store expert log-probabilities $\ell_t^T = \log \pi_{\phi_{\rho_t}}(u_t \mid h_t)$ at selected positions.
  Add trajectories, $\widehat{A}_t$, ρ_t, ℓ_t^T, and old-policy log-probabilities to $\mathcal{D}$.
end for
Set λ_k for this rollout iteration and hold it fixed across actor updates.
for each actor update on $\mathcal{D}$ **do**
  Evaluate current actor log-probabilities; form detached Δ_t. {Eq. (14)}
  Compute ω_t, $\widetilde{A}_t$. {Eq. (15), (17)}
  Update θ via $\mathcal{J}_{\text{RLRD}}$ using the frozen old-policy denominator. {Eq. (18)}
end for

B.1 Visual-Score Identity and Clipped Surrogate

Let $p(u) = \pi_{\theta_{\text{old}}}(u \mid h_t)$ and $q(u) = \pi_{\theta_{\text{old}}}(u \mid \widetilde{h}_t)$. For $u \sim p$ and $\delta(u) = \log q(u) - \log p(u)$,

$$\mathbb{E}_{u\sim p}[e^{\delta(u)} - \delta(u) - 1] = \sum_u q(u) - \mathbb{E}_{u\sim p}[\delta(u)] - 1 \tag{23}$$

$$= \mathrm{KL}(p\|q). \tag{24}$$

Therefore, the unclipped quantity underlying the visual-sensitivity score in the main paper is a non-negative, unbiased estimator of the KL divergence from the original to the perturbed next-token distribution. In practice, we clamp the log-likelihood shift to $[-20, 20]$ and the resulting score to $[0, 10]$.

Table 10 reports marginal clipping rates for the selected 3B run during rollout iterations 1–50, when $\lambda_k > 0$. Each rate describes how often a score or weight reaches its corresponding clamp.

Table 10: Average clipping rates during rollout iterations 1–50 of the selected 3B run under the default routed configuration.

Statistic	Rate	Comment
$\lvert\bar{\delta}_t\rvert > 20$	1.9%	raw log-shift clipping
$d_t^v = 10$	2.7%	visual-score upper clipping
$\omega_t = 1 - \epsilon_w$	6.4%	lower weight clip
$\omega_t = 1 + \epsilon_w$	7.1%	upper weight clip

Table 11: Routing-budget and overlap-resolution sensitivity for Qwen2.5-VL-3B. P-Avg and R-Avg average the three perception-oriented and three reasoning-oriented benchmarks, respectively.

Setting	Priority rule	P-Avg	R-Avg	Mean overlap fraction
$(p_v, p_r) = (0.5, 0.3)$	perception priority	**62.90**	**51.13**	0.11
$(p_v, p_r) = (0.5, 0.3)$	reasoning priority	62.08	50.44	0.11
$(p_v, p_r) = (0.5, 0.5)$	perception priority	61.84	50.76	0.19
$(p_v, p_r) = (0.3, 0.5)$	perception priority	60.98	50.61	0.14

The reported saturation rates describe the default ceilings $(\delta_{\max}, d_{\max}) = (20, 10)$. At these ceilings, every token with $|\bar{\delta}_t| > 20$ also satisfies $d_t^v = 10$, since $e^{\delta_t} - \delta_t - 1 > 10$ at $\delta_t = \pm 20$. Thus, log-shift clipping implies visual-score saturation. These rates do not establish how changing the ceilings would affect the selected tokens. Table 12 separately reports the perturbation-parameter sweep.

B.2 Routing Budgets, Overlap, and Sensitivity

For a response containing T_m valid tokens, the perception and reasoning budgets are

$$K_v = \min(T_m, \max(1, \lceil p_v T_m \rceil)), \tag{25}$$

$$K_r = \min(T_m, \max(1, \lceil p_r T_m \rceil)). \tag{26}$$

We form the two masks independently with `TopK`. If they overlap, the main-paper routing rule assigns the shared positions to perception. The final perception route therefore contains K_v positions, the reasoning-only route contains $K_r - O$ positions, and their union contains $K_v + K_r - O$ positions, where $O = |\mathcal{M}_v \cap \mathcal{M}_r|$.

The routing ablations in the main paper match the teacher-token budget separately for every response. Let $U_i = |\mathcal{M}_{v,i} \cup \mathcal{M}_{r,i}|$ be the default routed-token union for response i. The `w/o perception` variant sets $|\mathcal{M}_{v,i}| = 0$ and expands the reasoning route to exactly U_i positions. The `w/o reasoning` variant symmetrically expands the perception route to U_i positions. The random-routing control samples U_i valid positions uniformly and then assigns the first $|\mathcal{M}_{v,i}|$ sampled positions to the perception role and the remaining $|\mathcal{M}_{r,i} \setminus \mathcal{M}_{v,i}|$ positions to the reasoning role. It thus matches both the routed-token union and the final role counts of the default route.

All valid response tokens are eligible, including delimiters and boxed-answer tokens. Formatting and answer tokens receive no special treatment and are selected whenever their routing score falls within the corresponding budget.

Table 11 varies the routing budgets and overlap rule. At the default budgets, perception priority exceeds reasoning priority by 0.82 P-Avg and 0.69 R-Avg points. This comparison motivates the fixed priority in the tested configuration; it does not establish the best expert for every mixed-role token. Larger reasoning budgets do not improve the reported group averages. Soft mixing and dynamic arbitration are not evaluated here.

B.3 Perturbation Sensitivity

The perception router divides the image into non-overlapping $P \times P$ pixel patches and blackens each patch independently with probability p_b. The launch defaults are $P = 14$ and $p_b = 0.5$; boundary patches are clipped to the image extent. Table 12 reports the perception-token AUPRC values, whose range is 2.44 percentage points.

B.4 Routed Advantage Reweighting

The routing label selects the perception expert gap Δ_t^v, the reasoning expert gap Δ_t^r, or zero at an unrouted position. With role weights α_v and α_r,

$$\Delta_t = \begin{cases} \alpha_v \, \Delta_t^v, & t \in \mathcal{M}_v, \\ \alpha_r \, \Delta_t^r, & t \in \mathcal{M}_r \setminus \mathcal{M}_v, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Table 12: Visual-sensitivity routing under different perturbation parameters. The AUPRC range is $0.9240 - 0.8996 = 0.0244$ (2.44 percentage points).

P	p_b	**Perception-token AUPRC**
8	0.5	0.8996
14	0.3	0.9071
14	0.5	0.9118
14	0.7	0.9240

Hence, unrouted positions recover the original GRPO update exactly.

We use the sign-aligned, clipped correction

$$\omega_t = \mathrm{clip}\Big(e^{\,\mathrm{sgn}(\widehat{A}_t)\,\Delta_t}, 1-\epsilon_w, 1+\epsilon_w\Big) \tag{28}$$

and the annealing schedule

$$\lambda_k = \lambda_0 \cdot \max\big(0,\, 1-(k-1)/K_{\mathrm{dec}}\big). \tag{29}$$

Here k indexes rollout iterations, and λ_k is fixed across the actor updates within each iteration. The default reweighting hyperparameters are $\epsilon_w = 0.2$, $\lambda_0 = 0.5$, and $K_{\mathrm{dec}} = 50$ rollout iterations.

Table 13 reports sensitivity to the clipping width and decay horizon around the default. Varying K_{dec} changes both the duration of guidance and the sum of its interpolation coefficients; the sweep does not compare early and late guidance at matched cumulative strength. The joint `w/o Stabil.` ablation in the main paper removes clipping and annealing together, leaving their individual effects unresolved.

Table 13: Sensitivity to the correction clip and annealing horizon K_{dec} (rollout iterations).

Setting	**P-Avg**	**R-Avg**
$\epsilon_w = 0.1$	62.21	50.88
$\epsilon_w = 0.2,\ \lambda_0 = 0.5,\ K_{\mathrm{dec}} = 50$	**62.90**	**51.13**
$\epsilon_w = 0.3$	62.34	50.97
$K_{\mathrm{dec}} = 25$	62.40	50.82
$K_{\mathrm{dec}} = 100$	62.73	51.01

C Routing and Expert Analyses

C.1 Routing-Role Localization

The routing-role analysis uses 400 correct responses randomly sampled from test-set outputs of the 3B student. Human annotations provide the reference perception and reasoning roles for evaluating the visual-sensitivity and entropy scores. Results are averaged over responses, giving each response equal weight.

Table 14 summarizes the AUPRC values in the routing-analysis figure. Visual sensitivity reaches 0.912 on the perception target, compared with 0.440 for entropy; entropy reaches 0.824 on the reasoning target, compared with 0.345 for visual sensitivity. The figure shows 95% bootstrap intervals, label-prevalence baselines, and the default routing-budget markers at $p_v = 0.5$ and $p_r = 0.3$. This analysis measures agreement with annotated roles in correct responses. The model-annotated likelihood analysis below uses a separate sample.

C.2 Expert Likelihood by Token Role

The likelihood analysis uses 400 correct, format-compliant responses from the base Qwen2.5-VL-3B-Instruct model on MathVista, LogicVista, MMStar, and RealWorldQA, generated at temperature 0.7 and evaluation seed 0. Response texts are re-tokenized for scoring because the original sampled token IDs are unavailable. Labels and validity judgments are model-generated, not human

Table 14: Routing-role AUPRC on 400 human-annotated correct 3B responses, averaged over responses. Bold marks the role-matched signal.

Signal	P target	R target
Visual sensitivity d_t^v	**0.912**	0.345
Reasoning uncertainty s_t^r	0.440	**0.824**

annotations. The labels distinguish perception, reasoning, both, and other; this comparison uses the exclusive perception and reasoning roles.

Both experts score the same response set for the comparison in Table 15. We average NLL over each role within a response, then within a prompt, and finally equally across prompts. The same prompts are sampled jointly for all cells in 10,000 bootstrap replicates (seed 20270728); intervals are percentile intervals. Both experts score the same token IDs and prefixes. This analysis uses base-student outputs rather than final RLRD outputs, but its restriction to selected correct responses and model-generated labels limits generalization.

Table 15: Prompt-macro token NLL on 400 model-annotated responses. Brackets show cell-wise 95% intervals from paired prompt bootstrap. Bold marks the lower NLL in each row.

Labelled role	Perception expert NLL	Reasoning expert NLL
Perception	$\mathbf{1.108}$ $[0.812, 1.384]$	1.611 $[1.432, 1.857]$
Reasoning	1.684 $[1.426, 1.965]$	$\mathbf{1.247}$ $[0.946, 1.358]$

Define the matched-expert advantage as the mismatched expert's NLL minus the matched expert's NLL. Positive values favor the intended assignment. The differences are 0.503 for perception and 0.437 for reasoning, with an equally weighted mean of 0.470. The crossover in expert ordering supports role-matched likelihood preferences in both token categories on these responses. The intervals in Table 15 describe individual cell means; confidence intervals for the paired differences are not reported. Likelihood agreement does not by itself establish token correctness or the causal benefit of using an expert during training.

D Sign-Preserving Correction: Mathematical Illustration

GRPO assigns the same advantage to every valid token of one response. The four rows of Table 16 therefore describe possible conditions across responses, not tokens with different advantage signs within one response. Let $\gamma_t = (1 - \lambda_k) + \lambda_k \omega_t$, so that $\widetilde{A}_t = \gamma_t \widehat{A}_t$. For positive λ_k and clipping width, the sign of the expert gap determines whether the common response advantage is amplified or attenuated at a selected position.

Table 16: Algebraic behavior of the correction, not an empirical trajectory. The expert gap compares token likelihoods, not token correctness.

$\widehat{A}$	Δ_t	γ_t	Effect on advantage
> 0	> 0	> 1	larger positive advantage
> 0	< 0	< 1	smaller positive advantage
< 0	> 0	< 1	smaller negative magnitude
< 0	< 0	> 1	larger negative magnitude

For $\Delta_t = 0$ or $\lambda_k = 0$, the multiplier equals one. A zero original advantage remains zero. More generally, $\gamma_t \in [1 - \lambda_k \epsilon_w, 1 + \lambda_k \epsilon_w]$; with $\lambda_0 = 0.5$ and $\epsilon_w = 0.2$, the multiplier in the first rollout iteration is between 0.9 and 1.1. This is a bound derived from the update rule, not a measured distribution. The mechanism refines relative magnitudes within a response while preserving its outcome-derived direction, so it cannot reverse an individual token's advantage or assign a nonzero update when the response advantage is zero.

E REPRODUCIBILITY NOTES

Post-trained main results summarize Avg@8 over three independent training runs with seeds 42, 43, and 44 as means and standard deviations. For each run, checkpoints are selected by Geo3K strict accuracy on the shared one-epoch validation grid. Base evaluates the pretrained checkpoint without further training, so it is reported without a training-run standard deviation. Ablations retain single training runs and evaluation seed 0 unless stated otherwise.

The Qwen2.5-VL-3B/7B-Instruct students and both 7B experts share the Qwen2.5-VL tokenizer family, allowing direct scoring of student-sampled tokens without vocabulary remapping.

Tables 5 and 7 list the default routing and reweighting settings. The `w/o Stabil.` variant removes clipping and holds $\lambda_k \equiv 0.5$. The audited software environment uses Python 3.12.3, PyTorch 2.9.0, CUDA 12.8, NCCL 2.27.5, Transformers 4.57.6, vLLM 0.12.0, Datasets 4.4.1, and verl 0.8.0.dev0.

Repeated ablation runs and standalone expert evaluations would clarify ablation stability and expert capability.

F Complete Main Results

Tables 17 and 18 report post-trained means and standard deviations across three independent training runs (seeds 42, 43, and 44); Base reports accuracy only. Bold/underlined means mark best/second-best per backbone here and in the main table. Benchmark subsets and scoring rules are specified in Appendix A.1.

Table 17: Full perception-oriented results (Avg@8 accuracy, %).

Method	RWQA	MMStar	HalluBench	P-Avg
		Qwen2.5-VL-3B		
Base	59.70	52.80	60.04	57.51
GRPO	60.44 ± 0.58	55.73 ± 0.63	62.17 ± 0.70	59.45 ± 0.43
DAPO	59.39 ± 0.62	56.58 ± 0.54	64.13 ± 0.76	60.03 ± 0.46
VPPO	61.85 ± 0.51	$\underline{58.50} \pm 0.47$	64.17 ± 0.66	61.51 ± 0.38
OPD	59.84 ± 0.55	56.93 ± 0.43	63.92 ± 0.60	60.23 ± 0.36
ToR	61.96 ± 0.46	57.26 ± 0.52	64.20 ± 0.58	61.14 ± 0.34
RLSD	60.94 ± 0.42	57.49 ± 0.50	64.17 ± 0.53	60.87 ± 0.32
RLRD-P	$\underline{62.08} \pm 0.39$	56.87 ± 0.46	$\underline{65.74} \pm 0.51$	$\underline{61.56} \pm 0.29$
RLRD-R	59.22 ± 0.44	57.73 ± 0.42	63.12 ± 0.57	60.02 ± 0.33
RLRD-Dual	$\mathbf{63.40} \pm 0.38$	$\mathbf{58.93} \pm 0.46$	$\mathbf{66.37} \pm 0.41$	$\mathbf{62.90} \pm 0.29$
		Qwen2.5-VL-7B		
Base	67.52	63.13	65.50	65.38
GRPO	68.60 ± 0.49	64.06 ± 0.56	68.80 ± 0.64	67.15 ± 0.39
DAPO	67.60 ± 0.57	61.72 ± 0.68	69.03 ± 0.59	66.12 ± 0.44
VPPO	69.08 ± 0.45	65.18 ± 0.50	69.40 ± 0.55	67.89 ± 0.35
OPD	68.48 ± 0.48	64.22 ± 0.53	67.12 ± 0.47	66.61 ± 0.34
ToR	68.70 ± 0.41	65.42 ± 0.44	$\underline{70.60} \pm 0.50$	68.24 ± 0.31
RLSD	68.14 ± 0.43	64.88 ± 0.40	69.44 ± 0.52	67.49 ± 0.30
RLRD-P	69.74 ± 0.36	$\underline{65.76} \pm 0.39$	$\underline{70.60} \pm 0.45$	$\underline{68.70} \pm 0.27$
RLRD-R	$\underline{69.75} \pm 0.37$	64.78 ± 0.42	69.52 ± 0.46	68.02 ± 0.28
RLRD-Dual	$\mathbf{71.28} \pm 0.34$	$\mathbf{67.42} \pm 0.36$	$\mathbf{71.46} \pm 0.40$	$\mathbf{70.05} \pm 0.24$

Table 18: Full reasoning-oriented results (Avg@8 accuracy, %).

Method	MathVista	LogicVista	MathVerse	R-Avg
		Qwen2.5-VL-3B		
Base	57.80	37.28	30.84	41.97
GRPO	63.70 ± 0.74	39.06 ± 0.61	36.29 ± 0.67	46.35 ± 0.49
DAPO	60.40 ± 0.69	39.17 ± 0.73	34.80 ± 0.58	44.79 ± 0.48
VPPO	64.60 ± 0.59	43.07 ± 0.55	37.53 ± 0.62	48.40 ± 0.41
OPD	65.70 ± 0.57	40.79 ± 0.64	35.17 ± 0.50	47.22 ± 0.40
ToR	65.80 ± 0.63	43.04 ± 0.49	$\underline{38.50} \pm 0.57$	49.11 ± 0.39
RLSD	66.10 ± 0.52	42.79 ± 0.47	37.92 ± 0.55	48.94 ± 0.36
RLRD-P	67.20 ± 0.48	42.19 ± 0.53	37.15 ± 0.45	48.85 ± 0.34
RLRD-R	$\underline{67.50} \pm 0.46$	$\underline{44.68} \pm 0.51$	38.14 ± 0.48	$\underline{50.11} \pm 0.32$
RLRD-Dual	$\mathbf{67.90} \pm 0.39$	$\mathbf{45.98} \pm 0.44$	$\mathbf{39.50} \pm 0.42$	$\mathbf{51.13} \pm 0.30$
		Qwen2.5-VL-7B		
Base	68.00	44.49	41.11	51.20
GRPO	70.40 ± 0.61	46.08 ± 0.58	44.28 ± 0.66	53.59 ± 0.45
DAPO	69.10 ± 0.64	45.76 ± 0.70	42.58 ± 0.54	52.48 ± 0.47
VPPO	72.30 ± 0.57	43.44 ± 0.62	42.12 ± 0.48	52.62 ± 0.40
OPD	71.20 ± 0.54	44.12 ± 0.59	42.68 ± 0.52	52.67 ± 0.39
ToR	72.60 ± 0.49	46.22 ± 0.55	44.30 ± 0.46	54.37 ± 0.35
RLSD	72.10 ± 0.50	45.74 ± 0.46	43.94 ± 0.49	53.93 ± 0.34
RLRD-P	72.50 ± 0.44	45.92 ± 0.48	44.23 ± 0.41	54.22 ± 0.32
RLRD-R	$\underline{73.30} \pm 0.43$	$\underline{46.84} \pm 0.47$	$\underline{45.34} \pm 0.39$	$\underline{55.16} \pm 0.31$
RLRD-Dual	$\mathbf{74.20} \pm 0.38$	$\mathbf{48.14} \pm 0.42$	$\mathbf{46.36} \pm 0.37$	$\mathbf{56.23} \pm 0.28$