RESIDUAL VISUAL CREDIT OPTIMIZATION: CONSERVED EVIDENCE ROUTING FOR MULTIMODAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards scales multimodal reasoning, but an outcome reward says how much a trajectory is worth, not how that value should be spread over the decisions that produced it. We introduce Residual Visual Credit Optimization (RVCO), which treats token credit as a conserved routing problem. A controlled visual intervention yields a per-token evidence response; robust within-trajectory coordinates remove incidental scale; and a budgeted entropic router distributes a fixed amount of sequence utility according to perceptual dependence. A residual support path guarantees positive credit at every valid position, and an analytic correction restores the prescribed credit mass exactly. The resulting field is selective, bounded, full-support, and invariant to response-local score shifts, and recovers hard token selection as a limiting case. Across four model families and seven reasoning benchmarks, RVCO improves accuracy over strong RLVR baselines while maintaining late-stage optimization stability, corruption robustness, and competitive training cost. Rewards, rollouts, and the group-relative advantage estimator are unchanged; only the geometry of token-level credit differs.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) (Lightman et al., 2024; Shao et al., 2024; Guo et al., 2025) has emerged as a key mechanism for scaling reasoning in large language models. It turns automatically verified outcomes into effective policy updates without requiring dense process supervision (Wen et al., 2025; Yu et al., 2025). The extension of RLVR to large vision-language models has shown strong potential for visual reasoning (Huang et al., 2025). Mathematical diagrams, charts, documents, and structured visual questions often support objective final-answer verification (Meng et al., 2025; Liu et al., 2025). Dense annotation of intermediate reasoning steps remains expensive for these tasks (Wang et al., 2025). Modern multimodal backbones (Bai et al., 2023; 2025b;a) further make this regime practical at increasing context length (Gemini Team et al., 2024) and model scale (Zhu et al., 2025; Dubey et al., 2024; Li et al., 2024).

The central optimization problem, however, is richer than assigning a scalar score to a completed response. A multimodal trajectory is a structured computation in which visually grounded decisions, linguistic transitions, symbolic manipulations, and later conclusions remain coupled across hundreds or thousands of generated positions. Once a sequence-level utility is available, the policy still needs an allocation law that decides *where* that utility enters the autoregressive computation. Uniform broadcast is simple and stable, yet it ignores heterogeneous visual dependence. Perception-aware RL therefore increasingly introduces visual interventions (Wang et al., 2026c; Huang et al., 2026; Ye et al., 2026), confidence signals (Xiao et al., 2026), process rewards (Min et al., 2026), hidden-state similarity (Li et al., 2026b; Wang et al., 2026b), or token-level advantage reshaping (Li et al., 2026a). This body of work establishes token credit as an important interface between outcome supervision and grounded multimodal computation, but it leaves open what a credit operator should guarantee: existing designs fix either the active support or the score transform, and the total amount of update they inject is a consequence of that choice rather than a controlled quantity.

A visual-credit operator must reconcile three properties that are individually natural but jointly nontrivial. *Selectivity* requires assigning greater credit to decisions whose likelihood depends more

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.

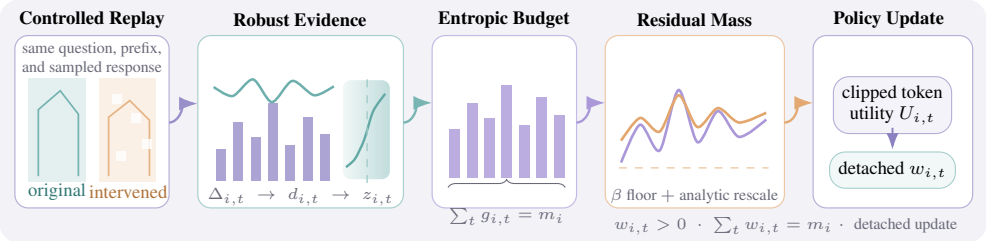

Figure 1: Conserved evidence routing in RVCO. Controlled visual replay produces a sampled-token evidence response without changing the generated trajectory. Response-local calibration removes incidental scale, the entropic router concentrates a fixed mass m_i, and residual mixing preserves positive support before analytic correction restores the exact budget. Only the detached token weights enter policy optimization; rewards and advantages remain unchanged.

strongly on visual evidence. *Support continuity* requires maintaining a learning signal along the surrounding reasoning trajectory, since visual evidence can influence distributed spans rather than isolated tokens (Sun et al., 2025; Favero et al., 2024). *Utility conservation* requires controlling the total credit allocated to a trajectory independently of response length, evidence scale, and training stage. Hard top-ρ routing provides explicit selectivity and budget control but introduces discontinuous support (Rao et al., 2021; Shi et al., 2023); independent soft gates preserve continuous support but allow the total credit mass to vary with score calibration. Existing approaches therefore leave unresolved how to achieve selective, continuous, and budget-preserving credit within a single operator.

We formulate visual credit as *conserved evidence routing*. Residual Visual Credit Optimization (RVCO) first compares each sampled token under the original image and a controlled visual intervention while holding the textual state fixed. The resulting likelihood response is transformed into a robust within-trajectory coordinate and passed through an entropy-regularized router that allocates a fixed credit budget according to visual dependence. A residual path then guarantees positive support for every valid token, while a closed-form correction restores the prescribed mass exactly. The reward, rollout procedure, and group-relative advantage estimator remain unchanged; RVCO modifies only the geometry of token-level credit. The resulting construction provides a continuous bridge between sparse and dense credit allocation while separating the roles of its components. The evidence coordinate determines where credit is allocated, the gate temperature τ controls how sharply it concentrates, the residual ratio β controls how broadly support is maintained, and the visual budget ρ fixes how much total credit mass is distributed. These roles are separated explicitly rather than entangled in an unconstrained score transformation. We evaluate RVCO across multiple model families and scales on visual mathematics, geometry, logical and multidisciplinary reasoning, together with corruption robustness and out-of-domain VQA.

The main contributions of this paper are summarized as follows:

- **Conserved Evidence Routing.** We formulate multimodal token credit as a constrained allocation problem that jointly enforces visual selectivity, continuous support, and a fixed response-level credit budget.
- **Residual Visual Credit Optimization.** We introduce RVCO, which combines robust intervention-based evidence, budgeted entropic routing, residual full-support credit, and exact mass correction. The resulting credit field is bounded, positive at every valid token, order-preserving, and exactly budget-preserving.
- **Evaluation across Models and Reasoning Tasks.** Across multiple VLM families and scales, RVCO consistently improves over strong RLVR and perception-aware baselines, with additional analyses demonstrating robustness to visual corruption and improved late-stage optimization stability.

2 Related Work

Multimodal RLVR. RLVR enables policy optimization from automatically verifiable outcomes without dense preference or process supervision. GRPO (Shao et al., 2024) provides a critic-free

group-relative estimator, while DAPO (Yu et al., 2025) develops a scalable online training recipe. Multimodal extensions build on this paradigm through curated reasoning data, rule-based rewards (Huang et al., 2025; Meng et al., 2025), self-reflection, and visual reinforcement fine-tuning (Liu et al., 2025; Wang et al., 2025). Failure-centric resampling and correction-specific rollouts further improve the construction of informative training trajectories (Wang et al., 2026a; Ding et al., 2026). These methods primarily improve the data, rollouts, rewards, or sequence-level optimization. In contrast, RVCO holds the sequence-level utility fixed and studies how it should be allocated across token decisions within a trajectory.

Perceptual credit assignment. Recent multimodal RL methods increasingly differentiate token updates according to visual dependence. PAPO (Wang et al., 2026c) derives an implicit perception objective by contrasting original and corrupted visual inputs; VPPO (Huang et al., 2026) identifies perceptually pivotal tokens and concentrates updates on a sparse subset; and PGPO (Ye et al., 2026) and PRPO (Li et al., 2026a) reshape token-level advantages using estimates of visual dependency. PEPO (Li et al., 2026b) combines perception priors with token entropy, VGPO (Wang et al., 2026b) addresses temporal visual forgetting, and VL-Calibration (Xiao et al., 2026) separates visual from reasoning confidence. Process-reward and attribution methods provide complementary forms of dense supervision (Min et al., 2026; Wu et al., 2026). Collectively, these methods establish the value of identifying *where* visual evidence matters. RVCO addresses the complementary question of *how* such evidence should be converted into token-level credit: it treats allocation as a constrained routing problem with explicit control over support, concentration, and total credit mass.

Visual grounding and intervention. Counterfactual saliency and visual contrastive decoding use controlled changes to visual input to probe or strengthen the dependence of model predictions on image evidence (Wang et al., 2023; Leng et al., 2024; Favero et al., 2024). Related benchmarks such as HallusionBench and SimpleVQA expose failures in visual grounding, illusion resistance, and multimodal factuality (Guan et al., 2024; Cheng et al., 2025). These works motivate visual intervention as a useful diagnostic signal. RVCO uses intervention differently: rather than modifying decoding or defining an auxiliary reward, it measures the response of the *sampled token likelihood* under controlled visual perturbation and uses that response only to construct a detached credit field.

Credit geometry and optimization stability. Token-level credit rules induce different optimization geometries even when they are constructed from similar visual-dependence signals. Hard selection provides explicit sparsity and budget control but changes support discontinuously at rank boundaries, whereas unconstrained continuous reweighting smooths token allocation while allowing the effective credit mass to vary with score calibration. Related work on dynamic token sparsification shows that changing active sets can itself become an optimization variable (Rao et al., 2021; Shi et al., 2023), while analyses of multimodal policy optimization emphasize the importance of late-stage stability (Huang et al., 2026; Wang et al., 2026c). RVCO separates three properties that are coupled in these alternatives: the visual budget ρ controls total credit mass, the gate temperature τ controls routing concentration, and the residual ratio β controls support. This decomposition motivates our analyses of optimization stability, response length, intervention choice, and data efficiency.

3 Residual Visual Credit Optimization

3.1 Problem Setup

For an image-question pair $x = (I, q)$, a behavior policy $\pi_{\bar{\theta}}$ samples a group of G responses $y^{(i)} = (y_1^{(i)}, \ldots, y_{T_i}^{(i)}) \sim \pi_{\bar{\theta}}(\cdot \mid I, q)$. A verifier returns a sequence reward R_i, and group-relative normalization produces an advantage A_i. Standard clipped policy optimization then broadcasts A_i across all valid positions (Schulman et al., 2017; Shao et al., 2024). RVCO instead constructs a detached token field $w_i \in \mathbb{R}^{T_i}$ under a visual budget $m_i = \lceil \rho T_i \rceil$. The visual budget ρ controls the total credit mass allocated to the trajectory: it matches the mass of a top-ρ sparse allocation without requiring the support itself to be binary.

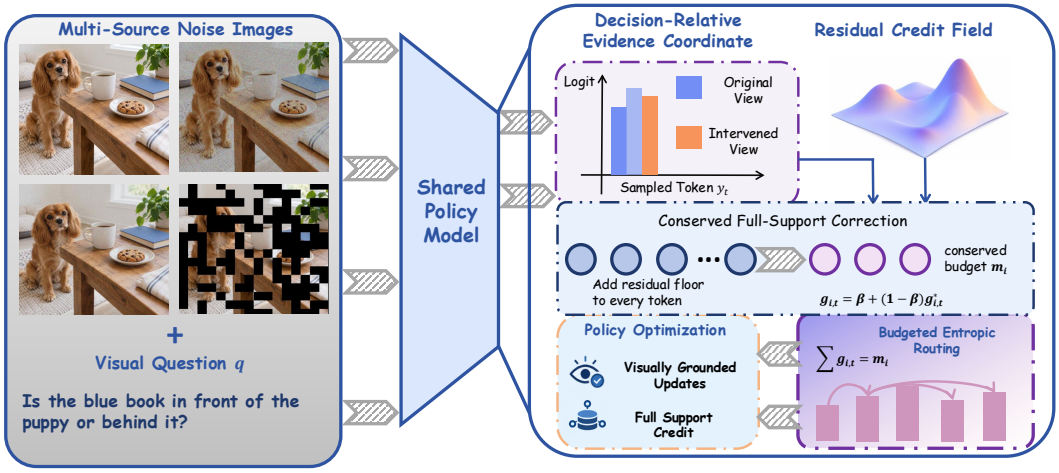

Figure 2: Main architecture of RVCO. Multiple controlled visual interventions are evaluated by a shared behavior policy while the visual question and generated response state are held fixed. The resulting original/intervened token responses define a decision-relative evidence coordinate, which is transformed into a residual credit field and routed under the exact budget constraint $\sum_t g_{i,t} = m_i$. The conserved field then modulates the ordinary policy update, emphasizing visually grounded decisions while retaining full-support credit across the response.

The desired operator is characterized by four response-local conditions, written compactly as

$$0 < w_{i,t} \le 1, \qquad \sum_{t=1}^{T_i} w_{i,t} = m_i, \qquad z_{i,t} \ge z_{i,s} \Rightarrow w_{i,t} \ge w_{i,s}, \qquad w_i = \mathcal{G}(z_i; \rho, \tau, \beta). \quad (1)$$

The first condition preserves support, the second fixes credit mass, the third preserves evidence order, and the last keeps allocation local to the sampled trajectory. No token label, learned critic, or auxiliary reward model is introduced.

3.2 Overview

Figure 2 shows the full computation. The sampled response is replayed under an intervened image with the question and prefix fixed, yielding one likelihood comparison per generated token. A monotone compression followed by IQR normalization produces a scale-free evidence coordinate. An entropy-regularized router then allocates exactly m_i units of credit, after which residual mixing keeps every token connected to the outcome signal. Analytic mass correction removes the mass increase caused by residual mixing, and the resulting detached weights multiply the ordinary clipped token utility. The same rollout, verifier, reward, and advantage tensors can therefore be passed through an unchanged RLVR trainer.

3.3 Visual Intervention as a Token Evidence Response

The policy itself provides a task-adaptive probe of visual dependence. For response i, let $\widetilde{I} = \mathcal{T}(I; \xi)$ denote an image intervention that independently suppresses visual patches with probability 0.5. The same sampled token and prefix are evaluated under both inputs,

$$\begin{aligned} \ell_{i,t} &= \log \pi_{\bar{\theta}} \left(y_t^{(i)} \mid I, q, y_{<t}^{(i)} \right), \\ \tilde{\ell}_{i,t} &= \log \pi_{\bar{\theta}} \left(y_t^{(i)} \mid \tilde{I}, q, y_{<t}^{(i)} \right), \\ \Delta_{i,t} &= \mathrm{clip} \left(\tilde{\ell}_{i,t} - \ell_{i,t}, -20, 20 \right). \end{aligned} \quad (2)$$

Rather than retaining a signed shift, RVCO maps it to a non-negative convex response

$$d_{i,t} = \exp(\Delta_{i,t}) - \Delta_{i,t} - 1, \qquad d_{i,t} \leftarrow \mathrm{clip}(d_{i,t}, 0, 10). \quad (3)$$

The function is zero when intervention leaves the sampled-token likelihood unchanged and increases for shifts in either direction. Measuring the sampled decision directly also keeps the evidence signal aligned with the action whose log-probability appears in the policy gradient, rather than summarizing changes over the entire vocabulary. Because only one scalar is retained per generated position, the construction avoids storing a full-vocabulary attribution tensor. The evidence response is detached before optimization and is not interpreted as a new reward or formal causal effect.

3.4 Scale-Free Evidence Coordinates

Raw likelihood responses vary with policy confidence, response composition, and training stage. Their order is often more transferable than their absolute magnitude. A monotone compression and robust response-local normalization therefore define

$$x_{i,t} = \log(1 + d_{i,t}), \qquad z_{i,t} = \mathrm{clip}\left(\frac{x_{i,t} - Q_{i,.50}}{\max(Q_{i,.75} - Q_{i,.25}, \varepsilon)}, -\kappa, \kappa \right), \quad \varepsilon = 10^{-6},\ \kappa = 8, \tag{4}$$

where $Q_{i,p}$ is the empirical p-quantile over valid tokens. This transformation preserves within-response evidence order, is unaffected by a shared shift in x_i, and prevents one extreme token from setting the scale for an entire trajectory. Using the interquartile range also makes the coordinate less sensitive to a small number of unusually large intervention responses, which can occur as the policy becomes increasingly confident during training.

3.5 Budgeted Entropic Routing

Credit is allocated jointly rather than by independent token gates. Let $g_i \in [0,1]^{T_i}$ and $h(a) = -a \log a - (1-a)\log(1-a)$. The routed field solves

$$g_i^* = \arg \max_{0 \le g_i \le 1,\ \mathbf{1}^\top g_i = m_i} \left[g_i^\top z_i + \tau \sum_{t=1}^{T_i} h(g_{i,t}) \right], \qquad g_{i,t}^* = \sigma\left(\frac{z_{i,t} - \lambda_i}{\tau} \right), \quad \sum_t g_{i,t}^* = m_i. \tag{5}$$

The scalar dual variable λ_i is obtained by 32 bisection steps. For $0 < m_i < T_i$, the mass equation is continuous and strictly decreasing in λ_i, yielding a unique finite solution. In the boundary case $m_i = T_i$, we directly set $g_{i,t}^* = 1$ for all valid positions. For $0 < m_i < T_i$, unlike independent gating, the equality constraint makes positions compete for the same response-level budget: increasing the allocation of one region necessarily changes how the remaining mass is distributed elsewhere. This coupling is mediated only by the scalar λ_i, keeping the routing computation simple despite the joint constraint. The gate temperature τ controls routing concentration but not total credit mass: decreasing τ approaches a rank selector, while increasing it spreads the same mass over a broader set of positions.

3.6 Residual Support and Exact Conservation

Finite-precision kernels and downstream thresholds can erase very small gate values. Positive support is therefore made explicit through residual mixing, followed by a closed-form correction:

$$\bar{g}_{i,t} = \beta + (1-\beta) g_{i,t}^*, \qquad w_{i,t} = \frac{m_i \bar{g}_{i,t}}{\sum_s \bar{g}_{i,s}} = \frac{m_i \left[\beta + (1-\beta) g_{i,t}^*\right]}{\beta T_i + (1-\beta) m_i}, \qquad 0 < \beta < 1. \tag{6}$$

This step separates guaranteed support from selectivity. The residual ratio β controls the strength of the full-support path. A small β preserves sparse-like contrast, whereas larger values move the field toward uniform credit without altering the prescribed mass. Importantly, the residual path is not another smoothing operator on z_i: it acts after routing and therefore ensures positive support while preserving the evidence ordering produced by the entropic gate.

Proposition 1 (conserved full support). For $0 < \beta < 1$ and $0 < m_i \le T_i$, Equation 6 satisfies $0 < w_{i,t} \le 1$ and $\sum_t w_{i,t} = m_i$. It is monotone in $g_{i,t}^*$, invariant to additive shifts of z_i, and converges to the top-m_i selector as $\tau \to 0$ and $\beta \to 0$ when the coordinates are distinct.

Proof sketch. Positivity follows from $\beta > 0$, while the denominator is exactly $\sum_t \bar{g}_{i,t} = \beta T_i + (1 - \beta) m_i$. Since $m_i \le T_i$, the multiplicative correction is at most one and hence $w_{i,t} \le 1$. Monotonicity

is inherited from the sigmoid gate. Adding a constant to every $z_{i,t}$ shifts the dual variable by the same constant and leaves g_i^* unchanged. As $\tau \to 0$, the entropy term vanishes and the capped linear program selects the m_i largest coordinates; $\beta \to 0$ removes the residual floor. A complete derivation appears in Appendix A.

3.7 Full-Support Policy Optimization

For policy π_θ, define the importance ratio and asymmetric clipped token utility on one line,

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_t^{(i)} \mid I, q, y_{<t}^{(i)})}{\pi_{\bar{\theta}}(y_t^{(i)} \mid I, q, y_{<t}^{(i)})}, \qquad U_{i,t}(\theta) = \min\big(r_{i,t}A_i, \mathrm{clip}(r_{i,t}, 1-\epsilon_l, 1+\epsilon_h)A_i\big). \tag{7}$$

The final objective is

$$\mathcal{L}_{\mathrm{RVCO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\mathrm{stopgrad}(w_{i,t})\,U_{i,t}(\theta). \tag{8}$$

Because the weights sum to m_i, RVCO preserves the prescribed token-credit budget while keeping every valid token connected to sequence utility. The stop-gradient operation is important here: intervention-based evidence controls the relative weighting of each token utility but does not create an additional differentiation path through the credit-construction procedure. The additional operations are response-local quantiles, elementwise transforms, and scalar bisection. Standard attention architectures and AdamW optimization remain unchanged (Vaswani et al., 2017; Loshchilov & Hutter, 2019).

Overall Process of RVCO. RVCO converts intervention-based visual responses into a conserved token-level credit field through three steps: robust evidence calibration, budgeted entropic routing, and residual full-support correction. The resulting weights remain response-local and detached from policy optimization, leaving the verifier, group-relative advantage estimator, and PPO training pipeline unchanged. The complete optimization procedure is summarized in Algorithm 1.

Algorithm 1 Residual Visual Credit Optimization

Require: : behavior policy $\pi_{\bar{\theta}}$, policy π_θ, pair (I, q), visual budget ρ, gate temperature τ, residual ratio β
1: Sample a rollout group $\{y^{(i)}\}_{i=1}^{G}$, evaluate rewards, and compute group-relative advantages $\{A_i\}$
2: Construct an intervened image $\tilde{I} = \mathcal{T}(I;\xi)$ while keeping sampled responses and prefixes fixed
3: **for** each response i **do**
4: Compute $\Delta_{i,t}$, $d_{i,t}$, and $z_{i,t}$ for valid tokens using Eqs. 2–4
5: Set $m_i = \lceil \rho T_i \rceil$ and solve $\sum_t \sigma((z_{i,t} - \lambda_i)/\tau) = m_i$
6: Form $g_{i,t}^*$ and conserved full-support weights $w_{i,t}$ using Eq. 6
7: **end for**
8: Update θ with Eq. 8; no gradient flows through $w_{i,t}$

4 Experiments

4.1 Experimental settings

Models and training data. Our primary policy is Qwen3-VL-8B-Instruct (Bai et al., 2025a). Cross-family experiments use Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B (Zhu et al., 2025), and Llama-3.2-Vision-11B (Dubey et al., 2024); scale analysis additionally includes Qwen3-VL-4B and Qwen3-VL-32B. All models are trained on ViRL39K, a 38,870-example verifiable multimodal reasoning corpus (Wang et al., 2025). We evaluate on seven reasoning benchmarks: MathVerse (Zhang et al., 2024), DynaMath (Zou et al., 2025), MMK12 (Meng et al., 2025), Geo3K (Lu et al., 2021), VisNumBench (Weng et al., 2025), LogicVista (Xiao et al., 2024), and MMMU-Pro (Yue et al., 2025).

Baselines. We compare against general RLVR methods GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025); multimodal RL methods Vision-R1 (Huang et al., 2025), MM-Eureka (Meng et al., 2025), Visual-RFT (Liu et al., 2025), and VL-Rethinker (Wang et al., 2025); and perception-aware

objectives PAPO (Wang et al., 2026c), VPPO (Huang et al., 2026), PGPO (Ye et al., 2026), PRPO (Li et al., 2026a), VL-Calibration (Xiao et al., 2026), and VGPO (Wang et al., 2026b). Optimization-side baselines use the same Qwen3-VL trainer, rollout budget, verifier, and decoding configuration. Cross-model experiments retain each backbone's native visual encoder and tokenizer.

Optimization and metrics. Unless stated otherwise, we train for two epochs with AdamW at a learning rate of 10^{-6}, using rollout and global batch sizes of 384 and 128, eight responses per prompt, temperature 1.0, top-p 0.99, and maximum prompt/response lengths of 4096/8192. We use asymmetric clipping $(\epsilon_l, \epsilon_h) = (0.20, 0.28)$ and default RVCO parameters $(\rho, \tau, \beta) = (0.4, 0.12, 0.05)$, corresponding to the visual budget, gate temperature, and residual ratio. For evaluation, we report accuracy@8 (%), the percentage of examples with at least one exact-match correct answer among eight sampled responses under the official extraction rules. We additionally report credit stability, corruption robustness, training cost, and out-of-domain VQA metrics. Appendix B provides the complete protocol.

4.2 Comparison with State-of-the-Art Baselines

Qwen-Family Comparison. We first evaluate RVCO on Qwen3-VL-4B, 8B, and 32B using the same seven benchmarks and training protocol within each backbone. As shown in Table 1, RVCO achieves the highest average accuracy at all three scales, improving over VPPO by 2.1, 1.1, and 1.1 points on the 4B, 8B, and 32B models, respectively. The gains are distributed across benchmarks rather than driven by a single task. At 32B, VPPO is higher on DynaMath, while RVCO achieves the highest average and leads on the other six benchmarks. Additional Qwen3-VL-8B baselines are reported in Appendix Table 7.

Table 1: Benchmark-wise comparison across the Qwen3-VL family. All values are accuracy@8 (%); Avg. is the macro-average over the seven benchmarks with reported results. The same rollout budget, verifier, and decoding protocol are used for all methods within each backbone. Bold and underlined values denote the best and second-best results within each backbone.

Backbone	Method	MathV.	Dyna	MMK12	Geo3K	VisNum	Logic	MMMU-P	Avg.
Qwen3-VL-4B	GRPO	66.5	65.8	72.3	65.0	23.1	57.8	39.2	55.7
	Visual-RFT	80.7	<u>75.6</u>	<u>76.3</u>	67.6	<u>25.4</u>	59.1	49.3	<u>62.0</u>
	VPPO	<u>80.9</u>	75.3	76.2	<u>67.7</u>	24.8	<u>59.2</u>	<u>49.5</u>	61.9
	RVCO	**82.3**	**78.2**	**78.0**	**69.7**	**27.2**	**61.5**	**51.1**	**64.0**
Qwen3-VL-8B	GRPO	76.5	75.0	71.2	64.4	28.7	56.0	42.8	59.2
	Visual-RFT	81.0	78.1	77.2	69.1	<u>29.6</u>	<u>63.0</u>	49.3	63.9
	VPPO	<u>83.8</u>	<u>80.0</u>	<u>80.1</u>	<u>71.9</u>	29.1	62.3	<u>52.4</u>	<u>65.7</u>
	RVCO	**84.4**	**80.4**	**80.6**	**72.5**	**30.5**	**64.1**	**55.4**	**66.8**
Qwen3-VL-32B	GRPO	79.2	78.4	82.6	78.9	31.7	67.8	42.7	65.9
	Visual-RFT	83.0	84.0	84.8	79.2	32.3	67.4	<u>60.2</u>	70.1
	VPPO	<u>85.0</u>	**86.2**	<u>87.0</u>	<u>81.4</u>	<u>32.8</u>	<u>69.8</u>	59.7	<u>71.7</u>
	RVCO	**86.9**	<u>85.9</u>	**88.8**	**82.1**	**33.4**	**70.2**	**62.3**	**72.8**

Cross-Family Generalization. To test whether the gains are specific to Qwen3-VL, we apply RVCO to Qwen2.5-VL-7B, InternVL3-8B, and Llama-3.2-Vision-11B, retaining each backbone's native visual encoder and tokenizer. Table 2 shows consistent gains over VPPO, improving the macro-average by 1.1, 1.1, and 1.0 points, respectively. The largest improvements occur on MMMU-Pro, with gains of 3.0, 3.3, and 3.4 points. These results show that the gains extend across heterogeneous vision-language backbones rather than being specific to Qwen3-VL.

4.3 Ablation Study

Ablation Variants. We ablate each component of the RVCO credit field on Qwen3-VL-8B, holding the training data, reward, rollout budget, optimizer, and decoding protocol fixed. *w/o Robust Coordinate* bypasses Eq. 4 and uses the uncalibrated evidence response directly. *w/o Budgeted Routing* replaces the constrained gate in Eq. 5 with independent sigmoid gates followed by mass matching. *w/o Residual Support* sets $\beta = 0$ in Eq. 6, removing the explicit full-support path. *w/o Exact Conservation* replaces the analytic correction in Eq. 6 with unit-mean normalization, no longer preserving the prescribed credit budget m_i. The evidence response in Eq. 3 and policy objective in Eq. 8 remain unchanged across all variants.

Ablation Results. Table 3 shows that removing any individual component degrades overall performance, indicating that the four components provide complementary benefits. Removing the robust coordinate weakens performance across most benchmarks, while replacing budgeted routing further reduces the effectiveness of credit allocation. Removing residual support produces the largest degradation, highlighting the importance of maintaining full-support credit along the reasoning trajectory. Replacing exact conservation also consistently reduces performance, supporting the importance of controlling the total credit mass during optimization.

Table 2: Benchmark-wise comparison across vision-language model families. All values are accuracy@8 (%); Avg. is the macro-average over the seven benchmarks. Each backbone retains its native visual encoder and tokenizer. Bold and underlined values denote the best and second-best results within each backbone.

Backbone	Method	MathV.	Dyna	MMK12	Geo3K	VisNum	Logic	MMMU-P	Avg.
Qwen2.5-VL-7B	GRPO	75.0	72.9	69.5	62.0	26.4	54.0	40.7	57.2
	Visual-RFT	79.5	78.2	75.5	66.7	28.1	58.0	46.8	61.8
	VPPO	82.4	78.0	78.6	69.6	28.6	60.4	50.4	64.0
	RVCO	**83.1**	**78.5**	**79.1**	**70.3**	**29.0**	**62.3**	**53.4**	**65.1**
InternVL3-8B	GRPO	73.1	72.3	68.0	61.5	25.9	53.1	39.6	56.2
	Visual-RFT	77.4	75.2	73.8	66.0	27.3	56.9	45.5	60.3
	VPPO	79.9	76.8	76.4	68.5	27.0	58.9	48.1	62.2
	RVCO	**80.3**	**77.0**	**76.7**	**68.9**	**28.2**	**60.5**	**51.4**	**63.3**
Llama-3.2-Vision-11B	GRPO	70.5	69.9	66.5	59.6	23.8	51.2	37.5	54.1
	Visual-RFT	73.6	71.6	71.1	62.9	25.1	56.1	42.2	57.5
	VPPO	76.1	73.2	73.6	**66.4**	25.6	55.8	44.9	59.4
	RVCO	**76.7**	**73.6**	**74.2**	66.0	**26.2**	**57.6**	**48.3**	**60.4**

4.4 Hyperparameter Sensitivity

To assess whether RVCO depends on a narrowly tuned routing configuration, we independently vary the residual ratio $\beta \in \{0.01, 0.03, 0.05, 0.07, 0.09\}$, visual budget $\rho \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$, and gate temperature $\tau \in \{0.04, 0.08, 0.12, 0.16, 0.20\}$ on Qwen3-VL-8B, holding the other parameters at their defaults. We report training accuracy and MathVerse validation accuracy.

Table 3: Ablation study on Qwen3-VL-8B. Each variant changes one component while preserving the same reward, rollout budget, optimizer, and decoding protocol.

Variant	MathV.	Dyna	MMK12	Geo3K	VisNum	Logic	MMMU-P	Avg.
w/o Robust Coordinate	83.2	78.7	79.3	71.3	29.7	62.9	53.2	65.5
w/o Budgeted Routing	81.7	77.0	78.5	71.6	28.9	61.5	52.1	64.5
w/o Residual Support	80.9	76.4	77.2	69.0	27.8	60.6	49.3	63.0
w/o Exact Conservation	81.9	77.6	78.1	70.5	28.6	61.2	51.7	64.2
RVCO	**84.4**	**80.4**	**80.6**	**72.5**	**30.5**	**64.1**	**55.4**	**66.8**

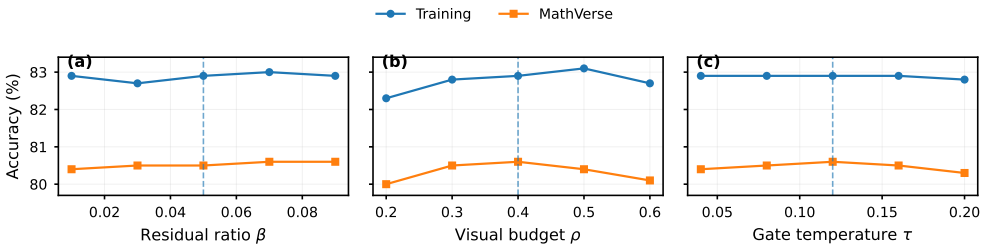

Figure 3: Hyperparameter sensitivity on Qwen3-VL-8B. We vary the residual ratio β, visual budget ρ, and gate temperature τ while holding the other two parameters fixed. Dashed lines mark the default configuration.

Figure 3 shows that RVCO is stable over a broad neighborhood of the default configuration. MathVerse accuracy varies by only 0.2 points across the β sweep, while changing τ from 0.08 to 0.20 changes accuracy by only 0.3 points. Performance is more sensitive to the visual budget ρ, with $\rho = 0.4$ achieving the highest validation accuracy. These trends are consistent with the distinct roles of the three parameters: ρ controls total credit mass, τ controls routing concentration, and β controls residual support.

4.5 Additional Analysis

Visual Robustness and Training Cost. We evaluate Qwen3-VL-8B under five image corruptions and report end-to-end training cost under the same rollout budget. For robustness, we compare GRPO, VPPO, and RVCO; for efficiency, we report wall-clock time and GPU-hours for representative RLVR methods.

Table 4: Visual corruption robustness and end-to-end training cost on Qwen3-VL-8B. Robustness reports exact-match accuracy (%).

Robustness

Corruption	GRPO	VPPO	RVCO
Patch dropout (15%)	58.9	64.2	**66.8**
Gaussian blur	57.4	61.9	**64.1**
Low contrast	55.8	60.5	**62.9**
Unseen crop	52.6	58.1	**60.7**
Gaussian noise	54.9	59.8	**62.5**

Training Efficiency

Method	Time	GPU-h
GRPO	46h	368
DAPO	44h	352
VPPO	47h	376
PGPO	47.5h	380
RVCO	48h	384

Table 4 shows that RVCO outperforms VPPO by 2.2–2.7 points across all five corruptions, including gains of 2.6 points on unseen crops and 2.7 points under Gaussian noise. RVCO requires 48 hours (384 GPU-hours), compared with 47 hours (376 GPU-hours) for VPPO. Thus, the robustness gains are obtained with modest additional training cost.

End-of-Training Optimization Stability. We analyze the final 20% of Qwen3-VL-8B training updates to compare the stability of different credit geometries. We evaluate Hard Top-ρ, PGPO, VPPO, entropic routing without residual support ($\beta = 0$), and the complete RVCO. Support flicker measures changes in active credit support, while credit continuity measures similarity between neighboring allocation fields; we additionally report policy-entropy variance, gradient-norm coefficient of variation, and held-out reward.

Table 5: Optimization stability diagnostics over the final 20% of Qwen3-VL-8B updates. Lower support flicker, entropy variance, and gradient CV are better.

Method	Flicker ↓	Continuity ↑	Ent. Var. ↓	Grad. CV ↓	Reward ↑
Hard Top-ρ	0.219	0.184	0.047	0.384	0.751
PGPO	0.143	0.231	0.031	0.319	0.766
VPPO	0.126	0.246	0.024	0.298	0.758
Entropic routing ($\beta = 0$)	0.073	0.267	0.018	0.251	0.775
RVCO	**0.041**	**0.281**	**0.012**	**0.214**	**0.784**

Stability Analysis. Table 5 shows a clear progression from discrete selection to smooth routing and finally to full-support credit. Compared with Hard Top-ρ, entropic routing ($\beta = 0$) reduces support flicker from 0.219 to 0.073 and increases credit continuity from 0.184 to 0.267, indicating that smooth routing substantially stabilizes the allocation field. Adding residual support further reduces flicker to 0.041 and increases continuity to 0.281. The complete RVCO also achieves the lowest entropy variance (0.012) and gradient CV (0.214), together with the highest held-out reward (0.784). These results indicate that residual support provides additional stability beyond smoothing the routing boundary alone.

5 Conclusion

We introduced Residual Visual Credit Optimization (RVCO), which formulates multimodal token credit as a conserved evidence-routing problem. Given a fixed sequence-level utility, RVCO uses controlled visual interventions to estimate decision-relative visual evidence and allocates credit through budgeted entropic routing with residual full-support correction and exact mass conservation. The reward, rollout procedure, and group-relative advantage estimator remain unchanged; only the token-level credit geometry is modified. Across multiple model families and scales, RVCO improves multimodal reasoning accuracy while maintaining robustness, optimization stability, and modest training overhead. These results highlight a complementary dimension of multimodal RLVR: beyond determining how much utility a trajectory receives, effective optimization also depends on how that utility is distributed across the decisions that produced it.

REFERENCES

Brandon Amos, Vladlen Koltun, and J Zico Kolter. The limited multi-label projection layer. *arXiv preprint arXiv:1906.08707*, 2019.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4637–4646, 2025.

Yi Ding, Ziliang Qiu, Bolian Li, and Ruqi Zhang. Learning self-correction in vision-language models via rollout augmentation. *arXiv preprint arXiv:2602.08503*, 2026.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, et al. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

Gemini Team, Machel Reid, Nikolay Savinov, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Daya Guo, Dehan Yang, Haowei Zhang, Junxiao Song, 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, et al. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, 2026.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Qiming Li, Tianlun Li, Xiaolong Cheng, Hangyu Li, et al. Prpo: Perception-reinforced policy optimization via token-level dynamic advantage reshaping. *arXiv preprint arXiv:2606.08708*, 2026a.

Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, et al. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, et al. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, et al. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, et al. 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*, pp. 6774–6786, 2021.

Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, et al. Calibrating multimodal learning. In *International Conference on Machine Learning*, pp. 23429–23450, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, et al. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, et al. Improving vision-language models with perception-centric process reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *International Conference on Learning Representations*, 2024.

Dachuan Shi, Chaofan Tao, Ying Jin, Zhendong Yang, et al. Upop: Unified and progressive pruning for compressing vision-language transformers. In *International Conference on Machine Learning*, pp. 31292–31311, 2023.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 5158–5171, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

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*, 2025.

Xue Wang, Zhibo Wang, Haiqin Weng, Hengchang Guo, et al. Counterfactual-based saliency map: Towards visual contrastive explanations for neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2042–2051, 2023.

Yongxin Wang, Zhicheng Yang, Meng Cao, Mingfei Han, et al. Care what fails: Contrastive anchored-reflection for verifiable multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2026a.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, et al. Visually-guided policy optimization for multimodal reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 6646–6664, 2026b.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026c.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. Visnumbench: Evaluating number sense of multimodal large language models. *arXiv preprint arXiv:2503.14939*, 2025.

Sitong Wu, Haoru Tan, Bin Xia, Xichen Zhang, et al. Unlocking token rewards via training-free reward attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5082–5091, 2026.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 44791–44815, 2026.

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

Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. Differentiable top-k with optimal transport. *Advances in neural information processing systems*, 33:20520–20531, 2020.

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, et al. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *arXiv preprint arXiv:2604.01840*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, 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*, pp. 15134–15186, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, 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, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, et al. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, 2025.

A Optimization Properties

The constrained entropic routing step used in RVCO is a temperature-scaled instance of the Limited Multi-Label (LML) projection (Amos et al., 2019). It also belongs to the broader class of regularized prediction maps (Blondel et al., 2020). Related differentiable top-k constructions can be obtained through entropic optimal transport (Xie et al., 2020). We therefore do not attribute novelty to the projection primitive itself. We restate its standard optimization properties for completeness and then analyze the residual-support and exact-conservation operations used to construct the credit field in RVCO. For a response of length T and budget m, define the capped fixed-mass polytope

$$\mathcal{L}_{T,m} = \left\{ g \in [0,1]^T : \mathbf{1}^\top g = m \right\}, \tag{9}$$

and the binary entropy

$$H_{\mathrm{b}}(g) = \sum_{t=1}^{T} h(g_t), \qquad h(a) = -a \log a - (1-a)\log(1-a). \tag{10}$$

The routing map is

$$g_\tau(z) = \underset{g \in \mathcal{L}_{T,m}}{\arg\max} \left\{ z^\top g + \tau H_{\mathrm{b}}(g) \right\}, \qquad \tau > 0. \tag{11}$$

Dividing the objective by τ gives

$$g_\tau(z) = \underset{g \in \mathcal{L}_{T,m}}{\arg\max} \left\{ \left(\frac{z}{\tau}\right)^\top g + H_{\mathrm{b}}(g) \right\}, \tag{12}$$

which is exactly the LML projection of the temperature-scaled score vector z/τ under the binary-entropy regularizer (Amos et al., 2019). Equivalently, with

$$\Omega_{\tau,m}(g) = -\tau H_{\mathrm{b}}(g) + I_{\mathcal{L}_{T,m}}(g), \tag{13}$$

where $I_{\mathcal{L}_{T,m}}$ is the convex indicator of $\mathcal{L}_{T,m}$, Equation equation 11 has the regularized-prediction form

$$g_\tau(z) = \arg\max_{g} \left\{ z^\top g - \Omega_{\tau,m}(g) \right\}. \tag{14}$$

A.1 Existence, uniqueness, and scalar-dual solution

Consider first the non-degenerate case $0 < m < T$. The feasible set $\mathcal{L}_{T,m}$ is compact and convex, and the objective in Equation equation 11 is continuous. Hence a maximizer exists. Binary entropy is strictly concave, so the objective is strictly concave on the feasible affine set. The maximizer is therefore unique.

Moreover,

$$h'(a) = \log \frac{1-a}{a} \tag{15}$$

diverges to $+\infty$ as $a \to 0^+$ and to $-\infty$ as $a \to 1^-$. Thus the optimum lies in the relative interior for $0 < m < T$, and the box constraints are inactive at finite τ. Introducing a multiplier λ for $\mathbf{1}^\top g = m$, the Lagrangian is

$$\mathcal{J}(g, \lambda) = z^\top g + \tau H_{\mathrm{b}}(g) - \lambda(\mathbf{1}^\top g - m). \tag{16}$$

Stationarity gives

$$z_t + \tau \log \frac{1-g_t}{g_t} - \lambda = 0, \tag{17}$$

and hence

$$g_t = \sigma\left(\frac{z_t - \lambda}{\tau}\right), \qquad \sigma(x) = \frac{1}{1+\exp(-x)}. \tag{18}$$

The equality constraint reduces the complete optimization problem to the scalar equation

$$F(\lambda) = \sum_{t=1}^{T} \sigma\left(\frac{z_t - \lambda}{\tau}\right) = m. \tag{19}$$

Its derivative is

$$F'(\lambda) = -\frac{1}{\tau}\sum_{t=1}^{T} g_t(1-g_t) < 0. \tag{20}$$

Furthermore,

$$\lim_{\lambda\to-\infty} F(\lambda) = T, \qquad \lim_{\lambda\to+\infty} F(\lambda) = 0. \tag{21}$$

Therefore Equation equation 19 has exactly one solution.

The same result follows directly from the one-dimensional dual. Taking the supremum of the Lagrangian over each coordinate gives

$$\phi(\lambda) = m\lambda + \tau\sum_{t=1}^{T} \log\left(1 + \exp\frac{z_t - \lambda}{\tau}\right). \tag{22}$$

Its first and second derivatives are

$$\phi'(\lambda) = m - F(\lambda), \tag{23}$$

$$\phi''(\lambda) = \frac{1}{\tau}\sum_{t=1}^{T} g_t(1-g_t) > 0. \tag{24}$$

Thus ϕ is strictly convex and has a unique minimizer. Bisection on Equation equation 19 therefore recovers the unique primal solution. This is the scalar-dual structure of the LML projection (Amos et al., 2019).

For the boundary case $m = T$, the feasible set contains only $g = \mathbf{1}$, so uniqueness and conservation hold trivially.

A.2 Shift invariance and order preservation

The common dual variable makes the invariances of the routing map explicit. Consider a shifted score vector

$$z' = z + a\mathbf{1}, \tag{25}$$

for an arbitrary scalar a. If λ satisfies Equation equation 19 for z, then

$$\lambda' = \lambda + a \tag{26}$$

satisfies the same mass equation for z', because

$$\frac{z'_t - \lambda'}{\tau} = \frac{z_t - \lambda}{\tau}. \tag{27}$$

Consequently,

$$g_\tau(z + a\mathbf{1}) = g_\tau(z). \tag{28}$$

The router therefore depends only on relative evidence scores and not on an arbitrary response-level offset.

Order preservation follows immediately from the strict monotonicity of the sigmoid. For any two coordinates r and s,

$$z_r > z_s \quad\Longleftrightarrow\quad \frac{z_r - \lambda}{\tau} > \frac{z_s - \lambda}{\tau} \quad\Longleftrightarrow\quad g_r > g_s. \tag{29}$$

Thus entropic smoothing changes the magnitude of the allocation but not the ranking induced by the evidence scores.

A.3 Jacobian and finite-temperature stability

The scalar-dual form also yields an explicit Jacobian. Define

$$q_t = g_t(1-g_t), \qquad Q = \sum_{r=1}^{T} q_r. \tag{30}$$

Differentiating the mass constraint $\sum_t g_t = m$ with respect to z_s gives

$$\frac{\partial \lambda}{\partial z_s} = \frac{q_s}{Q}. \tag{31}$$

Using Equation equation 18,

$$\frac{\partial g_t}{\partial z_s} = \frac{q_t}{\tau}\left(\mathbb{I}[t=s] - \frac{q_s}{Q}\right). \tag{32}$$

In matrix form,

$$J_g = \frac{1}{\tau}\left(D_q - \frac{qq^\top}{Q}\right), \qquad D_q = \mathrm{diag}(q). \tag{33}$$

For any $v \in \mathbb{R}^T$,

$$v^\top J_g v = \frac{1}{\tau}\left[\sum_{t=1}^{T} q_t v_t^2 - \frac{\left(\sum_{t=1}^{T} q_t v_t\right)^2}{Q}\right] \geq 0, \tag{34}$$

where the inequality follows from Cauchy–Schwarz. Hence J_g is positive semidefinite. It also satisfies

$$J_g \mathbf{1} = 0, \tag{35}$$

which is the differential form of shift invariance and exact budget conservation.

Because $0 < q_t \leq 1/4$,

$$0 \preceq J_g \preceq \frac{1}{\tau} D_q \preceq \frac{1}{4\tau} I, \tag{36}$$

and therefore

$$\|J_g\|_2 \leq \frac{1}{4\tau}. \tag{37}$$

Finite temperature thus produces a smooth allocation map with a controlled local sensitivity. Decreasing τ sharpens the response and increases sensitivity near rank boundaries.

Equation equation 32 also exposes the competition induced by the fixed budget:

$$\frac{\partial g_t}{\partial z_t} = \frac{q_t}{\tau}\left(1 - \frac{q_t}{Q}\right) > 0, \tag{38}$$

while for $s \neq t$,

$$\frac{\partial g_t}{\partial z_s} = -\frac{q_t q_s}{\tau Q} < 0. \tag{39}$$

Increasing one score therefore increases its own allocation while slightly reducing the allocation available to the remaining positions. This interaction is absent from independent sigmoid gating.

A.4 Conserved residual support

The method-specific credit field is obtained after the entropic projection. For $0 < \beta < 1$, define

$$\bar{g}_t = \beta + (1-\beta) g_t. \tag{40}$$

Since $\sum_t g_t = m$,

$$\sum_{t=1}^{T} \bar{g}_t = \beta T + (1-\beta) m. \tag{41}$$

The exact mass-correction factor is therefore

$$c = \frac{m}{\beta T + (1-\beta) m}, \tag{42}$$

and Equation 6 can be written as

$$w_t = c\left[\beta + (1-\beta) g_t\right]. \tag{43}$$

The total routed credit is exactly conserved:

$$\sum_{t=1}^{T} w_t = c \sum_{t=1}^{T} \big[\beta + (1-\beta) g_t\big] \tag{44}$$

$$= \frac{m}{\beta T + (1-\beta)m} \left[\beta T + (1-\beta)m\right] \tag{45}$$

$$= m. \tag{46}$$

Moreover, $m \le T$ implies

$$0 < c = \frac{m}{\beta T + (1-\beta)m} \le 1. \tag{47}$$

Since $0 \le g_t \le 1$,

$$c\beta \le w_t \le c \le 1. \tag{48}$$

Hence every token receives strictly positive credit for $\beta > 0$, while no coefficient exceeds one. This establishes both full support and boundedness.

The residual transformation is affine and strictly increasing in g_t:

$$\frac{\partial w_t}{\partial g_t} = c(1-\beta) > 0. \tag{49}$$

It therefore preserves the ordering produced by the entropic router. Combining this fact with Equations equation 28 and equation 29 gives

$$z_r > z_s \quad \Longrightarrow \quad w_r > w_s, \qquad w(z + a\mathbf{1}) = w(z). \tag{50}$$

Because c depends only on (T, m, β), the Jacobian of the final credit field is

$$J_w = c(1-\beta) J_g. \tag{51}$$

Consequently,

$$\|J_w\|_2 \le \frac{c(1-\beta)}{4\tau} < \frac{1}{4\tau}. \tag{52}$$

Residual support therefore retains the ranking and mass constraint of the projected field while attenuating its local sensitivity.

A.5 Sparse and uniform limits

The temperature parameter continuously connects a hard budgeted selection with a uniform allocation.

First consider $\tau \to 0$. Since binary entropy is bounded on the unit interval,

$$0 \le H_{\mathrm{b}}(g) \le T \log 2. \tag{53}$$

Hence

$$\sup_{g \in \mathcal{L}_{T,m}} |\tau H_{\mathrm{b}}(g)| \le \tau T \log 2 \longrightarrow 0. \tag{54}$$

The regularized objective therefore converges uniformly to the linear program

$$\max_{g \in \mathcal{L}_{T,m}} z^\top g. \tag{55}$$

Let

$$z_{(1)} > z_{(2)} > \cdots > z_{(T)} \tag{56}$$

denote distinct coordinates sorted in descending order. If m is an integer, Equation equation 55 has the unique solution

$$g^0_{(t)} = \begin{cases} 1, & t \le m, \\ 0, & t > m. \end{cases} \tag{57}$$

Thus

$$g_\tau(z) \longrightarrow g^0 \qquad \text{as } \tau \to 0, \tag{58}$$

which is the hard top-m selector.

For completeness, if a relaxed non-integer budget $m = k + r$ is used with $k = \lfloor m \rfloor$ and $0 < r < 1$, the limiting capped solution is

$$g^0_{(t)} = \begin{cases} 1, & t \le k, \\ r, & t = k+1, \\ 0, & t > k+1. \end{cases} \tag{59}$$

The integer-budget case reduces exactly to Equation equation 57.

Next consider the residual parameter. From Equation equation 42,

$$\lim_{\beta \to 0} c = 1, \tag{60}$$

and therefore

$$\lim_{\beta \to 0} w_t = g_t. \tag{61}$$

Combining the two limits gives the hard selector stated in Proposition 1 for distinct scores and integer m:

$$\lim_{\substack{\tau \to 0 \\ \beta \to 0}} w = g^0. \tag{62}$$

At the opposite extreme, $\tau \to \infty$ suppresses all finite score differences. From Equation equation 18,

$$\mathrm{logit}(g_t) - \mathrm{logit}(g_s) = \frac{z_t - z_s}{\tau} \longrightarrow 0. \tag{63}$$

All coordinates therefore approach a common value. The mass constraint determines this value uniquely:

$$g_t \longrightarrow \frac{m}{T}. \tag{64}$$

Substituting this limit into Equation equation 43 yields

$$w_t \longrightarrow \frac{m}{\beta T + (1-\beta)m} \left[\beta + (1-\beta)\frac{m}{T}\right] \tag{65}$$

$$= \frac{m}{T}. \tag{66}$$

Thus both the projected gate and the final conserved field converge to the same uniform allocation. The complete path from sparse selection to uniform credit preserves total mass m.

A.6 Exact conservation versus unit-mean normalization

The distinction between exact mass conservation and unit-mean normalization is algebraic rather than cosmetic. Let

$$a_t = \beta + (1-\beta)g_t, \qquad S = \sum_{t=1}^{T} a_t = \beta T + (1-\beta)m. \tag{67}$$

Exact conservation uses

$$w_t^{\mathrm{mass}} = \frac{m}{S} a_t, \qquad \sum_{t=1}^{T} w_t^{\mathrm{mass}} = m. \tag{68}$$

Unit-mean normalization instead gives

$$w_t^{\mathrm{mean}} = \frac{T}{S} a_t, \qquad \frac{1}{T}\sum_{t=1}^{T} w_t^{\mathrm{mean}} = 1. \tag{69}$$

The two fields have identical relative ordering, but their global mass differs by the exact factor

$$w_t^{\mathrm{mean}} = \frac{T}{m} w_t^{\mathrm{mass}}. \tag{70}$$

For $m = \rho T$, this becomes

$$w_t^{\text{mean}} = \frac{1}{\rho} w_t^{\text{mass}}. \tag{71}$$

Unit-mean normalization therefore fixes the total coefficient mass at T, whereas the proposed correction fixes it at the prescribed routing budget m. In a weighted policy objective, these choices induce different global scales even when their token rankings are identical.

The role of exact conservation in RVCO is therefore specific: the residual floor can restore support to low-ranked tokens without changing the total credit assigned to the response. The allocation geometry is inherited from the known entropic projection, while residual support and analytic mass correction determine how that projection is converted into a conserved credit field for multimodal RLVR.

B Detailed Experimental Setup

B.1 Training protocol

Each prompt produces eight candidate responses during online reinforcement learning. A rule-based verifier assigns binary accuracy rewards and response groups without reward variation are removed by online filtering. The vision tower remains trainable. Policy updates use lower and upper clipping ratios of 0.20 and 0.28, weight decay 10^{-2}, zero warmup, maximum gradient norm 1.0, and global batch size 128. Sampling uses temperature 1.0 and top-p 0.99. Prompt and response limits are 4096 and 8192 tokens. Training uses BF16, activation checkpointing, fully sharded data parallelism, and padding-free packing. The default visual intervention independently suppresses image patches with probability 0.5. Evidence responses are clipped before logarithmic compression; the robust scale is lower-bounded by 10^{-6} and coordinates by $[-8, 8]$. All gate roots use 32 bisection iterations.

Table 6: Additional reinforcement-learning hyperparameters.

Setting	Value	Setting	Value
Rollouts / prompt	8	Global batch	128
Rollout temperature	1.0	Clip lower / upper	0.20 / 0.28
Top-p	0.99	Weight decay	10^{-2}
Prompt length	4096	Grad. norm	1.0
Response length	8192	Precision	BF16
Vision tower	Trainable	Sharding	Full FSDP
Rollout batch	384	Epochs	2
Learning rate	10^{-6}	Bisection steps	32

B.2 Baseline implementation

GRPO uses group-relative reward normalization with the same asymmetric clipping as the primary trainer. DAPO adopts its sampling and clipping refinements without adding an auxiliary value model. PAPO is instantiated with its perception-aware objective on original and corrupted image pairs (Wang et al., 2026c). VPPO uses token visual dependency to shape trajectory advantages and select pivotal token gradients (Huang et al., 2026); PGPO applies token-level visual-dependency advantage reshaping with its mass-preserving design (Ye et al., 2026); PRPO uses bounded dynamic advantage reshaping (Li et al., 2026a); VL-Calibration combines visual certainty and reasoning confidence (Xiao et al., 2026); and VGPO applies intra- and inter-trajectory visual-guidance reweighting (Wang et al., 2026b). For objective-side comparisons, all methods share the same rollout samples, verifier, base learning rate, response limits, and optimizer so that the dominant difference is the credit mechanism.

B.3 Evaluation protocol

The seven in-domain benchmarks use their official answer-extraction procedures and exact-match scoring. MathVerse, DynaMath, MMK12, Geo3K, and VisNumBench emphasize visual mathematical, geometric, or numerical reasoning; LogicVista stresses visual logical inference; MMMU-Pro evaluates multidisciplinary multimodal understanding. Out-of-domain tests use A-OKVQA (Schwenk et al., 2022), SimpleVQA (Cheng et al., 2025), ChartQA (Masry et al., 2022), InfographicVQA (Mathew et al., 2022), and DocVQA (Mathew et al., 2021). Corruption diagnostics are generated from the evaluation images only and do not alter the ground-truth answer or textual question.

B.4 CREDIT DIAGNOSTICS

Support flicker compares matched top-ρ active sets under paired mild visual interventions. If S_i and $\widetilde{S}_i$ are the active sets induced by two interventions, flicker is $1 - |S_i \cap \widetilde{S}_i|/|S_i \cup \widetilde{S}_i|$. Credit continuity is the cosine similarity between the dense allocation fields. Gradient CV is the coefficient of variation of per-minibatch gradient norms over the final 20% of updates. Entropy variance is computed over the same window from the policy's generated-token entropy. These diagnostics are measured on held-out trajectories and do not enter the training objective.

B.5 FULL QWEN3-VL-8B BASELINE COMPARISON

To verify that the main improvement remains competitive against a broader set of multimodal RLVR objectives, we extend the baseline pool beyond the compact comparison in the main text. Qwen3-VL-8B is trained on ViRL39K and evaluated on the same seven benchmarks with Instruction, GRPO, DAPO, Vision-R1, MM-Eureka, PAPO, PGPO, PRPO, VPPO, and RVCO under the protocol in Appendix B.

Table 7: Extended comparison on Qwen3-VL-8B. Values are accuracy@8 (%); Avg. denotes the macro-average over all seven benchmarks. Bold and underlined values denote the best and second-best results, respectively.

Method	MathV.	Dyna	MMK12	Geo3K	VisNum	Logic	MMMU-P	Avg.
Instruction	67.7	72.2	62.0	57.5	25.0	50.1	33.7	52.6
GRPO	76.5	75.0	71.2	64.4	28.7	56.0	42.8	59.2
DAPO	78.2	76.1	73.0	66.0	28.4	60.2	46.1	61.1
Vision-R1	79.0	76.8	74.1	66.8	28.9	58.0	50.7	62.0
MM-Eureka	80.1	78.5	76.0	<u>72.3</u>	29.5	61.1	52.1	64.2
PAPO	82.1	79.0	78.7	70.8	29.4	61.4	51.0	64.6
PGPO	82.7	79.4	79.5	71.1	<u>30.0</u>	61.8	51.9	65.2
PRPO	82.5	79.2	79.0	71.0	29.6	61.6	51.5	64.9
VPPO	<u>83.8</u>	<u>80.0</u>	<u>80.1</u>	71.9	29.1	<u>62.3</u>	<u>52.4</u>	<u>65.7</u>
RVCO	**84.4**	**80.4**	**80.6**	**72.5**	**30.5**	**64.1**	**55.4**	**66.8**

Table 7 shows that RVCO achieves the strongest overall performance across the extended baseline pool. The improvement remains consistent across mathematical reasoning, logical reasoning, and multidisciplinary evaluation. The method obtains the best result on all seven benchmarks and the highest overall average.

C EXTENDED CREDIT DIAGNOSTICS

To separate support preservation from simple smoothing or changes in total credit mass, we first measure the geometry of the learned credit fields directly. On held-out Qwen3-VL-8B trajectories, Hard Top-ρ, independent sigmoid weighting, the entropic gate with $\beta = 0$, and RVCO are evaluated under paired mild interventions using continuity, support overlap, weight Gini, and mass error.

Table 8: Evidence-retention diagnostics on held-out trajectories. Credit continuity is cosine similarity between allocation fields under paired mild interventions; support overlap uses matched top-ρ sets.

Method	Continuity ↑	Support overlap ↑	Weight Gini ↓	Mass error ↓
Hard Top-ρ	0.184	0.621	0.604	0.000
Independent sigmoid	0.251	0.652	0.421	0.087
Entropic gate ($\beta = 0$)	0.267	0.638	0.468	0.000
RVCO	**0.281**	**0.807**	**0.392**	**0.000**

Table 8 shows that independent sigmoid weighting improves continuity to 0.251 but incurs 0.087 mass error, whereas Hard Top-ρ preserves mass exactly but has only 0.184 continuity. RVCO combines exact mass with the highest continuity (0.281), the highest support overlap (0.807), and the lowest weight Gini (0.392). The residual correction therefore improves support without changing the total routed mass.

To examine how this support difference evolves during optimization, we next track matched support overlap throughout training. The analysis uses Qwen3-VL-8B held-out trajectories and compares Hard Top-ρ, the entropic gate with $\beta = 0$, and RVCO at matched training checkpoints under the same paired perturbations.

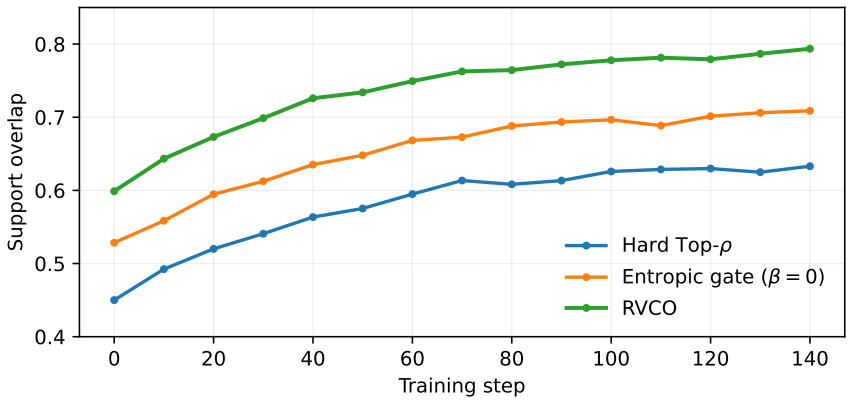

Figure 4: Support overlap during training under matched perturbations. The residual route maintains a larger stable intersection of credited positions as evidence becomes sharper late in training.

Figure 4 shows that the three methods are closer early in training, but the gap widens as optimization proceeds. RVCO retains the highest overlap in the late stage, while Hard Top-ρ remains substantially lower and the entropic-only variant lies between them. This trajectory-level pattern matches the final-window stability results in Table 5.

To determine whether improved support stability is accompanied by better optimization progress, we also track held-out reward over training. Qwen3-VL-8B is trained with PGPO, VPPO, and RVCO under the shared protocol, and the dashed marker denotes the same late-stage window used for Table 5.

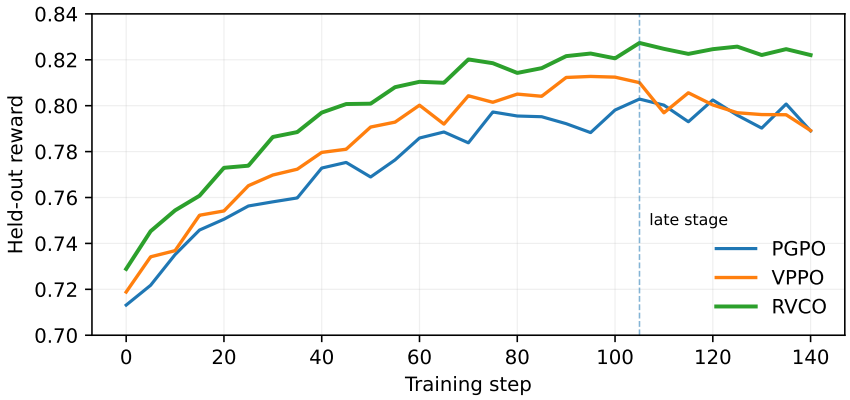

Figure 5: Held-out reward across training for representative perception-aware methods. The vertical marker denotes the late-stage window used for stability diagnostics.

Figure 5 shows similar early gains for all three methods, followed by clearer separation after the late-stage marker. RVCO continues to increase toward the highest held-out reward, while VPPO bends downward and PGPO saturates earlier. This behavior is consistent with the lower entropy variance and gradient CV reported for the complete method.

To test the role of routing sharpness directly, we sweep the gate temperature over a $16\times$ range. Qwen3-VL-4B is trained with $\tau \in \{0.03, 0.06, 0.12, 0.24, 0.48\}$ while all other settings remain fixed, and the training score is recorded through 140 steps.

Figure 6 shows that $\tau = 0.06$, 0.12, and 0.24 follow closely matched trajectories, while the sharpest setting $\tau = 0.03$ becomes more variable in the late stage and $\tau = 0.48$ is slightly more conservative. The result supports a broad intermediate regime rather than a single narrow optimum.

To provide the corresponding validation view, we repeat the one-at-a-time sweep of β, ρ, and τ used in the main text. The experiment uses Qwen3-VL-8B on ViRL39K and reports training and MathVerse accuracy with the other parameters fixed at their defaults.

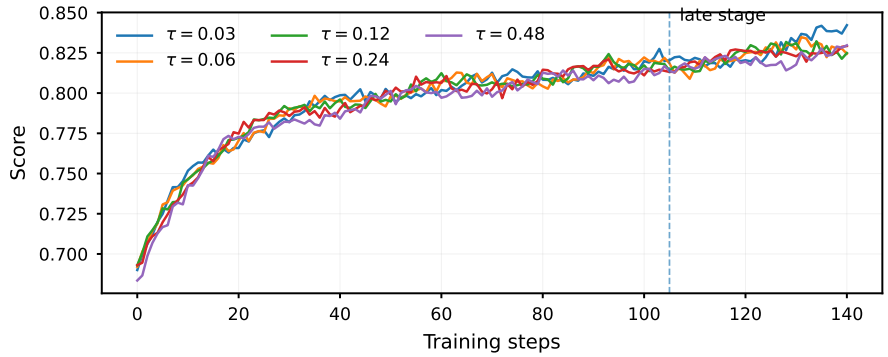

Figure 6: Training dynamics over a 16× range of routing temperatures on Qwen3-VL-4B. Moderate temperatures preserve evidence contrast without abrupt late-stage changes.

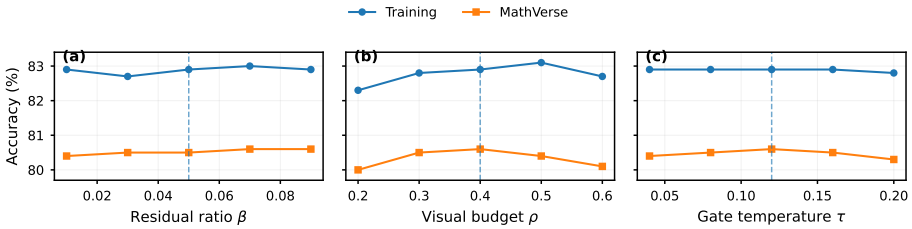

Figure 7: Sensitivity to residual ratio β, visual budget ρ,and gate temperature τ on Qwen3-VL-8B. The dashed line marks the default.

Figure 7 confirms the same broad plateau reported in the main text: performance changes little across the β and τ ranges, while $\rho = 0.4$ gives the strongest joint validation point. This agreement across training and validation metrics indicates that the default configuration is not selected from an isolated spike.

D Out-of-Domain Generalization

Out-of-Domain Generalization. To test whether reasoning-oriented credit optimization preserves visual-language capability outside the training-style reasoning suite, we evaluate unseen VQA and document tasks. Qwen3-VL-4B and Qwen3-VL-8B are evaluated on A-OKVQA, SimpleVQA, ChartQA, InfoVQA, and DocVQA using each benchmark's official primary metric, with the instruction model, GRPO, and RVCO compared under the same decoding protocol.

Table 9: Out-of-domain visual question answering and document understanding. Values use each benchmark's official primary metric (%).

Backbone	Method	A-OKVQA	SimpleVQA	ChartQA	InfoVQA	DocVQA	Avg.
Qwen3-VL-8B	Instruction	87.5	43.9	82.4	55.6	89.7	71.8
	GRPO	87.3	42.7	82.8	55.1	89.1	71.4
	RVCO	**88.7**	**47.8**	**84.9**	**59.4**	**91.0**	**74.4**
Qwen3-VL-4B	Instruction	86.3	38.4	79.6	51.2	87.1	68.5
	GRPO	86.2	37.1	79.9	50.8	86.5	68.1
	RVCO	**87.3**	**42.0**	**82.1**	**55.0**	**88.4**	**71.0**

Table 9 shows that RVCO raises the five-task average from 71.8 to 74.4 on Qwen3-VL-8B and from 68.5 to 71.0 on Qwen3-VL-4B relative to the instruction models. On the 8B backbone, the largest gains are on SimpleVQA (+3.9) and InfoVQA (+3.8), while DocVQA also improves by 1.3 points. The improvement therefore extends beyond the in-domain reasoning benchmarks rather than being confined to their answer format.

E Response-Length Analysis

To test whether preserving weak but nonzero credit becomes more important as reasoning trajectories grow longer, we stratify evaluation by generated-response length. Qwen3-VL-8B is evaluated on the seven-task suite with GRPO, VPPO, and RVCO, and responses are grouped into ≤ 512, 513–2048, and > 2048 token buckets after answer extraction.

Table 10: Accuracy by generated-response length on the seven-task reasoning suite. Length buckets are measured after answer extraction.

Length bucket	GRPO	VPPO	RVCO	Gain over VPPO
≤ 512 tokens	64.9	69.1	**70.4**	+1.3
513–2048 tokens	59.5	66.0	**67.6**	+1.6
> 2048 tokens	53.2	60.6	**63.0**	+2.4

Table 10 shows that RVCO improves VPPO by 1.3 points for responses up to 512 tokens, 1.6 points for 513–2048 tokens, and 2.4 points beyond 2048 tokens. Although accuracy falls with length for every method, the increasing margin indicates that full-support routing is most useful when the reasoning path contains more intermediate and revisited states.

F Rollout Group and Verifier Sensitivity

To determine whether the token router depends on a particular group-relative advantage resolution, we vary rollout-group size. Qwen3-VL-8B is trained with 4, 8, and 16 responses per prompt while approximately matching the total generated-token budget through gradient accumulation; GRPO, VPPO, and RVCO use the same settings at each group size.

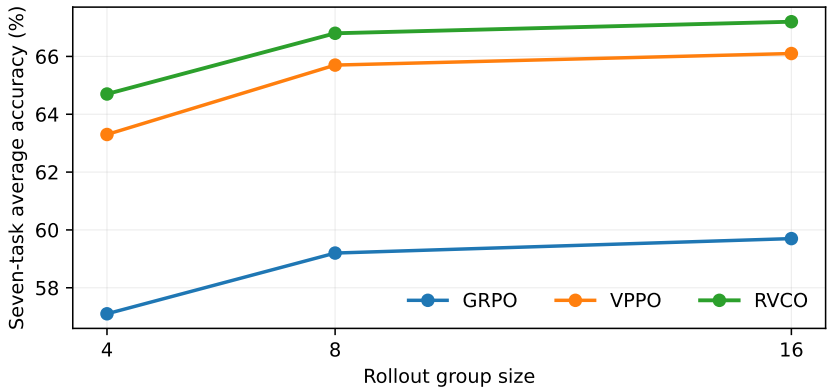

Figure 8: Sensitivity to rollout-group size under approximately matched generated-token budgets.

Figure 8 shows that RVCO improves from 64.7 to 67.2 as the group size increases and remains 1.1–1.4 points above VPPO at every setting. The stable margin indicates that token-level conserved routing complements, rather than replaces, the variance reduction obtained from larger rollout groups.

To test sensitivity to imperfect outcome supervision, we next corrupt the verifier labels directly. On Qwen3-VL-8B, binary rewards are flipped at 0%, 2%, 5%, and 10% before group-relative normalization; GRPO, VPPO, and RVCO receive the same corrupted reward stream at each flip rate.

Figure 9 shows the expected decline for all methods as label noise grows, but RVCO remains above VPPO by 1.1 points at 0% flips and 2.2 points at 10% flips. The method does not correct erroneous rewards; the persistent margin instead shows that the credit operator remains beneficial when the sequence-level signal is moderately noisy.

G Intervention Design and Corruption Severity

To determine whether the evidence coordinate depends on one particular visual perturbation, we compare several intervention operators. Qwen3-VL-8B is trained on ViRL39K with random patch

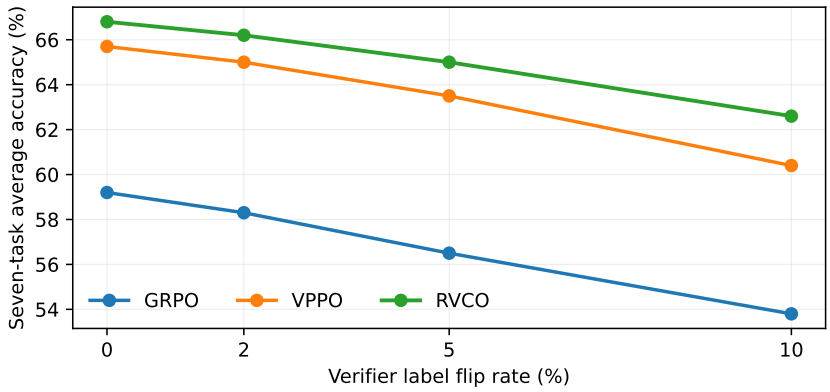

Figure 9: Sensitivity to controlled verifier label noise. All methods use the same corrupted reward labels at each flip rate.

masking, Gaussian noise, Gaussian blur, random cutout, or grayscale attenuation, and each variant is evaluated by seven-task average accuracy, credit continuity, and wall-clock overhead relative to GRPO.

Table 11: Sensitivity to the visual intervention used to construct evidence coordinates on Qwen3-VL-8B. Overhead is relative to GRPO wall-clock time.

Intervention	Avg. acc. ↑	Continuity ↑	Overhead
Random patch mask (default)	**66.8**	**0.281**	4.3%
Gaussian noise	66.3	0.275	4.1%
Gaussian blur	66.0	0.270	4.0%
Random cutout	66.1	0.272	4.2%
Grayscale attenuation	65.4	0.264	3.9%

Table 11 shows that random patch masking gives the best average accuracy (66.8) and continuity (0.281), but Gaussian noise, blur, and cutout remain within 0.5–0.8 accuracy points of the default. All five interventions add about 4% wall-clock overhead. The method therefore does not depend on a single corruption family, although localized patch masking provides the strongest overall result.

To test robustness beyond a single corruption operating point, we next vary the severity of patch corruption continuously. Qwen3-VL-8B is evaluated at 0%, 10%, 20%, 30%, and 40% patch corruption with GRPO, VPPO, and RVCO under identical decoding and answer extraction.

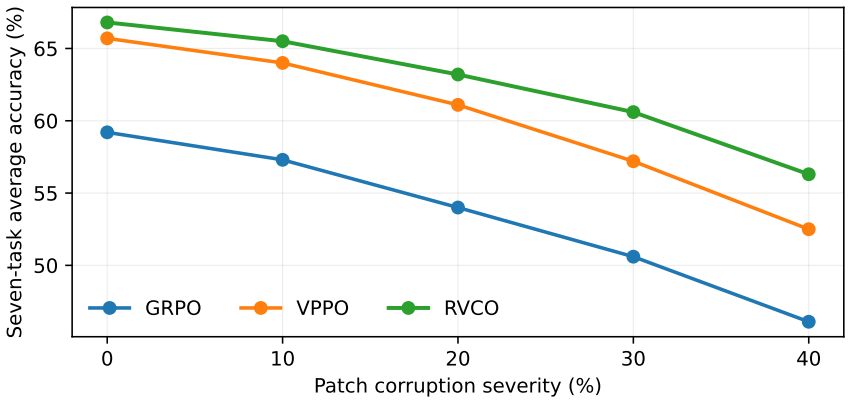

Figure 10: Accuracy under increasing patch corruption severity. The full-support field degrades more gradually as the visual signal becomes incomplete.

Figure 10 shows that RVCO declines from 66.8 to 56.3 as corruption increases to 40%, while VPPO falls from 65.7 to 52.5. The margin therefore grows from 1.1 points without corruption to 3.8 points at the highest severity. Full-support routing does not recover missing pixels, but it degrades more gradually when the visual evidence becomes diffuse.

H Data Efficiency and Allocation Case Study

To verify the low-data trend independently of the main table, we plot reasoning accuracy across the full ViRL39K fraction sweep. Qwen3-VL-8B is trained with 25%, 50%, 75%, and 100% of the corpus using GRPO, VPPO, and RVCO under the same optimizer and seven-task evaluation protocol.

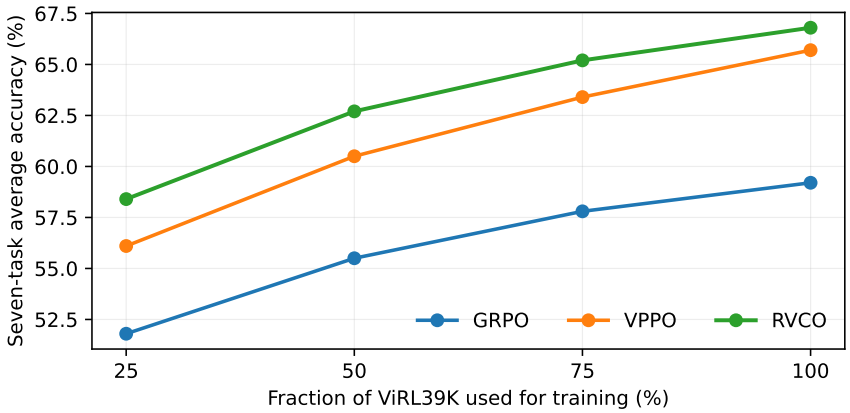

Figure 11: Reasoning accuracy as the ViRL39K training fraction increases. The proposed field retains an advantage throughout the data range.

Figure 11 shows that RVCO rises from 58.4 at 25% data to 66.8 at full data and remains above VPPO throughout. The gain over VPPO decreases from 2.3 points at 25% data to 1.1 points at full data, while accuracy remains monotonic for all methods. The larger low-data margin indicates that conserved routing is especially useful when verifiable supervision is limited.

To inspect how the same fixed budget is distributed at token level, we also perform a paired qualitative allocation analysis. A Qwen3-VL-8B response is held fixed and scored under Hard Top-ρ and RVCO with the same 40% total credit budget, so the comparison changes only the allocation rule rather than the generated text or reward.

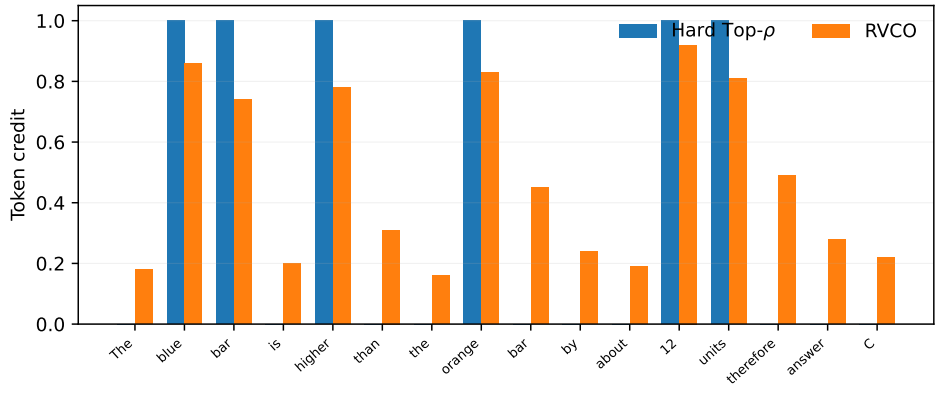

Figure 12: Representative token-credit allocation for a visual comparison response under the same 40% total budget. Hard routing alternates between unit and zero credit, whereas RVCO preserves graded support around visually grounded and connective tokens.

Figure 12 shows that the hard selector alternates between unit and zero weights, including abrupt changes around adjacent response tokens. RVCO assigns the largest weights to visually grounded content while retaining smaller positive weights on connective tokens under the same total budget. The example therefore illustrates full support without uniform credit: evidence-bearing tokens remain prioritized, but neighboring reasoning positions are not deleted from the update.

I Additional Implementation Notes

The credit computation is implemented after original and intervened token log-probabilities have been materialized. Valid-token masks are reused from the PPO batch; quantile statistics ignore padding and prompt tokens. The bisection interval is initialized from the minimum and maximum calibrated evidence coordinates with a fixed numerical margin, which avoids a per-response search for custom bounds. Thirty-two iterations are sufficient for the mass error to fall below the precision of the subsequent BF16 policy update.

For distributed execution, the intervened forward pass can be scheduled alongside reward-side diagnostics because it does not require policy gradients. The resulting $d_{i,t}$ and $z_{i,t}$ tensors are detached before the PPO minibatch is formed. This keeps the computational graph identical to the base policy update and avoids an additional backward pass through the intervention branch. The extra persistent storage is one scalar evidence coordinate and one scalar weight per generated token; temporary full-vocabulary logits can be released immediately after the sampled-token log-probability is gathered.

The budget ratio ρ is applied to valid response positions rather than padded sequence length. For very short responses, $m_i = \lceil \rho T_i \rceil$ ensures at least one unit of routed mass. For long responses, the linear scaling keeps the average gradient contribution comparable across examples. Residual support does not change this amount because the analytic correction is computed from T_i, m_i, and β in closed form.

J Limitations and Broader Impact

Limitations. RVCO requires an additional likelihood evaluation under a visually intervened input, introducing modest overhead relative to outcome-only GRPO. The evidence response measures conditional sensitivity to a chosen intervention distribution rather than a formal causal effect, and the intervention may underrepresent visual evidence that is robust to masking or distributed across large image regions. The visual budget ρ, gate temperature τ, and residual ratio β are fixed throughout training; adaptive schedules or response-dependent settings may be beneficial for unusually long trajectories or tasks in which visual dependence is highly sparse. The present study focuses on static-image tasks with verifiable outcomes; extending conserved evidence routing to video, embodied interaction, multi-image settings, and open-ended preference-based rewards remains future work.

Broader impact. More stable multimodal post-training may improve visual reasoning, document analysis, scientific assistance, and educational systems. The same mechanisms can nevertheless reinforce verifier biases or visual misconceptions when the reward is incomplete. Full-support routing preserves access to evidence but does not itself guarantee faithful reasoning, so deployment should continue to include calibration, corruption testing, and hallucination evaluation (Ma et al., 2023; Guan et al., 2024; Cheng et al., 2025).

K Use of Large Language Models

Large language models assisted manuscript organization, language revision, citation-format normalization, figure ideation, and layout checking. The authors remain responsible for the research design, implementation, experiments, numerical results, and scientific claims.