SELECTIVE TRANSFER OF RL UPDATES FOR VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging provides a training-free way to transfer reasoning capabilities from language models to vision-language models (VLMs), but endpoint-based transfer can conflate pre-existing model differences with changes acquired during reasoning post-training. We instead formulate capability transfer around the training-stage update, isolating the parameter changes induced by reinforcement learning (RL). Yet transferring this update in full remains suboptimal: we find that its components differ substantially in cross-model transferability, with dominant directions transferring more effectively than the complete update. Based on this finding, we introduce Selective-RL, which isolates the RL-stage update, retains its dominant matrix-wise directions with magnitude preservation, and transfers them to the language modules of a VLM. Across three model families and five visual-reasoning benchmarks, Selective-RL improves full-update interpolation in 12 of 15 comparisons, including an 8.55 percentage-point MathVision gain on the Qwen recipient. Matched controls show that update magnitude or arbitrary low rank alone does not reproduce these gains. These results highlight a distinction between what is acquired during post-training and what remains transferable across models, providing a training-stage perspective on cross-model capability transfer. Code is available at https://anonymous.4open.science/r/selective-rl.

1 INTRODUCTION

Model merging transfers capabilities across pretrained models without additional training (Marczak et al., 2025; Fang et al., 2025), including reasoning transfer from language models to vision-language models (VLMs) (Chen et al., 2025; Hu et al., 2025; Huang et al., 2026a). Endpoint-based transfer constructs a donor–recipient displacement, while parameter selection, projection, and spectral filtering address interference (Gargiulo et al., 2025; Lee et al., 2025; 2026). Yet the donor endpoint alone cannot identify reasoning post-training changes. Analyses of base-model capabilities and reinforcement learning (RL) distinguish pre-existing reasoning behavior from post-training effects (Liu et al., 2025b; Yue et al., 2025; Yu et al., 2025). For a reasoning donor and a multimodal recipient, their displacement can therefore mix differences that predate RL with changes acquired during RL.

This distinction becomes explicit when we track parameter provenance, as in task-vector and alignment transfer (Ilharco et al., 2023; Huang et al., 2024) and interpolation between reward-specialized models (Ramé et al., 2023). Let V denote the multimodal recipient, B the donor's actual pre-RL checkpoint, and R its post-trained counterpart. On their aligned language parameters,

$$\underbrace{\mathrm{R}-\mathrm{V}}_{\text{endpoint displacement}} = \underbrace{\mathrm{B}-\mathrm{V}}_{\text{pre-RL displacement}} + \underbrace{\mathrm{R}-\mathrm{B}}_{\text{RL-stage update}} .$$

Here $\mathrm{B}-\mathrm{V}$ predates RL, whereas $\mathrm{R}-\mathrm{B}$ isolates RL-acquired changes. Merging depends on parameter importance and functional mismatch (Matena & Raffel, 2022; Jin et al., 2023; Daheim et al., 2024); task information can occupy localized directions or parameter supports (Ortiz-Jimenez et al., 2023; Wang et al., 2024a). Selective parameter transfer can also preserve existing multimodal capabilities (Zhu et al., 2024). These findings motivate isolating the RL-stage update before testing which components transfer; stage attribution alone does not establish that every component encodes transferable reasoning.

RL-stage isolation is necessary but insufficient: $\mathrm{R}-\mathrm{B}$ was optimized for the source language model, so its components may differ in usefulness to a multimodal recipient. This question is particularly

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.

relevant because RL updates in reasoning models exhibit pronounced spectral structure, with dominant directions capturing a disproportionate part of their learning dynamics (Cai et al., 2026). Prior work primarily examines this structure through source-model reconstruction or trajectory characterization; whether it identifies components compatible across models remains unclear. In the primary Qwen controls, the dominant matrix-wise direction outperforms the complete update at matched Frobenius magnitude; retaining more directions progressively reduces MathVista accuracy. Thus, components of the same post-training update can differ substantially in cross-model transferability, and preserving more of the source update need not be beneficial.

Motivated by this finding, we introduce **Selective-RL**, a training-free framework for transferring reasoning capabilities from RL-trained language models to multimodal recipients. It first extracts the RL-stage update $\mathrm{R} - \mathrm{B}$, excluding parameter differences that predate reasoning post-training. For each eligible language-layer matrix, it then retains the dominant singular direction and rescales it to preserve the magnitude of the original matrix-wise update. We compare direct injection and reconstructed-donor interpolation on aligned language modules, preserving visual components. The interpolation comparison holds pre-RL displacement fixed. Unlike multimodal post-training approaches (Deng et al., 2025; Huang et al., 2026b; Wang et al., 2025), Selective-RL requires no multimodal post-training, recipient gradients, additional modules, or target-side adaptation, and preserves the VLM's architecture and inference structure.

We evaluate Selective-RL on three multimodal recipients spanning the Qwen (Bai et al., 2025), LLaMA (Grattafiori et al., 2024), and Mistral (Jiang et al., 2023) families across five visual-reasoning benchmarks. It improves matched full-update interpolation in **12/15** backbone–benchmark comparisons (Table 1). On Qwen3-VL, MathVision rises $40.46 \rightarrow 49.01\%$ ($+8.55$ **pp**), MathVerse $45.30 \rightarrow 52.00\%$ ($+6.70$ pp), and MathVista $76.70 \rightarrow 81.00\%$ ($+4.30$ pp). On MathVista, norm-matched leading directions reach 81.00%, versus $74.05 \pm 0.45\%$ for ten matched random-direction runs and 73.10% for trailing directions (Table 2). Increasing rank from one to sixteen lowers accuracy to 78.30%, while gains over Full average 4.13 ± 0.15 pp across three donor seeds (Tables 8 and 22). These results distinguish source-update preservation from recipient improvement: selecting transferable structure can be more effective than retaining the complete update, with task-dependent exceptions.

2 Related Work

Model Merging. Task Arithmetic (Ilharco et al., 2023) composes checkpoint differences. TIES-Merging (Yadav et al., 2023) trims updates and resolves sign disagreement; DARE (Yu et al., 2024) drops and rescales delta parameters. TSV (Gargiulo et al., 2025) compresses task matrices and decorrelates singular vectors. LoRE-Merging (Liu et al., 2025a) estimates a reference and low-rank task vectors without the original base. STAR (Lee et al., 2025) truncates by cumulative spectral mass and restores nuclear norms. AdaRank (Lee et al., 2026) learns component masks by entropy minimization on unlabeled test inputs. We compare adaptive selection and scaling with fixed leading-component transfer of the actual RL update under Frobenius matching.

Multimodal Reasoning. OpenVLThinker (Deng et al., 2025) alternates supervised fine-tuning and RL. Vision-R1 (Huang et al., 2026b) combines a reasoning cold start, GRPO, and Progressive Thinking Suppression Training. VL-Rethinker (Wang et al., 2025) uses Selective Sample Replay and Forced Rethinking. Bring Reason to Vision (Chen et al., 2025) merges language parameters while retaining visual components. IP-Merging (Hu et al., 2025) identifies, rescales, and projects reasoning-associated parameters into recipient subspaces. FRISM (Huang et al., 2026a) learns subspace coefficients through label-free multimodal self-distillation. We separate RL updates from initialization displacement without recipient gradients, retaining validation-selected coefficients.

RL Updates. Source recovery, generation behavior, and recipient improvement require separate evidence. DeepSeek-R1 (DeepSeek-AI et al., 2025) contrasts direct RL with multistage training, making the reference checkpoint essential. DAPO (Yu et al., 2025) studies decoupled clipping and dynamic sampling. Dr. GRPO (Liu et al., 2025b) examines base-model behavior and response-length biases. Large-budget pass@k analysis (Yue et al., 2025) distinguishes sampling efficiency from solution coverage. Cai et al. (2026) study low-rank RL reconstruction and introduce AlphaRL for trajectory extrapolation. Frankenstein-style analysis (Li et al., 2026) probes visual-RL refinements

through causal interventions, checkpoint exchange, and freezing. We control direction, magnitude, and initialization displacement; spectral concentration alone identifies neither causal mechanisms nor RL-exclusive semantic components.

3 METHODOLOGY

Our goal is to transfer the capability acquired during the donor's RL stage while separating it from parameter differences that predate RL. We isolate the RL-stage update from initialization displacement (Section 3.1), select its dominant matrix-wise direction at controlled magnitude (Section 3.2), and compose it with the recipient through two transfer routes (Section 3.3). Section 3.4 characterizes update scale and resource use.

3.1 PROBLEM DEFINITION AND STAGE ISOLATION

We begin by separating two sources of parameter change that are mixed in conventional donor–recipient merging. Let $V = (V_j)_{j\in\mathcal{I}}$ denote the multimodal recipient, B the actual checkpoint from which RL training starts, and R the resulting reasoning donor. Let $\mathcal{S} \subseteq \mathcal{I}$ contain the transferable attention and MLP matrices that are aligned between donor and recipient, with $V_j, B_j, R_j \in \mathbb{R}^{m_j\times n_j}$ for $j \in \mathcal{S}$. We define $\Gamma_j := B_j - V_j$ and $\Delta_j := R_j - B_j$ for $j \in \mathcal{S}$.

Here, Γ_j captures parameter differences that already exist before RL, whereas Δ_j isolates the update acquired during RL. This separates RL-acquired changes from pre-existing donor–recipient mismatch.

The support $\mathcal{S}$ contains the aligned $q/k/v/o$ projections in attention blocks and the gate/up/down projections in MLP blocks, matched by layer and module with consistent orientation and head ordering. Parameters outside $\mathcal{S}$, including the recipient's visual components, remain unchanged.

To express different transfer strategies in a common form, we define zero extension for any collection $H = (H_j)_{j\in\mathcal{S}}$, $[\mathcal{E}_{\mathcal{S}}(H)]_j = H_j$ for $j \in \mathcal{S}$ and 0 otherwise, and write the merged model as

$$W(\alpha, \beta; Z) = V + \mathcal{E}_{\mathcal{S}}(\alpha\Gamma + \beta Z), \tag{1}$$

where Z is the RL-derived signal to be transferred. The coefficients α and β independently control the initialization displacement and the RL-stage contribution, enabling us to separate their effects.

Conventional interpolation (Chen et al., 2025) is recovered by setting $Z = \Delta$ and $\alpha = \beta = \lambda$:

$$W_j^{\text{full}} = V_j + \lambda\Gamma_j + \lambda\Delta_j = (1-\lambda)V_j + \lambda R_j, \qquad j \in \mathcal{S}. \tag{2}$$

Equation (2) makes the coupling explicit: standard interpolation simultaneously moves the recipient toward the donor's pre-RL initialization through Γ and transfers the entire RL-stage update through Δ. Isolating Δ identifies what was acquired during RL; the remaining question is whether all of this update should be transferred.

Figure 1 summarizes the stage-isolation, direction-selection, normalization, and composition steps.

3.2 FIXED MATRIXWISE SPECTRAL SELECTION

Having isolated the RL-stage update, we next ask **which part of it transfers effectively**. The complete update Δ_j was optimized for the source language model, so its constituent directions need not be equally compatible with a multimodal recipient. We therefore decompose each update matrix independently and study its spectral directions under controlled magnitude.

For a nonzero Δ_j with rank r_j, the compact singular value decomposition is $\Delta_j = U_j\Sigma_j Q_j^\top$, where $U_j = [u_{j,1}, \ldots, u_{j,r_j}]$ and $Q_j = [v_{j,1}, \ldots, v_{j,r_j}]$ contain the left and right singular vectors, respectively, and $\sigma_{j,1} \geq \cdots \geq \sigma_{j,r_j} > 0$, with $U_j^\top U_j = Q_j^\top Q_j = I_{r_j}$. For a positive integer k, we retain the leading k singular components, $\mathcal{P}_k(\Delta_j) := \Delta_{j,k} = \sum_{i=1}^{\min(k,r_j)} \sigma_{j,i}u_{j,i}v_{j,i}^\top$, with $\mathcal{P}_k(0) = 0$ and empty sums equal to zero. For nonzero updates, these approximations satisfy

$$\|\Delta_j - \Delta_{j,k}\|_F^2 = \min_{\text{rank}(A)\leq k} \|\Delta_j - A\|_F^2. \tag{3}$$

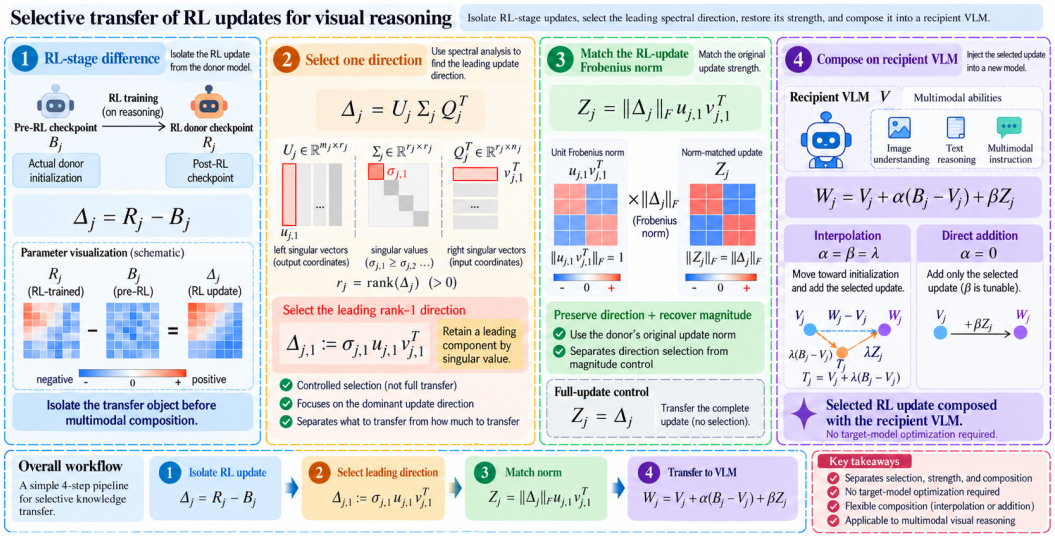


Figure 1: **Selective RL-update transfer.** Four stages isolate $R - B$, select a matrixwise direction, restore its Frobenius norm, and compose it with the recipient. Interpolation includes initialization displacement; direct addition omits it. Matrix patterns are schematic.

Equation (3) characterizes optimal low-rank reconstruction of the source update, which need not maximize transfer to another model. Spectral ordering defines controlled candidate directions whose transferability we evaluate empirically.

A second issue is that truncation changes both direction and magnitude. To avoid confounding these effects, we measure the fraction of squared Frobenius energy retained by the rank-k approximation, $\rho_{j,k} := \|\Delta_{j,k}\|_F^2 / \|\Delta_j\|_F^2$ for $\Delta_j \neq 0$, and rescale the truncated update to match the Frobenius norm of the original matrix (Cai et al., 2026): $\widehat{\Delta}_{j,k} := \mathcal{N}_k(\Delta_j)$, $\mathcal{N}_k(\Delta_j) = \rho_{j,k}^{-1/2} \Delta_{j,k}$ for $\Delta_j \neq 0$, and $\mathcal{N}_k(0) = 0$. Thus, $\|\widehat{\Delta}_{j,k}\|_F = \|\Delta_j\|_F$. The full and selected updates therefore have the same matrix-wise Frobenius magnitude while differing in directional content. Unless otherwise specified, Selective-RL uses $k = 1$ independently for each matrix, giving, for $\Delta_j \neq 0$, $\widehat{\Delta}_{j,1} = \|\Delta_j\|_F u_{j,1} v_{j,1}^\top$.

Rank one is a fixed empirical rule, not an assumption that one direction constitutes reasoning capability. We vary rank, direction, and normalization separately to assess their effects on transfer.

3.3 Controlled Transfer and Composition

Spectral selection determines what within the RL update is transferred. We next separate this choice from **how the update is composed with the recipient**, since conventional interpolation also introduces the initialization displacement Γ.

We first consider reconstructed-donor interpolation, which keeps the initialization displacement identical between full and selected transfer. Setting $Z = \widehat{\Delta}_1$ and $\alpha = \beta = \lambda$ gives $W_j^{\mathrm{rank1}} = V_j + \lambda(\Gamma_j + \widehat{\Delta}_{j,1})$, and $W_j^{\mathrm{rank1}} - W_j^{\mathrm{full}} = \lambda(\widehat{\Delta}_{j,1} - \Delta_j)$.

Both variants share initialization displacement $\lambda\Gamma_j$ and differ only in the RL contribution (Δ_j or $\widehat{\Delta}_{j,1}$), providing a controlled comparison of spectral selection.

We additionally consider direct RL-update transfer by setting $\alpha = 0$: $W_j^{\mathrm{direct}} = V_j + \beta\widehat{\Delta}_{j,1}$, $j \in \mathcal{S}$. This isolates RL-update transfer to test whether selection remains beneficial without initialization displacement.

For both routes, we compare $Z = \widehat{\Delta}_1$ against the corresponding full-update control $Z = \Delta$, holding transfer support, numerical precision, and coefficients fixed. Under direct transfer, $Z = \Delta$ reduces to RL-update Task Arithmetic (Ilharco et al., 2023). An initialization-only control instead sets $\beta = 0$

while retaining $\alpha\Gamma$. Together, these routes separate three otherwise entangled factors: **stage isolation, spectral selection, and update composition**. Construction requires no recipient gradient training; the coefficient is selected on H1 held-out validation and frozen for H2/H3. Algorithm 1 in Appendix A materializes the update.

3.4 Update Scale and Resource Use

Frobenius matching controls the magnitude of the **RL component**, but not necessarily the total displacement from the recipient. Under interpolation, the selected update also interacts with Γ. Let $\langle A, H\rangle_F := \mathrm{tr}(A^\top H)$. From Eq. (1),

$$\|W_j(\alpha, \beta; Z) - V_j\|_F^2 = \alpha^2\|\Gamma_j\|_F^2 + \beta^2\|Z_j\|_F^2 + 2\alpha\beta\langle\Gamma_j, Z_j\rangle_F. \tag{4}$$

Matching $\|Z_j\|_F$ therefore controls matrix-wise intervention strength while preserving directional differences through the cross term.

For rank-one selection with $\Delta_j \neq 0$, $\|\widehat{\Delta}_{j,1}\|_2/\|\Delta_j\|_2 = \rho_{j,1}^{-1/2} \geq 1$. Thus, Frobenius matching amplifies the retained direction by an amount determined by the spectral concentration of the original update. This motivates separate controls for rank, direction, global coefficient, and matrix-wise rescaling rather than attributing performance changes to selection alone.

All decomposition, normalization, coefficient selection, and checkpoint construction are performed offline. The checkpoint remains a standard dense VLM without extra inference-time modules.

4 Experiments

We evaluate **RQ1**, method and backbone transfer; **RQ2**, direction, scale, and composition; **RQ3**, source reconstruction and transfer boundaries; and **RQ4**, output behavior and construction cost.

4.1 Experimental Setup

Models and selection. H1 uses **Qwen3-VL-8B-Instruct**, H2 **LLaVA-NeXT-LLaMA3-8B**, and H3 **Idefics2-8B**, with Qwen3, LLaMA3, and Mistral language backbones (Bai et al., 2025; LMMs-Lab, 2024; Laurençon et al., 2024). Matched DAPO-style donors are trained on DAPO-Math-17K (Yu et al., 2025), starting from the actual pre-RL initialization. With primary seed 42, a 150-step budget, and checkpoints every ten steps, source-side validation selects steps **120/100/110** for H1/H2/H3; VLM benchmarks are not used for donor selection. H1 held-out validation selects $\lambda = 0.05$, which is frozen for H2/H3. The support contains $252/224/224$ attention/MLP matrices, respectively, and excludes visual components, embeddings, output heads, LayerNorm, and biases. All groups use bf16 and primary rank 1 (Appendix A).

Tasks and comparisons. The primary tasks are MathVista (N = 1,000), MathVerse (1,000), MathVision (304), DynaMath (5,010), and MMStar (1,500) (Lu et al., 2024; Zhang et al., 2025; Wang et al., 2024b; Zou et al., 2025; Chen et al., 2024). Additional H1 evaluations cover TextVQA, MMMU, OCRBench, and POPE (Singh et al., 2019; Yue et al., 2024; Liu et al., 2023; Li et al., 2023). We compare **Task Arithmetic**, **Full interpolation**, **TIES**, **DARE**, **TSV**, **LoRE**, **IP-Merging**, **FRISM**, **STAR**, and **AdaRank** (Ilharco et al., 2023; Chen et al., 2025; Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025; Liu et al., 2025a; Hu et al., 2025; Huang et al., 2026a; Lee et al., 2025; 2026). Full shares the proposed method's support. FRISM and AdaRank additionally adapt on the target side. We report single-run scores, donor seeds, and random-direction trials separately.

4.2 Comparing Methods and Backbones (RQ1)

Observation 1: Gains transfer across backbones, with task-dependent exceptions. Selective-RL improves Full in **12/15** group–task comparisons. H1 gains are $\uparrow 4.30/\uparrow 6.70/\uparrow 8.55/\uparrow 3.99/\uparrow 3.73$ points from paired counts. On H1 MathVista, $81.00 > 79.70 > 79.10 > 76.70$ for Selective, AdaRank, STAR, and Full. On MMStar, AdaRank leads ($60.07 > 59.67$). H2 improves four tasks but loses 0.20 MMStar points; H3 improves MathVista/MathVerse/DynaMath by $1.40/1.80/1.60$, while MathVision/MMStar decline $0.33/0.07$. These reversals are included in Table 1 and Figure 4.

Table 1: **Cross-backbone transfer and paired H1 outcomes.** Scores (%) for methods evaluated on all five benchmarks; extended comparisons: Table 6. Bold/underline: best/second within displayed groups; arrows: changes from Receiver.

Method	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
H1: Qwen3-VL-8B-Instruct					
Receiver	75.00	42.00	37.50	65.03	54.80
Task Arithmetic	76.10$_{\uparrow 1.10}$	44.50$_{\uparrow 2.50}$	39.47$_{\uparrow 1.97}$	65.67$_{\uparrow 0.64}$	54.67$_{\downarrow 0.13}$
Full interpolation	76.70$_{\uparrow 1.70}$	45.30$_{\uparrow 3.30}$	40.46$_{\uparrow 2.96}$	66.07$_{\uparrow 1.04}$	55.93$_{\uparrow 1.13}$
TIES	76.40$_{\uparrow 1.40}$	44.80$_{\uparrow 2.80}$	39.80$_{\uparrow 2.30}$	65.77$_{\uparrow 0.74}$	55.53$_{\uparrow 0.73}$
DARE	76.30$_{\uparrow 1.30}$	45.00$_{\uparrow 3.00}$	40.13$_{\uparrow 2.63}$	65.97$_{\uparrow 0.94}$	55.73$_{\uparrow 0.93}$
IP-Merging	77.40$_{\uparrow 2.40}$	46.60$_{\uparrow 4.60}$	42.76$_{\uparrow 5.26}$	66.87$_{\uparrow 1.84}$	56.67$_{\uparrow 1.87}$
FRISM†	78.30$_{\uparrow 3.30}$	49.20$_{\uparrow 7.20}$	45.07$_{\uparrow 7.57}$	68.06$_{\uparrow 3.03}$	57.73$_{\uparrow 2.93}$
STAR	79.10$_{\uparrow 4.10}$	49.80$_{\uparrow 7.80}$	47.70$_{\uparrow 10.20}$	68.62$_{\uparrow 3.59}$	58.27$_{\uparrow 3.47}$
AdaRank†	<u>79.70</u>$_{\uparrow 4.70}$	<u>50.60</u>$_{\uparrow 8.60}$	<u>48.36</u>$_{\uparrow 10.86}$	<u>69.12</u>$_{\uparrow 4.09}$	**60.07**$_{\uparrow 5.27}$
SELECTIVE-RL	**81.00**$_{\uparrow 6.00}$	**52.00**$_{\uparrow 10.00}$	**49.01**$_{\uparrow 11.51}$	**70.06**$_{\uparrow 5.03}$	<u>59.67</u>$_{\uparrow 4.87}$
H2: LLaVA-NeXT-LLaMA3-8B					
Receiver	37.40	20.10	13.82	22.69	43.80
Full interpolation	<u>39.10</u>$_{\uparrow 1.70}$	<u>22.40</u>$_{\uparrow 2.30}$	<u>15.13</u>$_{\uparrow 1.31}$	<u>24.07</u>$_{\uparrow 1.38}$	**44.33**$_{\uparrow 0.53}$
SELECTIVE-RL	**42.20**$_{\uparrow 4.80}$	**25.80**$_{\uparrow 5.70}$	**18.09**$_{\uparrow 4.27}$	**26.45**$_{\uparrow 3.76}$	<u>44.13</u>$_{\uparrow 0.33}$
H3: Idefics2-8B					
Receiver	51.80	19.40	17.11	21.80	**49.53**
Full interpolation	<u>52.60</u>$_{\uparrow 0.80}$	<u>20.40</u>$_{\uparrow 1.00}$	**17.76**$_{\uparrow 0.65}$	<u>22.55</u>$_{\uparrow 0.75}$	<u>49.47</u>$_{\downarrow 0.06}$
SELECTIVE-RL	**54.00**$_{\uparrow 2.20}$	**22.20**$_{\uparrow 2.80}$	<u>17.43</u>$_{\uparrow 0.32}$	**24.15**$_{\uparrow 2.35}$	49.40$_{\downarrow 0.13}$
H1: paired Full–Selective outcomes					
Task	N	Full correct	Selective correct	Gain (pp)	95% CI
MathVista	1,000	767	810	↑4.30	[2.14, 6.46]
MathVerse	1,000	453	520	↑6.70	[4.31, 9.09]
MathVision	304	123	149	↑8.55	[3.82, 13.29]
DynaMath	5,010	3,310	3,510	↑3.99	[2.10, 5.86]
MMStar	1,500	839	895	↑3.73	[2.23, 5.23]

†: target-side adaptation. H1/H2/H3 donor steps: 120/100/110; $\lambda = 0.05$. Paired gains use integer counts; DynaMath intervals resample seed questions (Appendix E).

Observation 2: Broader capability gains are substantial but not uniform. H1 gains over Full are $2.58/3.60/0.70/0.50$ points on TextVQA/MMMU/OCRBench/POPE. Relative to Receiver, OCRBench changes $78.60 \rightarrow 78.70$ and POPE $86.90 \rightarrow 86.80$. The broader scores appear in Table 28.

4.3 Isolating Selection and Composition (RQ2)

Observation 3: Selection helps under both composition routes. At $\beta = 0.05$, direct/interpolated selection gains $3.10/4.30$ MathVista points and $5.70/6.70$ MathVerse points over matched Full (Table 2). Initialization displacement lowers MathVista to 74.70 alone but raises selected direct transfer $79.20 \rightarrow 81.00$, depending on the paired update. Increasing Full's coefficient does not close the gap: at $\lambda = 0.2$, Rank-1 $78.90 >$ Receiver $75.00 >$ Full 74.40 (Figure 2).

Observation 4: Direction selection adds value beyond a tuned global or matrixwise scale. Unscaled rank-1 with validation-tuned $\beta = 0.50$ reaches MathVista 78.50; applying rank-one-derived matrixwise gains to Full reaches 77.60, versus matched selection 81.00. Ten matched random directions yield 74.05 ± 0.45, below Receiver 75.00; the trailing direction reaches 73.10. On DynaMath, tuned unscaled/full-reweighted/matched-leading give $67.52/66.93/70.06$. These controls separate direction from global and matrixwise scale; semantic attribution requires separate analysis.

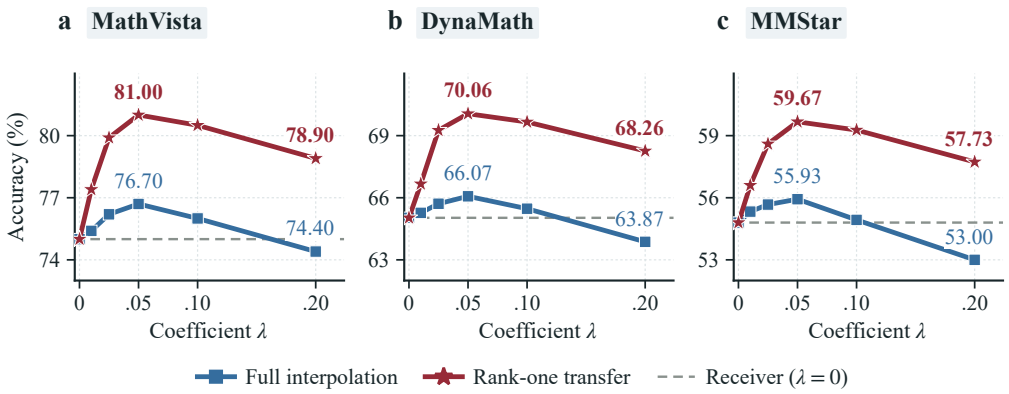


Figure 2: **Coefficient sensitivity on H1.** Full and rank-one interpolation for $0 \leq \lambda \leq 0.2$; dashed lines mark Receiver accuracy. Validation selects 0.05; full sweeps: Table 9.

Table 2: **Composition, direction controls, and generation behavior on H1.** Panels (a,b): four tasks with complete controls; five-task results: Table 7. Panel (c): 1,000 MathVista examples.

Construction	α	β	Math Vista	Math Verse	DynaMath	MMStar
(a) Composition: initialization displacement and RL update						
Receiver	0.00	0.00	75.00	42.00	65.03	54.80
Initialization only	0.05	0.00	74.70	41.50	64.65	54.07
Direct full	0.00	0.05	76.10	44.50	65.67	54.67
Direct rank-1	0.00	0.05	79.20	50.20	69.06	58.80
Interpolation full	0.05	0.05	76.70	45.30	66.07	55.93
Interpolation rank-1	0.05	0.05	81.00	52.00	70.06	59.67
(b) Direction and scale: fixed initialization displacement						
Leading, unscaled	0.05	0.05	75.20	42.60	65.27	54.80
Leading, tuned β	0.05	0.50	78.50	47.90	67.52	56.87
Full, matrixwise gains	0.05	0.05	77.60	46.50	66.93	56.33
Random, seed 42	0.05	0.05	73.70	40.50	63.67	52.93
Random, 10 seeds	0.05	0.05	74.05 ± 0.45	40.92 ± 0.51	64.11 ± 0.35	53.07 ± 0.33
Trailing, matched	0.05	0.05	73.10	39.60	62.87	52.33
Leading rank-1 matched	0.05	0.05	81.00	52.00	70.06	59.67
(c) H1 MathVista: output lengths and generation events						
Method			Mean tokens	P95 tokens	Repetition	Extract fail	At cap
Receiver			460	1,550	8	5	6
Full			575	2,200	15	8	14
SELECTIVE-RL			430	1,350	3	2	3

Random, 10 seeds: mean $\pm$ sample SD. Full matrixwise gains use $\rho_{j,1}^{-1/2}\Delta_j$; leading matched uses $\widehat{\Delta}_{j,1}$. Unscaled β is validation-selected. Generation events can overlap.

H1 median ρ_1 is 0.18 (IQR 0.11–0.31), with median amplification 2.36×; H2/H3 medians are 0.13/0.09. Attention retains more leading-component energy than MLP ($0.24 > 0.15$ in H1). Rank 1 gives MathVista 81.00, followed by 80.70/80.30/79.50/78.30 at ranks 2/4/8/16 (Figure 5). The rank-1/rank-2 gap is 0.30; fixed rank one is an empirical choice.

4.4 Source Reconstruction and Transfer Boundaries (RQ3)

Observation 5: Source recovery and recipient gains are distinct measurements. H1 rank-one reconstruction attains MATH 75.80 versus donor 76.20, recovering 97.18% of its gain over 62.00 pre-

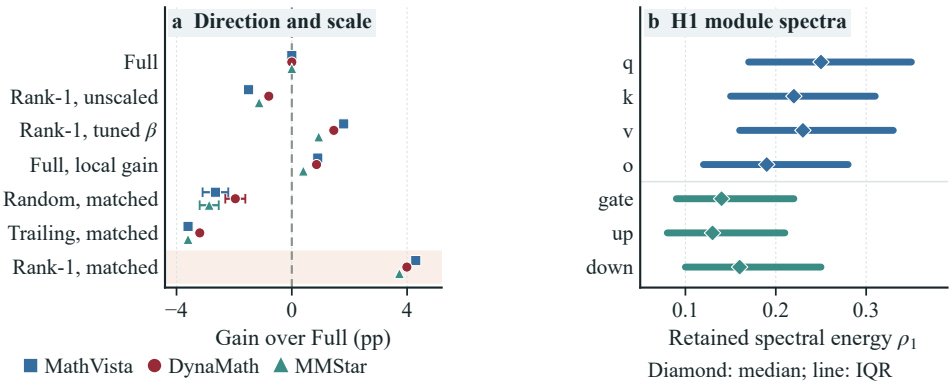


Figure 3: **Direction, scale, and spectral concentration.** Left: gains over Full under the H1 controls; only random-direction means carry sample-SD bars from ten runs. Right: module-wise median retained energy ρ_1 and interquartile ranges across eligible matrices.

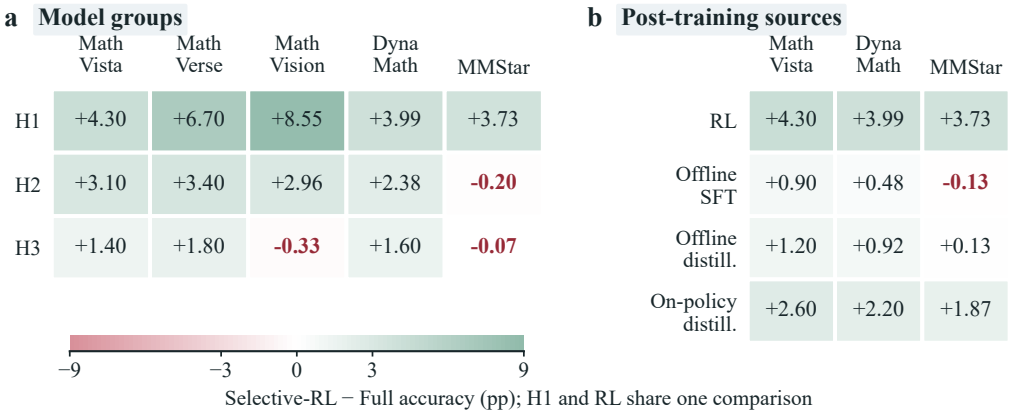


Figure 4: **Transfer scope and its limits.** Rank-1 minus Full in percentage points: three receiver backbones (left) and four post-training sources on H1 (right). Negative cells retain regressions; the H1/RL setting is shared across panels.

RL (Table 19). The earlier-reference control lowers selective MathVista $81.00 \rightarrow 77.80$, supporting explicit stage isolation. Three donor seeds give MathVista gains $4.10/4.30/4.00$ (4.13 ± 0.15). Offline SFT/distillation/on-policy distillation gain $0.90/1.20/2.60$ MathVista points; SFT loses 0.13 MMStar points. The effect is not exclusive to RL.

Primary DynaMath gains 200 correct variants (3.99 points), with a reported seed-cluster 95% interval $[2.10, 5.86]$; all-ten-correct rises $29.1 \rightarrow 34.3\%$ (Appendix E).

4.5 Generation Behavior and Cost (RQ4)

Observation 6: Shorter outputs accompany gains beyond obvious generation failures. Mean output length falls $575 \rightarrow 430$ tokens ($\downarrow 25.22\%$); repetition/extraction-failure/cap-hit counts fall $15/8/14 \rightarrow 3/2/3$. Of 83 Selective-only MathVista successes, $8/6/4$ correspond to Full repetition/truncation/extraction pathology and 65 (78.3%) to ordinary reasoning or perception errors. Mechanism attribution requires separate analysis. In the 200-example evaluator audit per model, disagreement is $3.0/3.5/2.5\%$ for Receiver/Full/Selective.

Construction costs $90 = 55 + 35$ s for decomposition and writing, versus Full's 35 s, STAR's 395 s, and AdaRank's 2,045 s. Validation search is separate (3.0 GPU-h for Full and Selective). The dense parameter ratio remains $1.00\times$; no inference module is added (Appendix F).

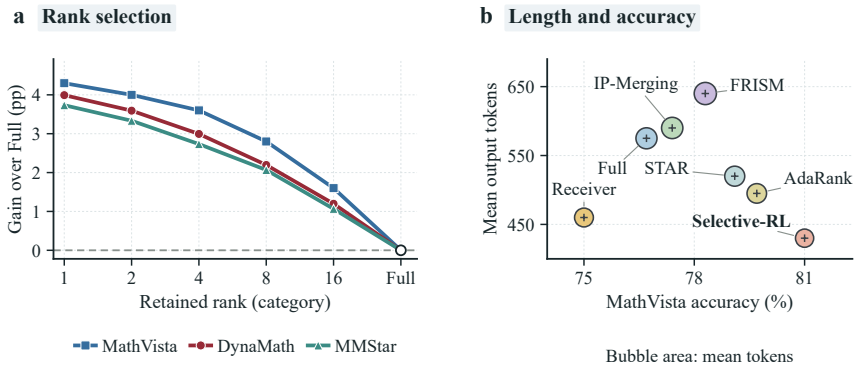


Figure 5: **Rank and output efficiency.** Left: gains over Full for finite ranks on three H1 tasks. Right: MathVista accuracy versus mean output tokens; bubble area encodes mean tokens. Points show measurements without fitted trends.

4.5.1 Qualitative Cases and Evaluation Boundaries

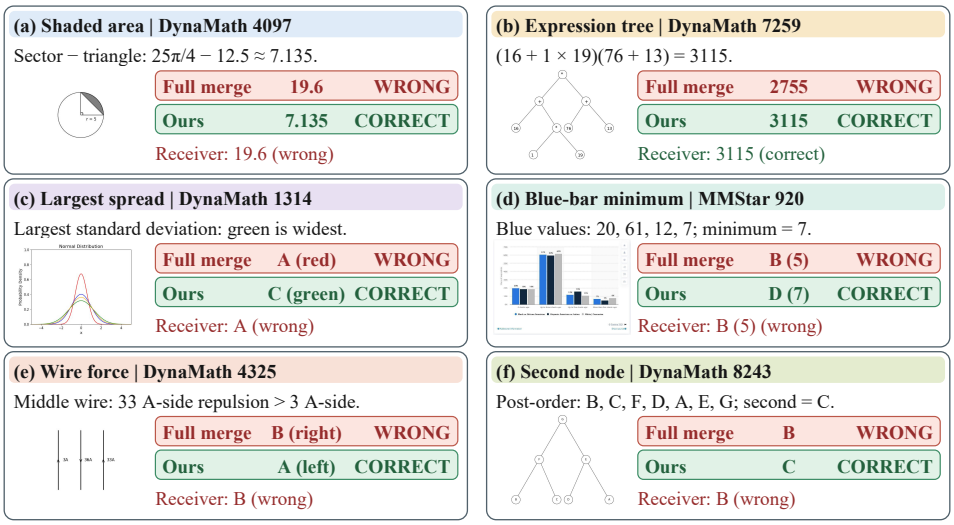


Figure 6: **Six Full errors corrected by Selective-RL.** Receiver also fails except in (b). Cases are outcome-selected; (f) uses left-to-right post-order. Records were retained during qualitative analysis.

Observation 7: Selection can correct a shared error or avoid a merge regression. Figure 6 shows six corrected Full errors: five shared failures and a merge regression in (b), where Receiver already answers 3115. Attribution to a singular component's visual skill requires separate analysis.

These outcome-selected Qwen-family cases serve qualitative analysis; benchmark performance is evaluated separately. Appendix G reports case provenance and further response comparisons, complementing the quantitative audit.

5 Conclusion

In this paper, we propose Selective-RL, a selective transfer algorithm that adapts RL-stage updates for visual reasoning. The method isolates the donor's RL update, selects each eligible matrix's leading spectral direction, and restores its magnitude through Frobenius matching before composing it with a recipient VLM. This construction separates direction selection, update strength, and initialization displacement without recipient gradient training or additional inference modules. Experiments across visual reasoning benchmarks and VLM backbones show effective selective transfer; controlled ablations characterize the roles of direction, normalization, and parameter support.

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance with technical tasks. We have not used generative AI tools for other applicable tasks with required disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability. We have reviewed all AI-assisted work. All AI-generated suggestions were independently reviewed, revised, and validated by the authors before inclusion in the final work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing model checkpoints and standard research benchmarks and does not involve the collection of private data or human-subject experiments. We follow the applicable licenses and usage conditions of the models, datasets, and software used in this study.

REPRODUCIBILITY STATEMENT

We provide the methodological details, experimental settings, evaluation protocols, and hyperparameters needed to understand and reproduce the reported experiments in the main paper and appendices. An anonymous code repository with the corresponding implementation and supporting materials is also provided.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631v2*, 2025. URL https://arxiv.org/abs/2511.21631v2.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Benyi Zhang, Zhenyu Tan, Guiquan Liu, and Junfeng Fang. On Predictability of Reinforcement Learning Dynamics for Large Language Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4c454d34f3a4c8d6b4ca85a918e5d7ba-Abstract-Conference.html.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring Reason to Vision: Understanding Perception and Reasoning through Model Merging. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 9803–9817. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25cm.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Nico Daheim, Thomas Möllenhoff, Edoardo M. Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model Merging by Uncertainty-Based Gradient Matching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/327b9b8d4e45c3f81568e11ffc505f77-Abstract-Conference.html.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex Vision-Language Reasoning via Iterative SFT-RL Cycles. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b33286edd8602ead8cb966358b01d115-Abstract-Conference.html.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024. URL https://arxiv.org/abs/2407.11691v1.

Zitao Fang, Guodong Du, Shuyang Yu, Yifei Guo, Yiwei Zhang, Yiyao Cao, Jing Li, Ho-Kin Tang, and Sim Kuan Goh. To See a World in a Spark of Neuron: Disentangling Multi-Task Interference for Training-Free Model Merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15720–15740. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.793. URL https://aclanthology.org/2025.emnlp-main.793/.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task Singular Vectors: Reducing Task Interference in Model Merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18695–18705, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Yijie Hu, Zihao Zhou, Kaizhu Huang, Xiaowei Huang, and Qiufeng Wang. Can MLLMs Absorb Math Reasoning Abilities from LLMs as Free Lunch? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bdcdf38389d7fcefc73c4c3720217155-Abstract-Conference.html.

Chenyu Huang, Peng Ye, Xudong Tan, Jinhan Mu, Shenghe Zheng, Li Shen, and Tao Chen. FRISM: Fine-Grained Reasoning Injection via Subspace-Level Model Merging for Vision-Language Models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=F13K6NNeAz.

Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung-yi Lee. Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://aclanthology.org/2024.acl-long.590/.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=UZIjskfbfU.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089v3.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless Knowledge Fusion by Merging Weights of Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FCnohuR6AnM.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.02246.

Chanhyuk Lee, Jiho Choi, Chanryeol Lee, Donggyun Kim, and Seunghoon Hong. AdaRank: Adaptive Rank Pruning for Enhanced Model Merging. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0b77a07c3c12b52121b3f42f96875726-Abstract-Conference.html.

Yu-Ang Lee, Ching-Yun Ko, Tejaswini Pedapati, I-Hsin Chung, Mi-Yen Yeh, and Pin-Yu Chen. STAR: Spectral Truncation and Rescale for Model Merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 496–505. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.42. URL https://aclanthology.org/2025.naacl-short.42/.

Xirui Li, Ming Li, and Tianyi Zhou. What does RL improve for Visual Reasoning? A Frankenstein-Style Analysis. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2602.12395. Accepted paper; full text available as arXiv:2602.12395.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.20/.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. URL https://arxiv.org/abs/2305.07895.

Zehua Liu, Han Wu, Yuxuan Yao, Xiaojin Fu, Ruifeng She, Xiongwei Han, Tao Zhong, and Mingxuan Yuan. LoRE-Merging: Exploring Low-Rank Estimation For Large Language Model Merging. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21919–21926. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-emnlp.1195. URL https://aclanthology.org/2025.findings-emnlp.1195/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. In *Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=5PAF7PAY2Y.

LMMs-Lab. LLaVA-NeXT Llama-3 8B Model Card, 2024. URL https://huggingface.co/lmms-lab/llama3-llava-next-8b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02255v3.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No Task Left Behind: Isotropic Model Merging with Common and Task-Specific Subspaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 43177–43199. PMLR, 2025. URL https://proceedings.mlr.press/v267/marczak25a.html.

Michael S. Matena and Colin A. Raffel. Merging Models with Fisher-Weighted Averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d28077e5ff52034cd35b4aa15320caea-Abstract.html.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e12a3b98b67e8395f639fde4c2b03168-Abstract-Conference.html.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1904.08920.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c84844a559e4f962752570bff456ae4-Abstract-Conference.html.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing Task Information for Improved Model Merging and Compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 50268–50287. PMLR, 2024a. URL https://proceedings.mlr.press/v235/wang24k.html.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.01708v2.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 57755–57775, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16502.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/537d5aa768c2d534016a4d06f87bc8fb-Abstract-Conference.html.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *Computer Vision – ECCV 2024*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73242-3_10. URL https://link.springer.com/chapter/10.1007/978-3-031-73242-3_10.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model Tailor: Mitigating Catastrophic Forgetting in Multi-modal Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62581–62598. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhu24l.html.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

A Model Configurations and Evaluation Protocol

Algorithm 1 Fixed Matrixwise Rank-One Transfer

Input: recipient V, actual pre-RL checkpoint B, donor R, matrix support $\mathcal{S}$, coefficients α, β.
Verify aligned tensor coordinates and consistent support and updates for tied parameters.
$W \leftarrow \mathrm{copy}(V)$. `// copy all tensors`
for each $j \in \mathcal{S}$ **do**
  $\Gamma_j \leftarrow B_j - V_j$; $\Delta_j \leftarrow R_j - B_j$. `// isolate RL`
  $\eta_j \leftarrow \sqrt{\langle \Delta_j, \Delta_j \rangle_F} = \|\Delta_j\|_F$.
  if $\eta_j = 0$ **then** $Z_j \leftarrow 0$ `// zero RL term`
  else $(\sigma_{j,1}, u_{j,1}, v_{j,1}) \leftarrow \mathrm{svd}_1(\Delta_j)$
    $Z_j \leftarrow \eta_j u_{j,1} v_{j,1}^\top = \mathcal{N}_1(\Delta_j)$. `// match update norm`
  end if
  $W_j \leftarrow V_j + \alpha\Gamma_j + \beta Z_j$. `// compose both updates`
end for
Return W; $\forall j \notin \mathcal{S},\ W_j = V_j$.

Routes: interpolation $\alpha = \beta = \lambda$; direct addition $\alpha = 0$ with coefficient β.

Table 3: **Model groups and donor selection.** All donors use matched DAPO-style training on DAPO-Math-17K.

Configuration	H1: primary	H2: cross-family	H3: cross-backbone
Receiver	Qwen3-VL-8B-Instruct	LLaVA-NeXT-LLaMA3-8B	Idefics2-8B
LM family	Qwen3	LLaMA3	Mistral
Layers / matrices	36 / 252	32 / 224	32 / 224
Donor seed	42	42	42
Selected step	**120**	**100**	**110**
Training budget	150 steps	150 steps	150 steps
Save interval	10 steps	10 steps	10 steps
Donor selection	Source-side validation only		
Coefficient	0.05: H1 held-out selection, then frozen for H2/H3		

The actual initialization used by each donor defines B. All groups use bf16 and rank 1; visual components, embeddings, output heads, LayerNorm, and bias remain frozen.

Checkpoint identities. The receiver checkpoints and donor configurations used in the primary experiments are as follows:

- **H1:** receiver: Qwen/Qwen3-VL-8B-Instruct; pre-RL initialization B: Qwen3-8B-Base (Yang et al., 2025); donor run: qwen3_dapo_math17k_s42_step120.
- **H2:** receiver: lmms-lab/llama3-llava-next-8b; pre-RL initialization B: Meta-Llama-3-8B-Instruct; donor run: llama3_dapo_math17k_s42_step100.
- **H3:** receiver: HuggingFaceM4/idefics2-8b; pre-RL initialization B: Mistral-7B-v0.1; donor run: mistral7b_dapo_math17k_s42_step110.

Each donor is obtained from the corresponding pre-RL initialization through DAPO-style post-training. Checkpoints are saved every ten training steps within a 150-step budget, and the checkpoint used for each primary configuration is selected using source-side validation. The donor run identifiers denote the training runs used in the reported experiments.

Parameter support. The seven projections are `q/k/v/o_proj` and `gate/up/down_proj`. There are 36 language layers in H1 and 32 in H2/H3, giving 252/224/224 eligible matrices. Correspondence is by layer, module, and orientation, not a shared full parameter-name string. The recipient configuration and tensors outside this support remain fixed. Visual modules, embeddings, output heads, LayerNorm parameters, and biases are excluded, also avoiding transfer of incompatible multimodal vocabulary extensions. The anonymous repository includes inference and evaluation support based on VLMEvalKit (Duan et al., 2024).

Table 4: **Evaluation tasks and sample sizes.** Main results use the first five tasks; the remaining four assess broader capabilities.

Benchmark	N	Metric
MathVista	1,000	Accuracy (%)
MathVerse	1,000	Accuracy (%)
MathVision	304	Accuracy (%)
DynaMath	5,010	Accuracy (%)
MMStar	1,500	Accuracy (%)
TextVQA	1,000	VQA soft score
MMMU	1,000	Accuracy (%)
OCRBench	1,000	Score (%)
POPE	3,000	Accuracy (%)

Table 5: **H1 held-out coefficient selection.** $\lambda = 0.05$ leads both validation metrics and is then frozen for H2 and H3.

λ	MathVista-val	DynaMath-val
0.01	76.90	66.10
0.025	79.00	68.35
0.05	80.20	69.40
0.075	79.90	69.10
0.1	79.50	68.75

B Extended Method and Control Comparisons

Table 6 retains every reported baseline and available five-task score. The main comparison presents rows with all five measurements available; this appendix preserves the broader method coverage. Table 7 reports the five-task controls, including MathVision, while the main control block uses four tasks with complete coverage. — indicates an unavailable measurement.

Table 6: **Extended comparison across all reported merging methods.** Accuracy (%). Colored subscripts show changes from the Receiver within each group; changes use displayed scores. Bold/underline: best/second-best available score per group and task.

Method	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
H1: Qwen3-VL-8B-Instruct					
Receiver	75.00	42.00	37.50	65.03	54.80
Task Arithmetic	76.10$_{\uparrow 1.10}$	44.50$_{\uparrow 2.50}$	39.47$_{\uparrow 1.97}$	65.67$_{\uparrow 0.64}$	54.67$_{\downarrow 0.13}$
Full interpolation	76.70$_{\uparrow 1.70}$	45.30$_{\uparrow 3.30}$	40.46$_{\uparrow 2.96}$	66.07$_{\uparrow 1.04}$	55.93$_{\uparrow 1.13}$
TIES	76.40$_{\uparrow 1.40}$	44.80$_{\uparrow 2.80}$	39.80$_{\uparrow 2.30}$	65.77$_{\uparrow 0.74}$	55.53$_{\uparrow 0.73}$
DARE	76.30$_{\uparrow 1.30}$	45.00$_{\uparrow 3.00}$	40.13$_{\uparrow 2.63}$	65.97$_{\uparrow 0.94}$	55.73$_{\uparrow 0.93}$
TSV	77.90$_{\uparrow 2.90}$	47.80$_{\uparrow 5.80}$	—	—	57.07$_{\uparrow 2.27}$
LoRE	78.40$_{\uparrow 3.40}$	48.70$_{\uparrow 6.70}$	—	67.90$_{\uparrow 2.87}$	57.47$_{\uparrow 2.67}$
IP-Merging	77.40$_{\uparrow 2.40}$	46.60$_{\uparrow 4.60}$	42.76$_{\uparrow 5.26}$	66.87$_{\uparrow 1.84}$	56.67$_{\uparrow 1.87}$
FRISM†	78.30$_{\uparrow 3.30}$	49.20$_{\uparrow 7.20}$	45.07$_{\uparrow 7.57}$	68.06$_{\uparrow 3.03}$	57.73$_{\uparrow 2.93}$
STAR	79.10$_{\uparrow 4.10}$	49.80$_{\uparrow 7.80}$	47.70$_{\uparrow 10.20}$	68.62$_{\uparrow 3.59}$	58.27$_{\uparrow 3.47}$
AdaRank†	<u>79.70</u>$_{\uparrow 4.70}$	<u>50.60</u>$_{\uparrow 8.60}$	<u>48.36</u>$_{\uparrow 10.86}$	<u>69.12</u>$_{\uparrow 4.09}$	**60.07**$_{\uparrow 5.27}$
SELECTIVE-RL	**81.00**$_{\uparrow 6.00}$	**52.00**$_{\uparrow 10.00}$	**49.01**$_{\uparrow 11.51}$	**70.06**$_{\uparrow 5.03}$	<u>59.67</u>$_{\uparrow 4.87}$
H2: LLaVA-NeXT-LLaMA3-8B					
Receiver	37.40	20.10	13.82	22.69	43.80
Full interpolation	39.10$_{\uparrow 1.70}$	22.40$_{\uparrow 2.30}$	<u>15.13</u>$_{\uparrow 1.31}$	<u>24.07</u>$_{\uparrow 1.38}$	44.33$_{\uparrow 0.53}$
STAR	40.60$_{\uparrow 3.20}$	24.20$_{\uparrow 4.10}$	—	—	<u>44.53</u>$_{\uparrow 0.73}$
AdaRank†	<u>41.00</u>$_{\uparrow 3.60}$	<u>24.80</u>$_{\uparrow 4.70}$	—	—	**44.93**$_{\uparrow 1.13}$
SELECTIVE-RL	**42.20**$_{\uparrow 4.80}$	**25.80**$_{\uparrow 5.70}$	**18.09**$_{\uparrow 4.27}$	**26.45**$_{\uparrow 3.76}$	44.13$_{\uparrow 0.33}$
H3: Idefics2-8B					
Receiver	51.80	19.40	17.11	21.80	49.53
Full interpolation	52.60$_{\uparrow 0.80}$	20.40$_{\uparrow 1.00}$	**17.76**$_{\uparrow 0.65}$	<u>22.55</u>$_{\uparrow 0.75}$	49.47$_{\downarrow 0.06}$
STAR	53.20$_{\uparrow 1.40}$	21.40$_{\uparrow 2.00}$	—	—	<u>49.60</u>$_{\uparrow 0.07}$
AdaRank†	<u>53.60</u>$_{\uparrow 1.80}$	<u>21.70</u>$_{\uparrow 2.30}$	—	—	**50.00**$_{\uparrow 0.47}$
SELECTIVE-RL	**54.00**$_{\uparrow 2.20}$	**22.20**$_{\uparrow 2.80}$	<u>17.43</u>$_{\uparrow 0.32}$	**24.15**$_{\uparrow 2.35}$	49.40$_{\downarrow 0.13}$

†: target-side adaptation. — indicates an unavailable measurement. H1/H2/H3 use donor checkpoints 120/100/110, respectively, with $\lambda = 0.05$.

Table 7: **Five-benchmark composition and direction controls.** The same five benchmarks are used across controls. The ten-seed random row reports mean ± sample standard deviation.

Construction	α	β	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
(a) Composition: initialization displacement and RL update							
Receiver	0.00	0.00	75.00	42.00	37.50	65.03	54.80
Initialization only	0.05	0.00	74.70	41.50	37.17	64.65	54.07
Direct full	0.00	0.05	76.10	44.50	39.47	65.67	54.67
Direct rank-1	0.00	0.05	79.20	50.20	47.04	69.06	58.80
Interpolation full	0.05	0.05	76.70	45.30	40.46	66.07	55.93
Interpolation rank-1	0.05	0.05	81.00	52.00	49.01	70.06	59.67
(b) Direction and scale: fixed initialization displacement							
Leading, unscaled	0.05	0.05	75.20	42.60	38.16	65.27	54.80
Leading, tuned β	0.05	0.50	78.50	47.90	—	67.52	56.87
Full, matrixwise gains	0.05	0.05	77.60	46.50	—	66.93	56.33
Random, seed 42	0.05	0.05	73.70	40.50	35.86	63.67	52.93
Random, 10 seeds	0.05	0.05	74.05 ± 0.45	40.92 ± 0.51	—	64.11 ± 0.35	53.07 ± 0.33
Trailing, matched	0.05	0.05	73.10	39.60	34.54	62.87	52.33
Leading rank-1 matched	0.05	0.05	81.00	52.00	49.01	70.06	59.67

Full matrixwise gains use $\rho_{j,1}^{-1/2}\Delta_j$; leading matched uses $\widehat{\Delta}_{j,1}$. The unscaled coefficient is chosen on validation. — indicates an unavailable measurement.

C Additional Mechanism and Spectral Analyses

Tables 8–10 expand the rank, coefficient, and support comparisons. Tables 11–13 report retained-energy distributions without fitting an association across configurations.

Table 8: **Rank sensitivity on H1.** Frobenius matching is applied at every finite rank.

Rank	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
1	81.00	52.00	49.01	70.06	59.67
2	80.70	51.60	—	69.66	59.27
4	80.30	50.80	47.70	69.06	58.67
8	79.50	49.40	—	68.26	58.00
16	78.30	47.60	—	67.27	57.00
Full	76.70	45.30	40.46	66.07	55.93

Full retains the complete update. — indicates an unavailable measurement.

Table 9: **Complete test-set coefficient sweep.** The main coefficient 0.05 is selected using H1 validation, not this test sweep.

λ	MathVista Full	MathVista Rank-1	DynaMath Full	DynaMath Rank-1	MMStar Full	MMStar Rank-1
0	75.00	75.00	65.03	65.03	54.80	54.80
0.01	75.40	77.40	65.27	66.67	55.33	56.60
0.025	76.20	79.90	65.71	69.26	55.67	58.60
0.05	76.70	81.00	66.07	70.06	55.93	59.67
0.1	76.00	80.50	65.47	69.66	54.93	59.27
0.2	74.40	78.90	63.87	68.26	53.00	57.73
0.5	70.00	76.50	59.88	66.07	49.00	55.67
1	61.00	75.20	52.89	65.07	42.33	55.00

Table 10: **Eligible support on H1.** The remaining recipient parameters stay fixed.

Support	Matrices	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
QKV only	108	78.10	47.60	—	—	57.47
Attention only	144	79.40	49.30	—	68.52	58.47
MLP only	108	78.80	48.60	—	68.88	57.93
O + Down only	72	77.90	46.90	—	67.54	57.20
Attention + MLP	252	81.00	52.00	49.01	70.06	59.67

QKV: query/key/value projections; O: attention output; Down: MLP down projection. — indicates an unavailable measurement.

Table 11: **Spectral concentration by model configuration.** ρ_1 is retained squared Frobenius energy; amplification is $\rho_1^{-1/2}$.

Configuration	Attention median	MLP median	Overall	Q1	Q3	Median amplification
H1 Qwen RL	0.24	0.15	0.18	0.11	0.31	2.36
H2 LLaMA RL	0.17	0.11	0.13	0.08	0.24	2.77
H3 Mistral RL	0.12	0.08	0.09	0.05	0.17	3.33

Table 12: **H1 spectral energy by module.** Quartiles summarize the distribution across eligible matrices.

Module	Mean ρ_1	Median ρ_1	Q1	Q3
q_proj	0.27	0.25	0.17	0.35
k_proj	0.24	0.22	0.15	0.31
v_proj	0.25	0.23	0.16	0.33
o_proj	0.21	0.19	0.12	0.28
gate_proj	0.16	0.14	0.09	0.22
up_proj	0.15	0.13	0.08	0.21
down_proj	0.18	0.16	0.10	0.25

Table 13: **H1 spectral energy by depth.**

Layer region	Median ρ_1	Median amplification
Early 1-12	0.13	2.77
Middle 13-24	0.22	2.13
Late 25-36	0.19	2.29

Table 14: **Normalization ablation on H1.** Leading rank-one updates are compared at the same composition coefficients.

Normalization	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
None	75.20	42.60	38.16	65.27	54.80
Nuclear-norm match	80.00	50.10	—	69.16	58.67
Frobenius match	81.00	52.00	49.01	70.06	59.67

Frobenius and nuclear matching use the corresponding norm of the full matrix as the target. — indicates an unavailable measurement.

Table 15: **Stage-reference ablation.** The actual donor initialization is compared with an earlier reference checkpoint.

Reference control	Math Vista	Math Verse	Math Vision	DynaMath	MMStar
Full, actual pre-RL	76.70	45.30	40.46	66.07	55.93
Rank-1, actual pre-RL	81.00	52.00	49.01	70.06	59.67
Full, earlier reference	75.60	43.80	—	65.41	54.93
Rank-1, earlier reference	77.80	46.90	—	67.11	56.40

All variants use $\alpha = \beta = 0.05$. — indicates an unavailable measurement.

Table 16: **Unscaled leading rank-one coefficient sweep.** Initialization displacement remains fixed at $\alpha = 0.05$.

β	MathVista	DynaMath	MMStar
0.05	75.20	65.27	54.80
0.1	76.40	65.87	55.47
0.2	77.60	66.91	56.33
0.5	78.50	67.52	56.87
1	77.80	66.95	56.40
2	74.90	63.95	53.87

Table 17: **Separating direction selection from matrixwise amplification.**

Control	MathVista	DynaMath	MMStar
Full, global beta	76.70	66.07	55.93
Full, rank-1 matrixwise gains	77.60	66.93	56.33
Rank-1, tuned global beta	78.50	67.52	56.87
Rank-1, Frobenius matched	81.00	70.06	59.67

Table 18: **Random-direction robustness.** Ten matched rank-one direction draws at $\alpha = \beta = 0.05$.

Direction seed	MathVista	DynaMath	MMStar
0	74.60	64.43	53.27
1	73.80	63.87	52.87
2	74.10	64.07	53.13
3	73.40	63.55	52.47
4	74.70	64.51	53.60
5	73.90	63.91	52.93
6	74.30	64.23	53.20
7	73.50	63.67	52.73
8	74.40	64.35	53.40
9	73.80	64.49	53.07
Mean $\pm$ SD	74.05 ± 0.45	64.11 ± 0.35	53.07 ± 0.33

Summary statistics are recomputed from the displayed ten runs using sample SD.

D SOURCE RECOVERY, DONOR VARIATION, AND BOUNDARIES

Accuracy retention is A_{recon}/A_R and gain retention is $(A_{\text{recon}}-A_B)/(A_R-A_B)$, each expressed as a percentage. These describe recovery of donor behavior; recipient transfer is evaluated separately. Donor stages belong to one trajectory, while the seed comparison repeats donor training.

Table 19: **Source-side reconstruction.** MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) accuracy; retention ratios use the donor accuracy or its gain over the pre-RL model.

Reconstruction	MATH	GSM8K	Accuracy retention (%)	Gain retention (%)
Pre-RL	62.00	84.99	—	—
Donor	76.20	90.98	100.00	100.00
rank-1	75.80	90.74	99.48	97.18
rank-2	75.40	90.60	98.95	94.37
rank-4	76.00	90.82	99.74	98.59
rank-8	75.60	90.67	99.21	95.77

For MATH, rank-1 recovers $(75.80-62.00)/(76.20-62.00) = 97.18\%$ of the donor improvement. Retention is reported as a reconstruction-fidelity ratio.

Table 20: **H1 donor progression.** Stage labels refer to checkpoints within one RL training trajectory.

Stage	Source MATH	MathVista Full	MathVista Rank-1	DynaMath Full	DynaMath Rank-1	MMStar Full	MMStar Rank-1
Early	69.60	75.80	77.90	65.47	66.91	55.33	56.47
Middle	75.60	76.60	79.90	65.87	68.72	55.80	58.33
Late	76.20	76.70	81.00	66.07	70.06	55.93	59.67

Table 21: **Transfer from different post-training sources.** Within each source, Full and Rank-1 share the donor and support.

Source	Transfer	MathVista	DynaMath	MMStar	Median ρ_1
RL	Full	76.70	66.07	55.93	—
RL	Rank-1	81.00	70.06	59.67	0.18
Offline SFT	Full	77.10	66.67	56.27	—
Offline SFT	Rank-1	78.00	67.15	56.13	0.10
Offline distillation	Full	77.60	67.27	56.67	—
Offline distillation	Rank-1	78.80	68.18	56.80	0.12
On-policy distillation	Full	77.20	67.07	56.53	—
On-policy distillation	Rank-1	79.80	69.26	58.40	0.15

Table 22: **Donor-training seed robustness.** Each seed compares Full and Selective with the same trained donor.

Seed	Full MathVista	Selective MathVista	MathVista gain	DynaMath gain	MMStar gain
17	76.60	80.70	↑4.10	↑3.75	↑3.47
42	76.70	81.00	↑4.30	↑3.99	↑3.73
73	76.50	80.50	↑4.00	↑3.81	↑3.33
Mean ± SD	76.60 ± 0.10	80.73 ± 0.25	4.13 ± 0.15	3.85 ± 0.12	3.51 ± 0.20

Gains are in percentage points; standard deviations are sample SDs over the three donor seeds.

E Paired Outcomes and Variant Robustness

Let (a, b, c, d) denote both-correct, Full-only, Selective-only, and both-wrong counts. With $n = a + b + c + d$ and $\widehat{\delta} = (c - b)/n$, the paired normal interval for independent evaluation examples is

$$100 \left[\widehat{\delta} \pm 1.96 \sqrt{\frac{(b + c)/n - \widehat{\delta}^2}{n - 1}} \right]. \tag{5}$$

For DynaMath the resampling unit is the seed question. If $y^M_{s,v} \in \{0, 1\}$ is correctness for method M on variant v of seed s, then

$$d_s = \frac{1}{10} \sum_{v=1}^{10} (y^{\text{Selective}}_{s,v} - y^{\text{Full}}_{s,v}), \qquad \widehat{\delta} = \frac{1}{501} \sum_{s=1}^{501} d_s. \tag{6}$$

Seed-cluster bootstrap resamples the 501 questions with replacement, retaining the ten individual variant outcomes within each question. It does not impose equal outcomes within a question. The reported DynaMath interval is $[2.10, 5.86]$ percentage points. The paired counts yield 3,310 Full-correct and 3,510 Selective-correct variants; $113 + 313 = 426$ variants have different correctness outcomes.

Table 23: **Paired H1 outcomes and 95% confidence intervals.** Full is the baseline.

Task	a	b	c	d	N	Gain (pp)	95% CI
MathVista	727	40	83	150	1000	↑4.30	[2.14, 6.46]
MathVerse	410	43	110	437	1000	↑6.70	[4.31, 9.09]
MathVision	108	15	41	140	304	↑8.55	[3.82, 13.29]
DynaMath	3197	113	313	1387	5010	↑3.99	[2.10, 5.86]
MMStar	800	39	95	566	1500	↑3.73	[2.23, 5.23]

a: both correct; b: Full only; c: Selective only; d: both wrong. Gains use integer counts before rounding. DynaMath uses seed-cluster bootstrap; the other tasks use paired normal intervals.

Table 24: **DynaMath robustness across variants.** Each of the 501 seed questions has ten variants.

Statistic	Full	Selective-RL
All-ten-correct (%)	29.1	34.3
Median correct variants	7.0	8.0

All-ten-correct is a seed-level metric, distinct from mean accuracy across 5,010 variants. The ten outcomes within a seed are retained individually.

F Generation Diagnostics and Resource Accounting

Output diagnostics are measured on H1 MathVista. Repetition, extraction failure, and reaching the generation cap may overlap. The separate analysis of Selective-only wins uses the mutually exclusive categories in Table 26. The evaluator audit compares extracted labels with final-answer correctness on 200 manually inspected examples per model, separately from full benchmark scores.

Table 25: **Generation diagnostics on H1 MathVista.** Token counts and events over 1,000 examples.

Method	Mean tokens	P95 tokens	Repeat	Extract fail	At cap
Receiver	460	1550	8	5	6
Full	575	2200	15	8	14
IP-Merging	590	2050	11	6	10
FRISM	640	2200	7	4	8
STAR	520	1820	8	4	7
AdaRank	495	1740	6	4	6
SELECTIVE-RL	430	1350	3	2	3

Repetition, extraction failure, and reaching the token cap may overlap.

Table 26: **Error categories among 83 Selective-only MathVista successes.**

Full error category	Count	Share (%)
Repetition / loop	8	9.6
Token-cap truncation	6	7.2
Extraction pathology	4	4.8
Reasoning / perception	65	78.3

The categories are mutually exclusive; rounded shares may sum to 99.9%.

Table 27: **Final-answer/evaluator audit on H1 MathVista.** Each model has 200 manually audited examples.

Statistic	Receiver	Full	SELECTIVE-RL
Audited examples	200.0	200.0	200.0
Answer/extraction disagreement	6.0	7.0	5.0
False-positive extraction	2.0	3.0	2.0
False-negative extraction	4.0	4.0	3.0
Disagreement (%)	3.0	3.5	2.5

The audit is reported separately from the full-benchmark scores.

Table 28: **Broader H1 capabilities.** TextVQA uses the VQA soft score; other metrics follow Table 4.

Benchmark	Receiver	Full	SELECTIVE-RL	Δ vs Full
TextVQA	82.03	81.06	83.64	↑2.58
MMMU	60.00	59.20	62.80	↑3.60
OCRBench	78.60	78.00	78.70	↑0.70
POPE	86.90	86.30	86.80	↑0.50

OCRBench is reported on a 0–100 scale. POPE changes 86.90 → 86.80 relative to Receiver, so improvements are not uniform across perception metrics.

Table 29: **Offline construction costs.** Stage times and totals are in seconds; validation search is separate in GPU-hours.

Method	SVD	Other	Write	Calib.	Total	Search (GPU-h)
Receiver	0	0	0	0	0	0.0
Full	0	0	35	0	35	3.0
TIES	0	40	35	0	75	3.0
DARE	0	27	35	0	62	3.0
TSV	260	90	35	0	385	3.0
STAR	285	75	35	0	395	3.0
AdaRank	300	110	35	1600	2045	4.0
IP-Merging	310	175	35	0	520	3.0
FRISM	300	195	35	2400	2930	4.5
SELECTIVE-RL	55	0	35	0	90	3.0

Table 30: **Deployment footprint.** The merged model retains the recipient architecture.

Method	Dense ratio	Added module	VRAM (GB)
Receiver	1.00	No	24.60
Full	1.00	No	24.60
STAR	1.00	No	24.60
AdaRank	1.00	No	24.60
FRISM	1.00	No	24.60
SELECTIVE-RL	1.00	No	24.60

Peak VRAM is reported for batch size 1; completion time is a separate deployment metric.

Construction totals sum decomposition, other processing, writing, and calibration. Validation search is additional and reported in GPU-hours; donor training is a separate upstream cost. Parameter count and VRAM characterize the dense merged model. Timing and VRAM measurements use the same hardware/software environment. Output length measures token use; inference latency is a separate timing metric.

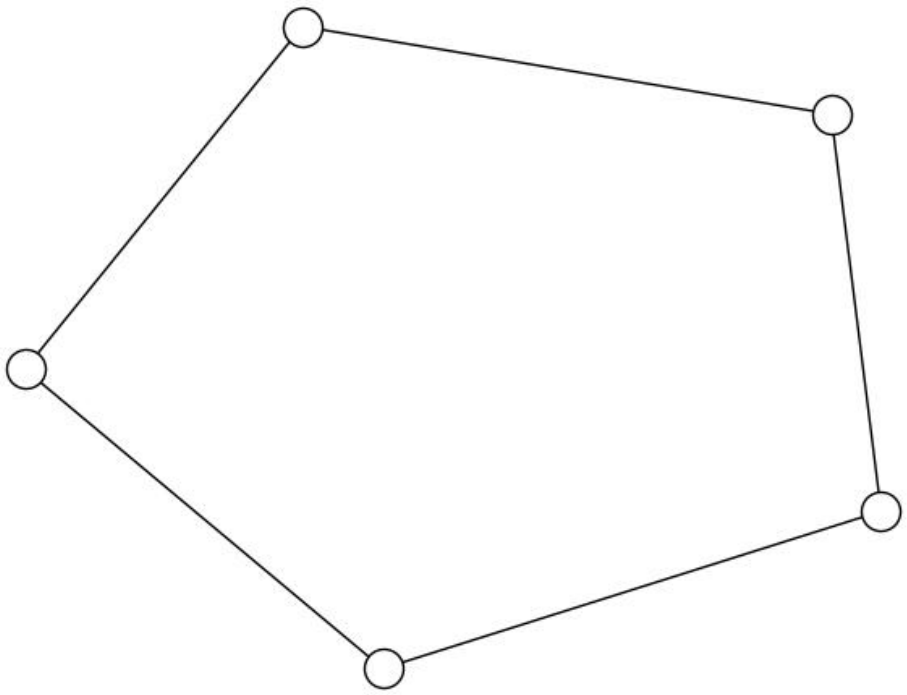
G Qualitative Case Audit and Output Analysis

Selection and duplicate handling. The analyzed collection contains 307 records organized for qualitative case analysis: 29 category-1 cases (both comparison models fail), 51 category-2 cases (the receiver succeeds and full merging fails), and 227 category-3 records excluded from the correction examples. The collection is outcome-selected, with other unsuccessful outputs of the proposed method outside its scope. Exact image-and-question matching yields 279 distinct records and 277 distinct decoded images. The displayed panels use distinct problems, with repeated runs grouped by image and question; semantic overlap across datasets may remain. Benchmark-level performance is evaluated separately in the H1–H3 experiments.

The main panels use DynaMath_4097 (source question 97, variant 5), DynaMath_7259 (question 259, variant 8), DynaMath_1314 (question 314, variant 2), MMStar_920, DynaMath_4325 (question 325, variant 5), and DynaMath_8243 (question 243, variant 9). Their decoded image hashes match the retained records. Model predictions, references, question identifiers, and evaluator-extracted answers were retained during the analysis and used to construct the qualitative examples reported in this appendix. Figure text summarizes these records. The retained run records correspond to the Qwen3-VL-8B family receiver and the associated Full and rank-one variants.

DynaMath 5234: reasoning about edge coloring

Question. Find the chromatic index of the following graph.



Correct answer. 3

Receiver	Full interpolation	Selective-RL
5.0 - wrong	5.0 - wrong	3.0 - correct

Reasoning summary. The graph is an odd cycle with maximum degree 2. Its edges cannot be colored with only two colors, but three suffice. Selective-RL correctly answers 3; Receiver and Full incorrectly answer 5.

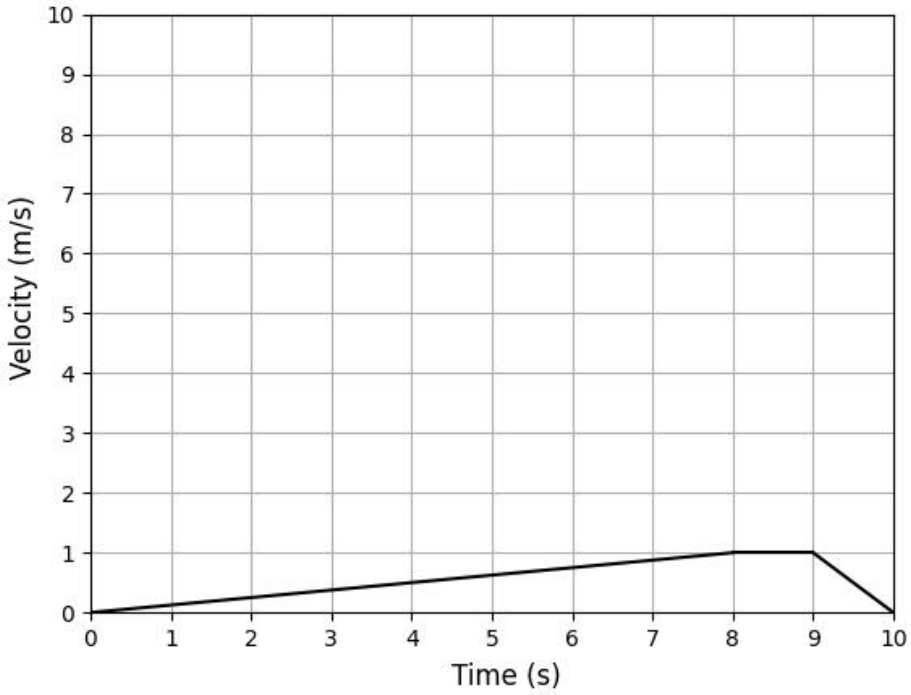
Displayed text summarizes the analyzed records; complete model outputs were retained during the analysis.

Figure 7: **An edge-coloring correction (DynaMath 5234; question 234, variant 6).** The graph is a five-vertex cycle, whose chromatic index is 3. The receiver and Full each answer 5. Rank-one merging correctly identifies that the odd cycle requires three edge colors.

Preserving correct receiver behavior. The following case illustrates a Full-interpolation error while the receiver and rank-one variant both answer correctly. Table 27 reports the separate H1 audit on 200 examples per model.

DynaMath 7268: integrating a velocity graph

Question. The motion of a car in a straight line is shown in the graph below. After 10 seconds, how far is the car from the starting point?



Correct answer. 5.5 m

Receiver	Full interpolation	Selective-RL
5.500 - correct	5.000 - wrong	5.500 - correct

Reasoning summary. The rising segment, constant-velocity interval, and falling segment contribute 4, 1, and 0.5 m, respectively. Selective-RL and Receiver correctly sum these to 5.5 m; Full treats 0-9 s as one rising segment and returns 5 m.

Displayed text summarizes the analyzed records; complete model outputs were retained during the analysis.

Figure 8: **Correct integration of a velocity graph (DynaMath 7268; question 268, variant 8).** The three segment areas give $4 + 1 + 0.5 = 5.5$ m. Full treats 0–9 s as a single rising segment and answers 5 m. Rank-one merging and the receiver both account for the constant-velocity interval and correctly answer 5.5 m.