WHEN GOOD VERIFIERS GO BAD: SILENT NEGATIVE TRANSFER IN VERIFIER-GUIDED VLM TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Verifier reliability is not portable across tasks. A verifier-guided self-DPO pipeline that produces genuine held-out gains on MathVista ($+9.6$ points on the self-training data and $+8.0$ points held out) can become harmful on MMMU. Crucially, the failure is not visible from the target-task self-training signal: averaged across six learner–verifier configurations, performance on the MMMU self-training data still improves by $+3.52$ points, while held-out performance drops by 1.42 points. We term this *silent negative transfer*:a verifier previously validated as useful can continue to exhibit the signatures of successful self-training after its induced update ceases to transfer to unseen data.

This raises a more basic question: how can we tell whether verifier supervision is reliable on the target task? A downstream failure alone cannot answer this question, because poor generalization may arise from two distinct sources: the verifier may induce an incorrect learning direction, or the correct learning direction itself may fail to generalize beyond the self-training distribution. We formalize these effects through *gradient fidelity* F, which measures alignment between the verifier-induced and correct training directions, and *gradient transferability* T, which measures alignment between the correct training and held-out directions. This decomposition yields a conservative Safe-Transfer Margin: positive first-order held-out alignment is certified when

$$\arccos F + \arccos T < \pi/2.$$

Motivated by this diagnosis, we introduce a simple yet effective precision-first filtering method, *Asymmetric Acceptance Gating* (AAG), which selects preference pairs according to the verifier's absolute confidence in the response receiving the positive update. Gradient-alignment analyses on MMMU show that AAG increases gradient fidelity from 0.29 to 0.42. In a separate same-cell analysis, the correct training direction remains positively aligned with the held-out direction ($T = 0.31$), yet raw verification rotates the induced update to negative held-out alignment (-0.13); AAG restores positive alignment ($+0.22$). Across all six MMMU learner–verifier configurations, AAG outperforms raw verifier-guided training on held-out evaluation.

These results redefine what it means for a verifier to be reliable: not merely that it has worked before, but that it induces the right update on the task at hand—and that this update remains useful beyond the data that generated it.

1 INTRODUCTION

Verifier-guided learning has become a scalable alternative to repeated human supervision (Bai et al., 2022; Lee et al., 2024). Model-based evaluators can provide preferences, rewards, critiques, or other judgments that are converted into optimization signals, enabling iterative alignment and self-improvement without requiring human labels at every round (Bai et al., 2022; Lee et al., 2024; Yuan et al., 2024). Related ideas also appear in scalable oversight and learned-verifier frameworks (Kenton et al., 2024; Zhang et al., 2025). These approaches share a central assumption: once model-generated judgments are sufficiently reliable, converting them into optimization signals should improve the model being trained (Sun et al., 2026; Zhao et al., 2026).

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.

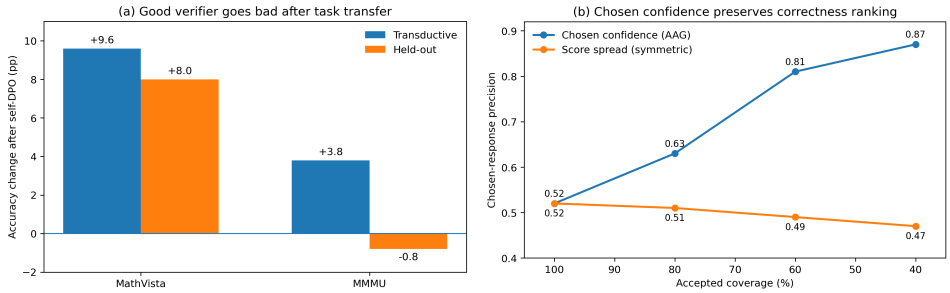

Figure 1: **Good verifiers can go bad under task transfer, while asymmetric chosen-confidence gating isolates cleaner preference supervision.** (a) For the same Qwen3-VL-4B learner–verifier configuration and self-DPO recipe, MathVista exhibits transferable improvement, whereas MMMU exhibits silent negative transfer: transductive performance improves while held-out performance declines. (b) On MMMU, increasing selectivity by chosen-response confidence raises chosen-response precision from 0.52 to 0.87. In contrast, symmetric score-spread filtering at matched coverage reduces precision from 0.52 to 0.47, showing that the gain comes from the asymmetric selection criterion rather than from filtering more aggressively alone.

Yet a reliable judge need not provide a reliable optimization signal. Prior work has studied the accuracy, calibration, and bias of model-based evaluators (Zheng et al., 2023; Stureborg et al., 2024). Learned judges and reward-model benchmarks further characterize the quality of model-generated evaluations (Kim et al., 2024a; Lambert et al., 2025). In parallel, preference-learning methods have examined robustness to noisy or unreliable supervision (Chowdhury et al., 2024; Wu et al., 2025).

More recent work has shown that verifier reliability can depend strongly on the evaluation regime, including problem difficulty, generator–verifier interactions, evaluator bias, and distribution shift (Zhou et al., 2026; Andrade et al., 2026). Related analyses show that reward and preference models can exploit signals that fail to generalize beyond the regime in which they were validated (Razin et al., 2026; Hwang et al., 2026). Theoretical studies of solver–verifier dynamics further suggest that verification quality and downstream self-improvement need not coincide (Sun et al., 2026).

These findings motivate a sharper question: *when does a verifier provide a reliable optimization signal?* We argue that judgment quality is only one part of the answer. Reliable verifier-guided learning requires two distinct directional agreements. First, the update induced by the verifier must align with the correct learning direction on the self-training distribution. Second, that correct learning direction must itself align with the direction that improves held-out performance. The two can fail independently: a verifier can rotate an otherwise useful update in the wrong direction, while even perfectly faithful supervision can recover a learning direction that fails to generalize.

These two failure modes correspond to two distinct geometric relationships. We formalize them through two alignments that are often conflated. Let g_V denote the update induced by verifier-generated preferences, g_P^* the counterfactual correct-preference update on the self-training distribution, and g_Q^* the corresponding correct-preference update on held-out data. We define

$$F = \cos(g_V, g_P^*), \qquad T = \cos(g_P^*, g_Q^*), \tag{1}$$

as *gradient fidelity* and *gradient transferability*, respectively. They satisfy

$$\cos(g_V, g_Q^*) \geq FT - \sqrt{1 - F^2}\sqrt{1 - T^2}, \tag{2}$$

which yields the *Safe-Transfer Margin*

$$M_{\text{safe}} = \frac{\pi}{2} - \arccos F - \arccos T. \tag{3}$$

A positive margin certifies positive first-order transfer under the correct-preference objective; a non-positive margin is deliberately inconclusive rather than predictive of failure. The margin is therefore a conservative diagnostic certificate rather than a downstream-accuracy predictor. The reference directions g_P^* and g_Q^* are used only for analysis and are never available to the training algorithm.

This geometry also explains why verifier accuracy alone can obscure optimization risk. A verifier used for training does not merely produce judgments; its decisions induce parameter updates. In particular, a *preference reversal*, in which an incorrect response is reinforced over a correct one, exactly negates the corresponding DPO update at initialization and can become increasingly influential as the policy separates the pair. Such errors are therefore directional optimization errors rather than exchangeable label noise.

This observation motivates *Asymmetric Acceptance Gating (AAG)*, inspired in part by DAPO's asymmetric clipping: rather than treating the two sides symmetrically, AAG explicitly prioritizes precision on the accepted positive. Instead of filtering by the score gap between chosen and rejected responses, AAG ranks pairs by the absolute confidence of the response that will receive the positive update. At fixed coverage, selecting the largest true posterior probabilities that the chosen response is correct maximizes selected-positive precision; AAG approximates this optimal selector when verifier confidence preserves that posterior ordering. This assumption requires rank consistency rather than numerical calibration and permits a label-free quantile implementation. It is directly supported by Fig. 1(b): chosen-response precision rises from 0.52 to 0.87 as chosen-confidence selection becomes stricter, whereas symmetric score-spread filtering reduces it from 0.52 to 0.47.

Our measurements support the predicted geometry on explicitly identified probes. On MMMU Qwen3-VL-4B/G3.1 at initialization, AAG increases gradient fidelity from 0.29 to 0.42. On a same-cell MMMU Qwen3-VL-8B/G2.5 probe, the correct direction has positive held-out alignment $T = 0.31$, yet raw verification yields $\cos(g_{\mathrm{Raw}}, g_Q^*) = -0.13$, while AAG restores it to $+0.22$. On the MathVista Qwen3-VL-4B/G3.1 positive control, a same-cell probe yields $F_{\mathrm{Raw}} = 0.76$, $F_{\mathrm{AAG}} = 0.86$, and $T = 0.78$, giving Safe-Transfer Margins of $+10.7^\circ$ and $+20.6^\circ$, respectively. Downstream, AAG outperforms raw verifier training on held-out MMMU in all six learner–verifier settings, shifting the mean effect from -1.42 to $+0.98$ points.

Our contributions are:

- **Silent negative transfer:** we identify a failure mode in which prior verifier success does not certify reliable supervision on a new task: self-training performance can continue to improve even as held-out performance deteriorates.
- **Directional diagnosis:** we disentangle gradient fidelity from gradient transferability and derive a conservative Safe-Transfer Margin.
- **Theory-motivated mitigation:** we introduce Asymmetric Acceptance Gating (AAG), a simple yet effective precision-first rule that improves gradient fidelity and consistently outperforms raw verifier-guided training on held-out data.

Reliable verifier-guided learning therefore requires not merely a good verifier, but a learning direction that is both faithful and transferable.

2 PRELIMINARIES

Let P denote the distribution used to construct self-training experience and Q a disjoint held-out distribution. For $x \sim P$, a learner π_θ samples K responses $\mathcal{Y}(x) = \{y_1, \ldots, y_K\}$. A frozen verifier V assigns score $s_V(x, y)$ and forms

$$y^+ = \arg\max_{y \in \mathcal{Y}(x)} s_V(x, y), \qquad y^- = \arg\min_{y \in \mathcal{Y}(x)} s_V(x, y). \tag{4}$$

We optimize the resulting pair $z = (x, y^+, y^-)$ using DPO (Rafailov et al., 2023):

$$\mathcal{L}_{\mathrm{DPO}}(z) = -\log \sigma(\beta \Delta_\theta(z)), \tag{5}$$

where

$$\Delta_\theta(z) = \log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\mathrm{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\mathrm{ref}}(y^- \mid x)}. \tag{6}$$

Neither raw self-DPO nor our method uses ground-truth preferences. Ground truth is used only for controlled baselines and post-hoc diagnostics.

For analysis, let g_V be the expected update induced by verifier-selected pairs on P. We define g_P^* and g_Q^* as the counterfactual update directions induced by correct preferences on P and Q, respectively. Equivalently, g_P^* and g_Q^* are gradients of the corresponding correct-preference objectives. These reference directions are never available to the training algorithm.

3 Gradient Geometry of Verifier-Guided Learning

3.1 Two alignments govern transfer

We distinguish two forms of alignment:

$$F = \cos(g_V, g_P^*), \qquad T = \cos(g_P^*, g_Q^*), \tag{7}$$

where F is *gradient fidelity* and T is *gradient transferability*. The former measures whether verifier supervision preserves the direction of correct learning; the latter measures whether that direction transfers beyond the improvement distribution.

Theorem 1 (Two-alignment transfer bound). For nonzero g_V, g_P^*, g_Q^*,

$$\cos(g_V, g_Q^*) \geq FT - \sqrt{1-F^2}\sqrt{1-T^2}. \tag{8}$$

Proof sketch. Let $\hat{v}, \hat{p}, \hat{q}$ be the corresponding unit vectors. Decompose

$$\hat{v} = F\hat{p} + \sqrt{1-F^2}\, a, \tag{9}$$

$$\hat{q} = T\hat{p} + \sqrt{1-T^2}\, b, \tag{10}$$

where $a, b \perp \hat{p}$ and $\|a\| = \|b\| = 1$. Then

$$\hat{v}^\top \hat{q} = FT + \sqrt{1-F^2}\sqrt{1-T^2}\, a^\top b, \tag{11}$$

and Eq. 8 follows from $a^\top b \geq -1$.

Corollary 1 (Safe-Transfer Margin). Define

$$M_{\text{safe}} = \frac{\pi}{2} - \arccos F - \arccos T. \tag{12}$$

If $M_{\text{safe}} > 0$, then $\langle g_V, g_Q^* \rangle > 0$, certifying positive first-order transfer under the held-out correct-preference objective. A non-positive margin is inconclusive rather than predictive of failure.

This decomposition exposes two distinct regimes: *transfer-limited* self-improvement, where the correct update itself transfers poorly, and *fidelity-limited* self-improvement, where a transferable update is corrupted by the verifier.

3.2 Preference reversal as an anti-gradient

Verifier errors differ in their optimization consequences. Define

$$h_\theta(z) = \nabla_\theta \log \pi_\theta(y^+ \mid x) - \nabla_\theta \log \pi_\theta(y^- \mid x). \tag{13}$$

The DPO update direction is

$$u_\theta(z) = \beta\sigma(-\beta\Delta_\theta(z))h_\theta(z). \tag{14}$$

Consider a *preference reversal*, where an incorrect response is chosen over a correct one, and let z^R denote the correctly ordered pair.

Proposition 1 (Reversal anti-gradient). At $\theta = \theta_{\text{ref}}$,

$$u_\theta(z) = -u_\theta(z^R). \tag{15}$$

Thus a reversal is an exact anti-gradient at DPO initialization.

More generally, if the policy assigns DPO gap $\Delta > 0$ to the correct ordering, then for the same pair geometry,

$$\frac{\|u_{\text{rev}}\|}{\|u_{\text{correct}}\|} = e^{\beta\Delta}. \tag{16}$$

Hence reversed supervision can become increasingly influential once the learner already favors the correct ordering.

Under the idealized model in which identical informative gradients are independently reversed with probability ρ,

$$\mathbb{E}[g_V] = (1 - 2\rho)g^*. \tag{17}$$

Eq. 17 describes cancellation of directional signal, not downstream accuracy. Structured verifier errors may additionally introduce orthogonal components and rotate the aggregate gradient, reducing fidelity.

4 Asymmetric Acceptance Gating

The geometry above suggests that selection should prioritize the reliability of the response receiving the positive update. Let

$$C(z) = \mathbb{1}\{y^+ \text{ is correct}\}, \qquad p(z) = \Pr(C(z) = 1 \mid z). \tag{18}$$

This is a selective-classification / reject-option rule (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017).

Theorem 2 (Optimal selection at fixed coverage). Among all selectors $A(z) \in \{0, 1\}$ satisfying $\Pr(A = 1) = \alpha$, chosen-response precision $\Pr(C = 1 \mid A = 1)$ is maximized by selecting the largest values of $p(z)$:

$$A^*(z) = \mathbb{1}\{p(z) \geq t_\alpha\}, \tag{19}$$

with boundary randomization when necessary.

Proof sketch. Since the acceptance mass is fixed, maximizing $\Pr(C = 1 \mid A = 1)$ is equivalent to maximizing $\mathbb{E}[A\,C] = \mathbb{E}[A\,p(z)]$, which is achieved by assigning the available acceptance mass to the largest posterior values.

The posterior $p(z)$ is unavailable during self-improvement. We therefore propose *Asymmetric Acceptance Gating (AAG)*. At fixed coverage α, AAG retains the top-confidence chosen responses:

$$A_\alpha(z) = \mathbb{1}\left\{s_V(x, y^+) \geq q_{1-\alpha}\right\}, \tag{20}$$

where $q_{1-\alpha}$ is the empirical $(1 - \alpha)$-quantile of chosen-response confidence over the eligible pair pool. Equivalently, one may use an absolute operating threshold,

$$A_\tau(z) = \mathbb{1}\left\{s_V(x, y^+) \geq \tau\right\}. \tag{21}$$

If chosen-response confidence preserves the ordering of $p(z)$, both parameterizations realize the same ranking as the precision-optimal selector. This assumption is weaker than calibration and is directly tested in Fig. 1(b).

AAG differs fundamentally from symmetric margin filtering,

$$A_\delta^{\text{sym}}(z) = \mathbb{1}\left\{s_V(y^+) - s_V(y^-) \geq \delta\right\}. \tag{22}$$

A large relative margin need not imply that y^+ itself is reliable; therefore symmetric margin has no corresponding fixed-coverage precision guarantee. AAG instead controls the quantity directly coupled to the positive DPO update: the trustworthiness of the response being reinforced.

5 Experiments

We evaluate four predictions of our analysis: (i) verifier reliability can fail to transfer across tasks even when self-training appears successful; (ii) harmful supervision is concentrated in directional preference errors; (iii) the resulting gradients exhibit the fidelity- and transfer-limited regimes predicted by Sec. 3; and (iv) asymmetric selection mitigates these failures for reasons beyond simply reducing the number of training pairs.

5.1 EXPERIMENTAL SETUP

Benchmarks and evaluation protocols. We study MathVista (Lu et al., 2024) and MMMU (Yue et al., 2024). MathVista provides a positive-control regime in which verifier-guided self-DPO genuinely transfers to unseen data. For MMMU, we evaluate two complementary settings. In the *transductive* setting, the prompts used to construct self-training experience are also used for evaluation. In the *held-out* setting, the 900 examples are split into 700 self-improvement prompts and a sealed 200-example test set; the latter never participates in candidate generation, pair construction, or optimization.

For MathVista, the positive-control experiment contains 500 transductive and 300 held-out examples.

Evaluation metric and pair universe. The primary MMMU held-out metric is macro-averaged accuracy over 30 subject groups. For pair-level analyses, we distinguish all generated prompt groups from the DPO-eligible pair pool. The Qwen3-VL-4B/G3.1 threshold intervention contains 230 eligible pairs after the training-pair validity checks. Table 3 uses this 230-pair training pool, whereas Table 13 uses the broader valid-confidence analysis pool before the DPO-eligibility restriction.

Learners and verifiers. We study Qwen3-VL-4B, Qwen2.5-VL-7B, and Qwen3-VL-8B with two frozen verifiers, Gemini-3.1-Pro (G3.1) and Gemini-2.5-Pro (G2.5), yielding six MMMU configurations. The learners span two model generations, so we do not interpret their differences as a pure scaling trend. Each prompt produces $K = 4$ candidates. Unless otherwise specified, self-DPO uses $\beta = 0.1$ and learning rate 10^{-5}. Qwen3-VL-4B multi-seed runs use seeds $\{17, 42, 123\}$; larger-model runs use Adafactor with gradient clipping at 0.5.

Verifier role and candidate availability. The verifier is used as a *candidate evaluator*, not as a direct answer generator: it scores learner-generated responses given the image and question without access to the ground-truth answer. Consequently, chosen-response precision is upper-bounded by candidate oracle availability,

$$\mathrm{pass@4}_{\mathrm{cand}} = \Pr(\exists\, y \in \mathcal{Y}(x) \text{ that is correct}) .$$

Appendix A.6 reports candidate availability and verifier selection efficiency. Directly distilling verifier-generated answers would study a different supervision channel; our question is whether an evaluator can be safely converted into an optimization signal over learner-generated experience.

Methods. *Frozen* denotes the pretrained learner before self-improvement. *Raw* constructs DPO preferences directly from verifier scores. *AAG* applies the asymmetric acceptance rule in Eq. 21; unless otherwise stated we use $\tau = 0.9$, fixed without access to held-out labels or held-out performance. We additionally evaluate the equivalent coverage-controlled rule in Eq. 20. *Correct-pref.* uses ground-truth correctness to order candidate responses and serves only as a diagnostic reference; it is not an upper bound on held-out performance. Neither Raw nor AAG accesses ground-truth labels.

Diagnostics. We report accuracy, chosen-response precision, and preference-reversal rate. A reversal occurs when the chosen response is incorrect while the rejected response is correct. For gradient diagnostics we measure

$$F = \cos(g_V, g_P^*), \qquad T = \cos(g_P^*, g_Q^*),$$

and direct held-out alignment $\cos(g_V, g_Q^*)$. Ground truth is used only to construct these post-hoc diagnostic reference directions.

5.2 VERIFIER TRUST DOES NOT TRANSFER ACROSS TASKS

We first test the premise behind Fig. 1(a). On MathVista, the Qwen3-VL-4B/G3.1 verifier–learner pair is a genuine positive control: raw self-DPO improves transductive accuracy from 0.654 to 0.750 ($+9.6$ points), while held-out accuracy improves from 0.650 to 0.730 ($+8.0$ points). Thus the verifier does not merely fit the experience used for self-training; its supervision produces transferable improvement.

Table 1: MMMU accuracy under transductive and held-out evaluation. Bold denotes the better of Raw and AAG. Correct-pref. is a diagnostic ground-truth-ordered reference. The Qwen3-VL-4B transductive rows report three-seed means; the Qwen3-VL-4B held-out rows use the canonical seed shared with the single-run Qwen2.5-VL-7B/Qwen3-VL-8B comparison, with three-seed held-out robustness in Appendix A.2.

Learner	Verifier	Frozen	Raw	AAG	Correct-pref.
Transductive					
Qwen3-VL-4B	G3.1	.495	.533	**.590**	.610
Qwen3-VL-4B	G2.5	.495	.535	**.560**	.558
Qwen2.5-VL-7B	G3.1	.550	.597	**.611**	.615
Qwen2.5-VL-7B	G2.5	.550	.605	**.610**	.622
Qwen3-VL-8B	G3.1	.550	**.593**	.588	.577
Qwen3-VL-8B	G2.5	.550	.538	**.583**	.608
Mean gain vs. frozen		–	+.035	+.059	+.067
Held-out					
Qwen3-VL-4B	G3.1	.514	.506	**.538**	.511
Qwen3-VL-4B	G2.5	.514	.521	**.527**	.517
Qwen2.5-VL-7B	G3.1	.476	.453	**.492**	.503
Qwen2.5-VL-7B	G2.5	.476	.459	**.484**	.494
Qwen3-VL-8B	G3.1	.533	.538	**.554**	.556
Qwen3-VL-8B	G2.5	.533	.484	**.510**	.513
Mean gain vs. frozen		–	-.014	+.010	+.008

The same recipe changes character on MMMU. For Qwen3-VL-4B/G3.1, transductive accuracy still improves by 3.8 points, yet held-out accuracy falls by 0.8 points. Table 1 shows that this discrepancy is systematic across learner configurations and verifiers.

Transductively, Raw improves five of six configurations and gains 3.52 points on average. AAG and Correct-pref. gain 5.87 and 6.67 points, respectively. Viewed only through this setting, verifier-guided self-improvement appears broadly successful.

Held-out evaluation gives the opposite diagnosis. Raw improves only two of six configurations and averages a 1.42-point regression. Its measured effect therefore changes from $+3.52$ points transductively to -1.42 points held out. AAG outperforms Raw in all six configurations and averages $+0.98$ points relative to frozen, a 2.4-point rescue over Raw. All six AAG–Raw held-out differences are positive (exact two-sided sign test, $p = 0.031$).

Correct-pref. supervision also contracts sharply, from $+6.67$ points transductively to $+0.80$ points held out. Thus the dominant MMMU bottleneck is transferability: even correct preferences on P produce little gain on Q. AAG should therefore be interpreted primarily as a *mitigation* mechanism: it removes verifier-induced directional damage, shifting the mean held-out effect from -1.42 to $+0.98$ points, rather than creating large transferable gains where the correct training direction itself transfers weakly.

5.3 AAG REMOVES DIRECTIONAL PREFERENCE ERRORS

We next examine whether AAG changes the quantity predicted by the pair-level theory: the reliability of the response receiving the positive update. Across the DPO-eligible pair pool, AAG increases chosen-response precision in every learner–verifier configuration (Table 2).

The improvement specifically suppresses wrong-direction pairs. For Qwen3-VL-4B/G3.1, reversal rate falls from 12.0% to 4.6%; for Qwen3-VL-8B/G2.5 it falls from 15.1% to 12.9%. These are precisely the pairs for which DPO reinforces an incorrect response while suppressing a correct one.

We test this mechanism more directly by intervening on the AAG threshold for Qwen3-VL-4B/G3.1. Table 3 shows a sharp improvement when moving away from raw supervision.

At $\tau = .50$, chosen precision rises from 0.409 to 0.606, reversal rate falls from 0.120 to 0.045, and the downstream gain increases from 2.0 to 10.7 points. Thresholds $\tau = .50$–$.90$ form a broad

Table 2: Chosen-response precision on the MMMU DPO-eligible pair pool. AAG increases the true correctness of the positive DPO target in all six learner–verifier configurations.

Learner	Verifier	Raw	AAG	Change
Qwen3-VL-4B	G3.1	.41	.64	+.23
Qwen3-VL-4B	G2.5	.31	.49	+.18
Qwen2.5-VL-7B	G3.1	.78	.83	+.05
Qwen2.5-VL-7B	G2.5	.60	.74	+.14
Qwen3-VL-8B	G3.1	.46	.69	+.23
Qwen3-VL-8B	G2.5	.35	.52	+.17

Table 3: Threshold intervention for Qwen3-VL-4B/G3.1 on the 230-pair DPO-eligible MMMU pool using seed 17. Δ accuracy is transductive; Table 1 reports the corresponding Qwen3-VL-4B transductive three-seed mean.

τ	Pairs	Precision	Reversal	Δ accuracy
0	230	.409	.120	+.020
.50	132	.606	.045	+.107
.70	125	.624	.045	+.108
.90	121	.636	.046	+.100
.99	119	.639	.047	+.078

plateau near $+10$ points. Because the $\tau = .99$ drop is smaller than observed across-seed variation, we do not interpret it as a distinct optimum.

The same conclusion holds when AAG is parameterized directly by coverage: transductive gains remain within 0.4 points of their optimum throughout $40\% - 70\%$ coverage, peaking at 60% coverage. This broad plateau indicates that the method does not rely on a single numerically tuned confidence threshold.

Eq. 17 should not be interpreted as a quantitative predictor of these accuracy gains. Its assumptions deliberately isolate directional cancellation, whereas real verifier errors vary in magnitude, geometry, and pair composition.

5.4 Real gradients exhibit both predicted failure modes

We keep all gradient measurements tied to an explicit task, learner, verifier, checkpoint, and analysis pool; Appendix A.8 collects the complete bookkeeping in one table. On *Probe A* (MMMU Qwen3-VL-4B/G3.1 at θ_0), AAG raises gradient fidelity from $F = .29$ to $.42$. On the same-cell *Probe B* (MMMU Qwen3-VL-8B/G2.5 at θ_0), the correct direction transfers positively ($T = .31$), while Raw gives $H = \cos(g_V, g_Q^*) = -.13$ and AAG restores $H = +.22$. The same probe also gives $F_{\mathrm{Raw}} = .38$, $F_{\mathrm{AAG}} = .66$, and Safe-Transfer Margins of $-49.6°$ and $-30.6°$, respectively. The negative MMMU margins are inconclusive despite AAG restoring positive direct held-out alignment.

On *Probe C* (MathVista Qwen3-VL-4B/G3.1 at θ_0), $F_{\mathrm{Raw}} = .76$, $F_{\mathrm{AAG}} = .86$, $T = .78$, and held-out alignments are $.61/.70$, yielding margins $+10.7°/+20.6°$. Thus the sufficient condition certifies MathVista while remaining inconclusive on MMMU.

5.5 Why asymmetric selection matters

Theorem 2 predicts that fixed-coverage selection should rank examples by chosen-response correctness rather than pairwise score spread. On MMMU Qwen3-VL-4B/G3.1, AAG precision rises from $.52$ to $.87$ as coverage falls from 100% to 40%, whereas symmetric filtering falls from $.52$ to $.47$. The same pattern holds for MMMU Qwen3-VL-8B/G2.5 and MathVista (Appendix A.10). These results support the rank-consistency assumption behind AAG and show that a large pairwise margin is not a reliable proxy for positive-target correctness.

5.6 Controls and independent replication

Matched-budget controls use separate reruns in which all methods within a row use the same retained-example budget; they are therefore not the full-data Raw runs reported elsewhere and should be compared within row. Under this control, MMMU Raw, Random-N, and AAG yield $+4.7$, $+6.5$, and $+8.7$ points, while MathVista Raw, Random-N, AAG, and Symmetric yield $+4$, $+4$, $+10$, and -2 points, respectively. Appendix A.5 further compares AAG with robust DPO, confidence-weighted DPO, a dual-threshold rule, positive-only AAG-SFT, and Correct-pref. restricted to the AAG-selected subset. Appendix A.6 separates verifier selection quality from candidate pass@4, and Appendix A.7 reports an independent ScienceQA visual-reasoning replication. Together, these controls test whether the effect comes from fewer examples, generic noise robustness, positive-only distillation, candidate availability, or a benchmark-specific artifact.

6 Concluding Remarks

Related work. Our work connects model-generated supervision, selective self-training, model-based evaluation, reward overoptimization, and robust preference optimization. Prior work studies how stronger models or iterative self-training procedures can provide preferences, critiques, rewards, or filtered supervision (Bai et al., 2022; Lee et al., 2024; Yuan et al., 2024; Zelikman et al., 2022; Singh et al., 2024), how reliably language and vision-language models evaluate other models (Zheng et al., 2023; Lambert et al., 2025; Kim et al., 2024b; Stureborg et al., 2024), and how preference optimization behaves under noisy or corrupted feedback (Rafailov et al., 2023; Chowdhury et al., 2024; Wu et al., 2025; Choi et al., 2025; Liu et al., 2026). Our setting is also related to reward overoptimization, where improving an imperfect proxy can degrade the underlying target (Gao et al., 2023). We study a complementary failure: after task transfer, verifier judgments can still support apparent transductive improvement while inducing a direction that fails on unseen data. The two-alignment view separates corruption of the verifier-induced gradient from non-transferability of the correct learning direction.

What counts as successful self-improvement? Improvement on prompts participating in the self-training loop is not, by itself, evidence of transferable progress: once those prompts generate responses that affect model parameters, they become part of the learning process even without observed ground-truth labels. Performance can therefore improve transductively while degrading on genuinely unseen data.

The two-alignment view separates two causes. High gradient fidelity cannot compensate for a learning direction that does not transfer, while high transferability cannot protect against a verifier that rotates an otherwise useful update. The Safe-Transfer Margin captures this interaction as a conservative local certificate rather than a universal predictor. Consistent with this view, MMMU is primarily transfer-limited: even Correct-pref. supervision transfers weakly, while AAG mainly removes additional verifier-induced damage.

Limitations and open questions. Our analysis is first-order and does not capture full nonlinear training dynamics. AAG assumes that verifier confidence preserves useful ordering of chosen-response correctness, which stronger distribution shift may violate. Due to limited GPU availability, we run multi-seed experiments only for Qwen3-VL-4B; Qwen2.5-VL-7B and Qwen3-VL-8B use a single reported seed after preliminary seed checks showed consistent results. G3.1 and G2.5 are closed-source APIs whose behavior may change across versions. Extending the analysis to online, agentic, open-ended, and multi-round settings remains future work.

Conclusion. We identify *silent negative transfer*: a verifier that produces genuine held-out gains on one task can become a harmful optimization signal on another while self-training still appears successful. Reliable verifier-guided learning requires both a faithful update direction and one that transfers to unseen data. This view explains preference reversal as a directional optimization error and motivates AAG as a precision-first mechanism for mitigating verifier-induced negative transfer.

More broadly, once verifier judgments enter an optimization loop, evaluation errors become gradient errors. A good verifier is therefore insufficient unless the learning direction it induces is both faithful and transferable.

REFERENCES

Moises Andrade, Joonhyuk Cha, Brandon Ho, Vriksha Srihari, Karmesh Yadav, and Zsolt Kira. Let's think in two steps: Mitigating agreement bias in MLLMs with self-grounded verification. In *International Conference on Learning Representations*, 2026.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Olivier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. In *International Conference on Learning Representations*, 2025.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42258–42274. PMLR, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Minjune Hwang, Yigit Korkmaz, Daniel Seita, and Erdem Bıyık. Causally robust reward learning from reason-augmented preference feedback. In *International Conference on Learning Representations*, 2026.

Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah. On scalable oversight with weak LLMs judging strong LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Seungone Kim, Jay Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan S. Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, 2024a.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353. Association for Computational Linguistics, 2024b.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797. Association for Computational Linguistics, 2025.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26874–26901. PMLR, 2024.

Xinxin Liu, Ming Li, Zonglin Lyu, Yuzhang Shang, and Chen Chen. Learning from noisy preferences: A semi-supervised learning approach to direct preference optimization. In *International Conference on Learning Representations*, 2026.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Noam Razin, Yong Lin, Jiarui Yao, and Sanjeev Arora. Why is your language model a poor implicit reward model? In *International Conference on Learning Representations*, 2026.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin F. Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Yifan Sun, Yushan Liang, Zhen Zhang, Xin Liu, and Jiaye Teng. Theoretical modeling of large language model self-improvement training dynamics through solver-verifier gap. In *International Conference on Learning Representations*, 2026.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *International Conference on Learning Representations*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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*, pp. 9556–9567, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *International Conference on Learning Representations*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Yefan Zhou, Austin Xu, Yilun Zhou, Janvijay Singh, Jiang Gui, and Shafiq Joty. Variation in verification: Understanding verification dynamics in large language models. In *International Conference on Learning Representations*, 2026.

A Additional experimental results

A.1 Three-seed transductive results

Table 4 reports the individual Qwen3-VL-4B transductive seed runs underlying the mean values in the main text. The reported standard deviations use the three experimental runs.

Table 4: Three-run transductive MMMU accuracy for the Qwen3-VL-4B learner.

Verifier	Method	Seed 17	Seed 42	Seed 123	Mean $\pm$ std.
G3.1	Correct-pref.	.620	.595	.615	$.610 \pm .011$
	AAG	.595	.615	.560	$.590 \pm .023$
	Raw	.515	.540	.545	$.533 \pm .013$
G2.5	Correct-pref.	.575	.535	.565	$.558 \pm .017$
	AAG	.575	.550	.555	$.560 \pm .011$
	Raw	.545	.520	.540	$.535 \pm .011$

For G3.1, the paired AAG–Raw differences are

$$+.080, \quad +.075, \quad +.015.$$

For G2.5 they are

$$+.030, \quad +.030, \quad +.015.$$

Thus AAG improves over Raw in all six paired runs.

A.2 Held-out multi-seed robustness

Table 5 repeats the Qwen3-VL-4B held-out comparison across three training seeds. The canonical seed-42 values match the Qwen3-VL-4B entries reported in the held-out panel of Table 1.

Table 5: Three-seed held-out MMMU accuracy for the Qwen3-VL-4B learner. Mean $\pm$ std. quantifies variation across training seeds, not sampling uncertainty of the fixed 200-example held-out set.

Verifier	Method	Seed 17	Seed 42	Seed 123	Mean $\pm$ std.
G3.1	Raw	.497	.506	.512	$.505 \pm .006$
	AAG	.531	.538	.550	$.540 \pm .008$
G2.5	Raw	.509	.521	.515	$.515 \pm .005$
	AAG	.538	.527	.540	$.535 \pm .006$

AAG exceeds Raw in all six paired seed runs, with mean advantages of 3.5 and 2.0 points for G3.1 and G2.5, respectively. These seed-level standard deviations characterize optimization variability; they should not be interpreted as confidence intervals for the finite held-out evaluation sample.

A.3 MATHVISTA POSITIVE CONTROL

Table 6 reports the raw self-DPO positive control used in Fig. 1(a). No Correct-pref. arm is used in this experiment.

Table 6: MathVista positive-control results for Qwen3-VL-4B/G3.1 raw self-DPO.

Evaluation	N	Frozen	Raw	Change
Transductive	500	.654	.750	+.096
Held-out	300	.650	.730	+.080

A.4 MATCHED-DATA DOWNSTREAM CONTROLS

Table 7 reports separate budget-matched reruns: within each row, Raw, Random-N, and selective methods use the same retained-example budget, set by the corresponding selective condition. These values are therefore not directly comparable to the full-data Raw gains in Tables 1 and 6; the relevant comparison is within each row.

Table 7: Downstream accuracy changes in separate matched-budget controls. All methods within each row use the same retained-example budget; these are not the full-data Raw runs reported in the main tables.

Setting	Raw	Random-N	AAG	Symmetric
MMMU matched-budget control	+.047	+.065	+.087	–
MathVista matched-budget control	+.040	+.040	+.100	-.020

Within the matched budget, generic subsampling helps on MMMU ($+.065$ versus $+.047$) but does not recover the AAG gain ($+.087$); on MathVista, AAG reaches $+.100$ while Raw and Random-N remain at $+.040$, and symmetric filtering regresses.

A.5 NOISE-ROBUST AND SOFT-SELECTION BASELINES

We compare AAG with alternatives that address preference noise without hard gating. For robust DPO (rDPO; Chowdhury et al., 2024), we sweep the flip-rate parameter $\epsilon \in \{.1, .2, .3, .4\}$ on a training-only calibration split and use $\epsilon = .1$ for the final runs. Confidence-weighted DPO minimizes

$$\mathcal{L}_{\text{wDPO}} = \frac{\sum_{z \in \mathcal{B}} s_V(x, y^+)\, \mathcal{L}_{\text{DPO}}(z)}{\sum_{z \in \mathcal{B}} s_V(x, y^+)}, \tag{23}$$

so confidence changes each pair's relative contribution without changing the minibatch loss scale. The dual-threshold baseline retains pairs with high chosen confidence and low rejected confidence, with thresholds set to match AAG's retained-pair count. *Correct-pref.-AAG-subset* uses the same AAG-selected prompt groups but applies the ground-truth ordering rule, isolating selection from preference-label quality.

The observed pattern separates two effects. Correct-pref. over all pairs can transfer less well than AAG because AAG also changes the training subset; on the identical AAG-selected subset, ground-truth ordering recovers the strongest result. The observed pattern is monotonic in supervision quality. Raw DPO regresses below the frozen learner, while noise-robust or confidence-aware variants recover above frozen: rDPO reaches .520, confidence-weighted DPO .524, and the dual-threshold rule .527. AAG-SFT reaches .531, showing that selection alone is useful, while AAG-DPO attains the strongest label-free point estimate at .540. On the identical AAG-selected subset, replacing verifier ordering with ground-truth ordering further increases accuracy to .548, isolating the remaining preference-label error.

Table 8: Three-seed held-out MMMU Qwen3-VL-4B/G3.1 ablations. Values are mean ± standard deviation across training seeds.

Method	Held-out accuracy
Frozen	.514
Raw DPO	$.505 \pm .006$
Correct-pref. (all pairs)	$.512 \pm .006$
rDPO	$.520 \pm .006$
Confidence-weighted DPO	$.524 \pm .007$
Dual-threshold DPO	$.527 \pm .006$
AAG-SFT	$.531 \pm .007$
AAG-DPO	$.540 \pm .008$
Correct-pref.-AAG-subset	$.548 \pm .007$

A.6 Candidate availability and verifier selection efficiency

To distinguish evaluator quality from candidate quality, we report candidate oracle availability and selection efficiency,

$$\text{Eff.} = \frac{\Pr(y^+ \text{ correct})}{\Pr(\exists\, y \in \mathcal{Y}(x) \text{ correct})}.$$

Table 9: Candidate availability and Raw verifier selection efficiency on MMMU.

Learner	Verifier	Candidate pass@4	Raw chosen precision	Eff.
Qwen3-VL-4B	G3.1	.58	.41	.71
Qwen3-VL-4B	G2.5	.57	.31	.54
Qwen2.5-VL-7B	G3.1	.88	.78	.89
Qwen2.5-VL-7B	G2.5	.86	.60	.70
Qwen3-VL-8B	G3.1	.72	.46	.64
Qwen3-VL-8B	G2.5	.71	.35	.49

These values make clear that chosen-response precision is not the verifier's direct MMMU answering accuracy: the evaluator cannot select a correct response when the learner-generated candidate set contains none.

A.7 Replication on ScienceQA visual reasoning

We replicate the protocol on an independent ScienceQA visual subset, using 600 self-training and 300 sealed held-out examples.

Table 10: Independent ScienceQA visual-reasoning replication.

Quantity	Raw	AAG
Transductive gain	+5.1 pp	+7.0 pp
Held-out gain	-1.2 pp	+2.3 pp
Chosen precision	.48	.66
Reversal rate	.108	.054

The point-estimate held-out rescue is $+3.5$ points.

A.8 Unified gradient-probe bookkeeping

A.9 Gradient geometry over training

We follow the same pre-specified MMMU Qwen3-VL-8B/G2.5 fidelity-limited probe as Probe B in Table 11; the 0% row therefore matches the same-cell initialization values exactly.

Table 11: Gradient probes used in the main text. Each same-cell row uses one task, learner/verifier pair, checkpoint, and intermediate reference pool throughout.

Probe	Task / config.	Ckpt.	Role	$F_{\mathrm{Raw}}/F_{\mathrm{AAG}}$	T	$H_{\mathrm{Raw}}/H_{\mathrm{AAG}}$	$M_{\mathrm{Raw}}/M_{\mathrm{AAG}}$
A	MMMU Qwen3-VL-4B/G3.1	θ_0	fidelity	.29/.42	–	–	–
B	MMMU Qwen3-VL-8B/G2.5	θ_0	fidelity-limited	.38/.66	.31	-.13/.22	$-49.6°/-30.6°$
C	MathVista Qwen3-VL-4B/G3.1	θ_0	positive control	.76/.86	.78	.61/.70	$+10.7°/+20.6°$

Table 12: Gradient fidelity F and direct held-out alignment $H = \cos(g_V, g_Q^*)$ over the MMMU Qwen3-VL-8B/G2.5 trajectory.

Progress	F_{Raw}	F_{AAG}	H_{Raw}	H_{AAG}
0%	.38	.66	-.13	+.22
25%	.34	.64	-.16	+.20
50%	.29	.61	-.19	+.18
100%	.24	.58	-.22	+.14

AAG remains more faithful throughout optimization, while Raw held-out alignment becomes increasingly negative and AAG remains positive.

A.10 Coverage robustness

Table 13: Chosen-response precision under matched coverage on the valid-confidence analysis pool.

Dataset / setting	Coverage	AAG	Symmetric	Raw
MMMU Qwen3-VL-4B/G3.1	100%	.52	.52	–
	80%	.63	.51	–
	60%	.81	.49	–
	40%	.87	.47	–
MMMU Qwen3-VL-8B/G2.5	100%	.54	.54	–
	80%	.64	.53	–
	60%	.77	.50	–
	40%	.78	.40	–
MathVista	60%	.75	.47	.52
	30%	.87	.45	.52

Table 14: Coverage-controlled AAG on MMMU Qwen3-VL-4B/G3.1; Δ accuracy is transductive.

Coverage	Chosen precision	Δ accuracy
80%	.63	+.089
70%	.72	+.104
60%	.81	+.108
50%	.84	+.107
40%	.87	+.102
20%	.92	+.071

Performance exhibits a broad transductive plateau between 40% and 70% coverage rather than depending on a uniquely tuned operating point.