CONSISTENCY AS AN INDUCTIVE BIAS: MITIGATING CROSS-VIEW RELIABILITY GAP IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inductive biases steer learning toward generalizable solutions by encoding task structure. In this work, we identify a crucial missing bias in MLLMs: cross-view consistency, *i.e.*, semantically invariant views of the same instance should lead to the same answer. Standard reinforcement learning with verifiable rewards (RLVR) objectives do not impose this constraint, but instead assign pointwise rewards to each visual input. Even with data augmentation (DA), transformed views are typically rewarded independently, providing little signal once within-view rewards saturate. We propose **ConsistRoll**, a simple but effective method that injects cross-view consistency into RLVR training by reusing the group-sampling mechanism of GRPO. Specifically, ConsistRoll places original and semantically invariant transformed views in the same generation group, assigning a joint reward when paired completions are both correct. In this way, ConsistRoll turns consistency into an online credit-assignment signal, **without extra annotations and overhead**. Theoretically, we prove that a reward computed from finitely many paired responses recovers the gradient of this long-run reliability target on average, and quantify how same-group centering modifies the signal. Comprehensive 14 benchmarks across math, general-purpose, and hallucination domains confirm that ConsistRoll achieves robust improvements in cross-view stability and multimodal reasoning.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Huang et al., 2026b; Fu et al., 2026) increasingly learn reasoning skills through RLVR (Guo et al., 2025; Liu et al., 2025c; Xiao et al., 2026). Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a representative RLVR method that samples multiple responses, checks their answers with a verifier, and increases the likelihood of the relatively successful trajectories. However, the verifier only checks single-view correctness, never testing whether the same reasoning remains reliable when the view changes but the problem does not.

A fundamental premise of visual intelligence is *viewpoint invariance* (Zaheer et al., 2017; Bronstein et al., 2021), *i.e.*, when an observation undergoes an answer-preserving visual transformation, *e.g.*, rotation or affine repositioning, the model's factual deduction should remain faithfully stable regardless of the incidental perspective from which the visual evidence is framed. Despite RLVR's empirical success, we expose a critical, objective-level blind spot: ***cross-view reliability gap***. As shown in Figure 1 (a), in standard GRPO scheme, the verifier checks correctness strictly pointwise on isolated viewpoints, which encourages the policy to memorize perspective-specific visual shortcuts, but remains oblivious to whether the same underlying reasoning survives a specific visual perturbation.

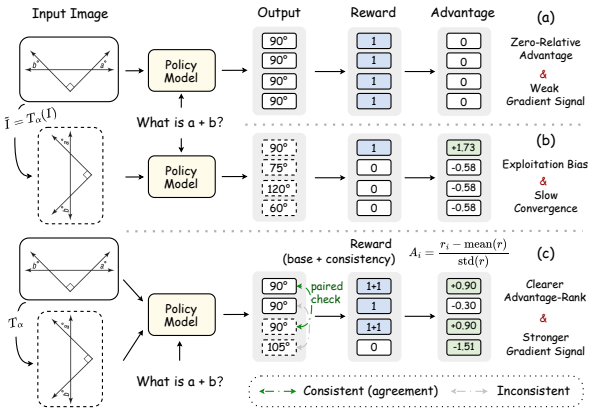


Figure 1: (a) Naive GRPO and (b) with DA strategy reward views independently, while (c) ConsistRoll turns cross-view agreement into on-policy credit assignment.

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.

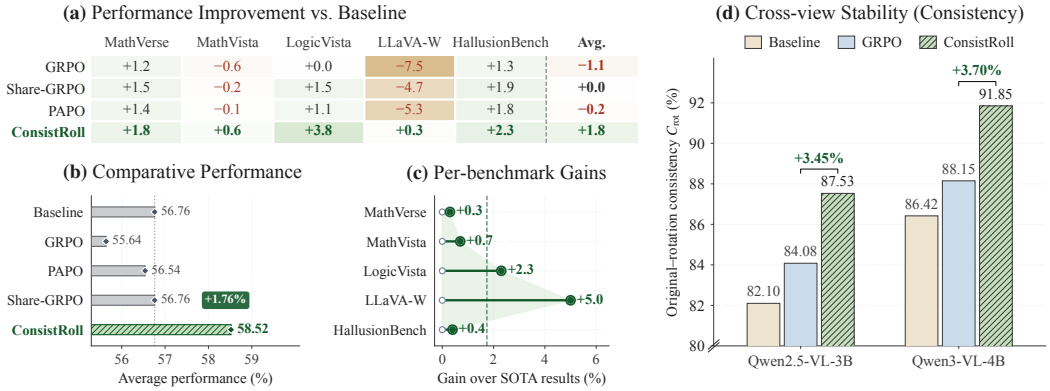


Figure 2: ConsistRoll improves performance and cross-view stability. (a-c) Under the same settings, ConsistRoll outperforms competing RLVR methods across different domains. (d) It improves cross-view stability over GRPO by 3.45% and 3.70% on Qwen2.5-VL-3B and Qwen3-VL-4B, respectively.

Motivated by this, we raise a fundamental research question:

How can we inject viewpoint invariance into multimodal RLVR as an intrinsic inductive bias, penalizing view fragility without sacrificing task accuracy or forcing artificial reasoning alignment?

A natural attempt to address this issue is data augmentation, *e.g.*, NoisyRollout (Liu et al., 2025a) simply feeds augmented viewpoints into the training pool. However, as illustrated in Figure 1 (b), merely exposing the policy to perturbed images within conventional RLVR falls short: because rewards are assessed pointwise and normalized independently within single-view rollouts, the optimizer cannot distinguish between a robust policy and a fragile one that achieves high average reward by dominating only canonical viewpoints while collapsing under perturbed ones. Conversely, strictly penalizing divergence between full reasoning traces or token distributions (Tarvainen & Valpola, 2017; Miyato et al., 2018) is overly restrictive for multimodal reasoning, as different visual perspectives often necessitate distinct spatial grounding orders and valid derivation paths to reach the same conclusion.

To resolve this dilemma, we propose **ConsistRoll**, a simple online RLVR framework for MLLMs that incorporates cross-view consistency directly into credit assignment without requiring additional annotations, external critics, or reasoning-trace alignment. The core insight of ConsistRoll is rooted in a simple missing statistic: *joint success* on verified view pairs. Rather than rewarding perspectives in isolation, ConsistRoll forms generation groups that include both an original image and the answer-preserving transformed image, under the same fixed rollout budget. Within each group, it assigns an intrinsic consistency reward exclusively when paired rollouts from both views are concurrently correct. In doing so, ConsistRoll transforms viewpoint invariance from passive evaluation into an active, on-policy optimization objective that favors reliable multimodal reasoning.

We substantiate ConsistRoll with theoretical grounding and empirical evidence. Theoretically, we show that optimizing joint success directly regularizes worst-view accuracy via Theorem 2.2 with an unbiased gradient estimator via Proposition 3.1, where correctness-gating naturally avoids trivial consensus without hurting task accuracy. Empirically, across 14 benchmarks covering math, perception, and hallucination, as shown in Figure 2, ConsistRoll consistently outperforms GRPO and strong baselines on Qwen2.5-VL and Qwen3-VL, effectively slashing cross-view reliability gaps.

2 Joint Success for Cross-View Reliability

We first formalize cross-view reliability and show that average correctness fails to distinguish robust policies from fragile ones. As Figure 3(a) illustrates, policies with equal mean accuracy can exhibit stark view sensitivity. To penalize such imbalance, we introduce joint success, requiring correct responses across independently sampled views, to reward balanced cross-view reliability.

2.1 The Missing Statistic: Joint Success

Let $Z = (x_0, x_1, y^*) \sim \mathcal{D}_{\text{pair}}$ denote a verified view pair, where $x_0 = (I, \xi)$ and $x_1 = (T(I), \xi)$ retain the same answer-relevant evidence and share the unique target y^*. We first condition on a fixed Z and use the same deterministic parser $\hat{a}$ on both views, and only after constructing its reliability score do we average over Z, which order prevents unrelated problems from being multiplied together.

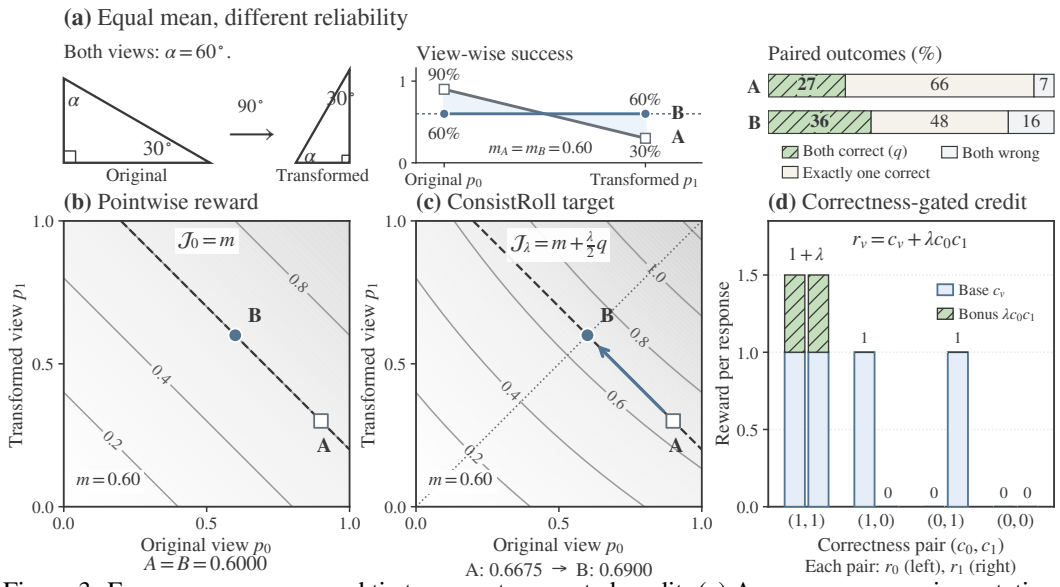


Figure 3: From an average-reward tie to correctness-gated credit. (a) An answer-preserving rotation, view-wise success profiles, and paired outcome distributions for illustrative policies with equal means. (b-c) Along the dashed $m = 0.60$ slice, pointwise reward ties A and B, whereas the raw ConsistRoll target favors B. (d) Rewarding response: only co-correct pairs receive the consistency bonus.

For $o_v \sim \pi_\theta(\cdot \mid x_v)$, let $c_v(o_v) = \mathbf{1}[\hat{a}(o_v) = y^*]$ record correctness and let $p_v(Z;\theta) = \mathbb{E}[c_v \mid Z]$ be the repeated-sampling success probability on view v. Write $m = (p_0 + p_1)/2$ for average success, $\Delta = |p_0 - p_1|$ for view imbalance, and $p_{\min} = \min(p_0, p_1)$ for weakest-view success. Mean success alone does not determine $p_{\min}$: a strong view can compensate for a weak one in the average. Joint success is a natural additional statistic because both responses must be correct, so its probability is limited by the weaker view's success rate. The following proposition makes this connection precise, *i.e.*, under independent response sampling, mean and joint success together recover $p_{\min}$ exactly.

Proposition 2.1 (Mean and joint success determine worst-view accuracy)**.** *Condition on a verified pair Z, draw one response independently from each view, and let $q(Z;\theta) = \mathbb{E}[c_0 c_1 \mid Z]$ be the probability that both are correct. Then*

$$q = p_0 p_1 = m^2 - \frac{\Delta^2}{4}, \qquad p_{\min} = m - \frac{\Delta}{2} = m - \sqrt{m^2 - q}. \tag{1}$$

Thus (m, q) exactly determines the unordered pair $\{p_0, p_1\}$ and worst-view correctness.

Specifically, joint success reveals how well it survives the view change. At fixed m, increasing q exactly shrinks the gap and improves the weaker view. In Figure 3a, the change from $q = 0.27$ to 0.36 identifies the improvement from $p_{\min} = 0.30$ to 0.60. Independence concerns sampled responses conditional on Z, not the images, which is an exact population identity for each pair, not a confidence certificate from one sampled pair or from corpus-level averages, and Appendix G.1 gives the proof.

2.2 An Accuracy-Monotone Target with a Balance Preference

Reducing the gap alone is insufficient: two consistently wrong views also have zero imbalance. We therefore retain mean correctness and add joint success. The coefficient $\lambda/2$ below is chosen so that λ is the bonus assigned to each response; Proposition 3.1 derives this correspondence.

Theorem 2.2 (The objective rewards both accuracy and balance)**.** *For $\lambda \geq 0$, define the pair-level target $\mathcal{J}_\lambda$ and its dataset-level reliability functional $\mathcal{R}_\lambda$ by*

$$\mathcal{J}_\lambda(p_0, p_1) := m + \frac{\lambda}{2} q = \underbrace{m + \frac{\lambda}{2} m^2}_{\textit{accuracy value}} - \underbrace{\frac{\lambda}{8}\Delta^2}_{\textit{imbalance cost}}, \tag{2}$$
$$\mathcal{R}_\lambda(\theta) := \mathbb{E}_{Z \sim \mathcal{D}_{\text{pair}}}[\mathcal{J}_\lambda(Z;\theta)].$$

The marginal weights satisfy

$$\frac{\partial \mathcal{J}_\lambda}{\partial p_v} = \frac{1}{2}\big(1 + \lambda p_{1-v}\big) > 0\,, \quad \frac{\partial \mathcal{J}_\lambda}{\partial p_1} - \frac{\partial \mathcal{J}_\lambda}{\partial p_0} = \frac{\lambda}{2}(p_0 - p_1). \tag{3}$$

Thus improving either view while holding the other fixed always increases the target, whose unique optimum on $[0,1]^2$ is $(1,1)$. At fixed mean and $\lambda > 0$, it strictly prefers more balanced success probabilities (strict Schur-concavity), with the unique maximum at $p_0 = p_1 = m$. When $p_0 > p_1$, the weaker view has the larger marginal weight. At $\lambda = 0$, both the balance preference and the marginal-weight difference disappear.

Reading the geometry. Figure 3 (b)(c) holds the mean fixed: the pointwise objective ties A and B, whereas $\mathcal{J}_\lambda$ increases from 0.6675 to 0.6900. The gain is exactly $\lambda\Delta_A^2/8$, with $\Delta_A = 0.6$. The target therefore prices fragility without imposing identical reasoning. These are preferences in probability space: shared model parameters can couple improvements and regressions, so the theorem does not guarantee that every training step improves both views, preserves mean accuracy, or reduces the gap. Appendix G.1 proves the claims and their limits.

Corollary 2.3 (K-view projection certificate)**.** *For $K \geq 2$, let $m = K^{-1}\mathbf{1}^\top \boldsymbol{p}$, $Q_K = \binom{K}{2}^{-1} \sum_{u<v} p_u p_v$, and $\mathbf{P}_\perp = \mathbf{I} - K^{-1}\mathbf{1}\mathbf{1}^\top$. Then*

$$\mathcal{J}_\lambda^{(K)} = m + \frac{\lambda}{2} Q_K = m + \frac{\lambda}{2} m^2 - \frac{\lambda \|\mathbf{P}_\perp \boldsymbol{p}\|_2^2}{2K(K-1)}, \quad p_{\min} \geq m - (K-1)\sqrt{m^2 - Q_K}. \tag{4}$$

For a fixed verified K-view instance, pairwise joint success measures the squared distance from equal reliability and lower-bounds weakest-view success. The bound is sharp; for $K = 2$ it is an equality.

Appendix G.2 proves the extension. The implementation uses two views: the remaining question is how to realize this probability-level preference using only sampled verifier outcomes.

3 ConsistRoll: From the Reliability Target to GRPO

Our ConsistRoll recipe is direct, *i.e.*, keep ordinary correctness, add a bonus when indexed responses on both views are correct, and return it only to the two responses. The flow of ConsistRoll in Fig. 4.

3.1 Verified Pairs with the Same Response Budget

ConsistRoll admits $T(I)$ only if a rule filter and image-aware verifier confirm preserved evidence and target; otherwise it uses ordinary GRPO. For each admitted pair, it fixes n responses, samples $M = n/2$ per view, and pairs by index, without text or reasoning-step alignment.

3.2 How Paired Rewards Give Response-Level Credit

Let $c_{v,i}$ be verified correctness, $b_i = c_{0,i}c_{1,i}$ observed joint success, and $r_{v,i} = c_{v,i} + \lambda b_i$ the reward for both responses. For the proof, $\psi_{v,i} = \nabla_\theta \log \pi_\theta(o_{v,i} \mid x_v)$ is standard response-likelihood sensitivity, so $r_{v,i}\psi_{v,i}$ assigns learning credit.

Proposition 3.1 (Finite rollouts recover the target gradient)**.** *Under conditionally independent response sampling, fixed verifier rewards, and Appendix G.1's differentiability and integrability conditions, every finite $M \geq 1$ satisfies*

$$g_M := \frac{1}{2M} \sum_{i=1}^{M} \sum_{v=0}^{1} r_{v,i}\psi_{v,i}, \quad \mathbb{E}[g_M \mid Z] = \nabla_\theta \mathcal{J}_\lambda(Z;\theta), \quad \mathbb{E}_{Z,o}[g_M] = \nabla_\theta \mathcal{R}_\lambda(\theta). \tag{5}$$

Although the long-run probabilities are unknown, g_M uses sampled zero-or-one results and is unbiased for the raw target gradient, first on each pair and then across the pair distribution. Averaging $2M$ response-local terms produces $\lambda/2$; globally assigning the pair mean would instead produce λ. Appendix G.1 proves both identities.

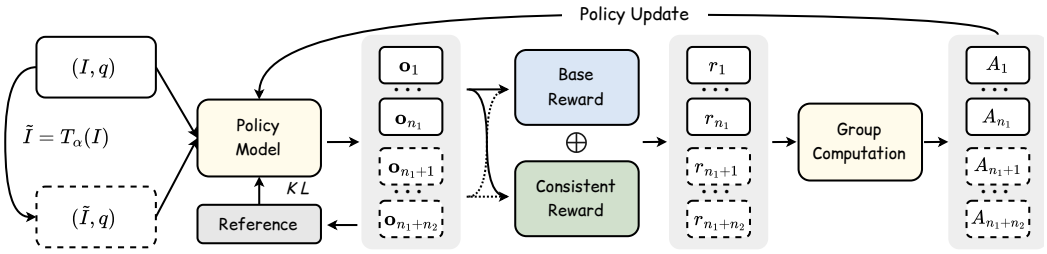


Figure 4: ConsistRoll pipeline. The consistency reward λb_i augments each response's correctness reward before pooled advantage estimation and the clipped update in Eq. (6).

3.3 What Practical GRPO Changes

GRPO compares responses within a sampled group: above-average rewards are reinforced, below-average rewards are suppressed, and a clipped likelihood ratio limits each token update. ConsistRoll changes only the group construction and rewards. It pools both views in $\mathcal{G} = \{(v, i) : v \in \{0, 1\}, i \in [M]\}$ and writes $\mathcal{O}_\mathcal{G} = \{o_{v,i}\}_{(v,i)\in\mathcal{G}}$. The frozen rollout law is $P_{\text{old}}(\mathcal{O}_\mathcal{G} \mid Z) = \prod_{(v,i)\in\mathcal{G}} \pi_{\theta_{\text{old}}}(o_{v,i} \mid x_v)$. For prefix $h_{v,i,t} = (x_v, o_{v,i,<t})$, define $\rho_{v,i,t}(\theta) = \pi_\theta(o_{v,i,t} \mid h_{v,i,t})/\pi_{\theta_{\text{old}}}(o_{v,i,t} \mid h_{v,i,t})$ and $\mathcal{K}_{v,i,t}(\theta) = D_{\text{KL}}(\pi_\theta(\cdot \mid h_{v,i,t}) \| \pi_{\text{ref}}(\cdot \mid h_{v,i,t}))$. With pooled reward mean $\bar{r}_\mathcal{G}$ and scale $s_\mathcal{G}$, the complete objective that is maximized is

$$\widehat{A}_{v,i} = \frac{r_{v,i} - \bar{r}_\mathcal{G}}{s_\mathcal{G} + \delta}, \quad \ell_\epsilon(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A\},$$
$$\mathcal{J}_{\text{ConsistRoll}}(\theta) = \mathbb{E}_{\substack{Z\sim\mathcal{D}_{\text{pair}},\\ \mathcal{O}_\mathcal{G}\sim P_{\text{old}}(\cdot|Z)}} \left[\frac{1}{2M} \sum_{(v,i)\in\mathcal{G}} \frac{1}{|o_{v,i}|} \sum_{t=1}^{|o_{v,i}|} \left(\ell_\epsilon(\rho_{v,i,t}, \widehat{A}_{v,i}) - \beta \mathcal{K}_{v,i,t} \right) \right]. \tag{6}$$

Here $\delta > 0$ stabilizes the reward scale, $\epsilon > 0$ controls clipping, and $\beta \geq 0$ weights the reference-policy KL. To expose the two sources of cross-view credit, let $\bar{c}_v = M^{-1} \sum_i c_{v,i}$, $\bar{c} = (\bar{c}_0 + \bar{c}_1)/2$, and $\bar{b} = M^{-1} \sum_i b_i$. Then

$$r_{v,i} - \bar{r}_\mathcal{G} = \underbrace{(c_{v,i} - \bar{c}_v)}_{\text{within-view credit}} + \underbrace{(\bar{c}_v - \bar{c})}_{\text{pooled view baseline}} + \underbrace{\lambda(b_i - \bar{b})}_{\text{paired joint success}}. \tag{7}$$

Independent within-view normalization retains the first term; pooling introduces the second, and paired co-correctness adds the third. Since $b_i = 1$ already means both canonical answers are correct, no textual-agreement gate is needed. Figure 4 maps theory to implementation. Before reward scaling, token averaging, clipping, and KL, same-group mean centering retains factors $1 - 1/n$ and $1 - 2/n$ in the expected raw pointwise and pair-gradient components, respectively. Proposition G.3 also gives $\text{Var}(\widehat{q}_M) = q(1-q)/M$ for $\widehat{q}_M = M^{-1} \sum_i b_i$.

4 Experiments

We test three backbones across 14 benchmarks, then examine view reliability, reward components, hyperparameters, response budgets, run variation, and measured cost.

4.1 Experimental Setup

We train Qwen2.5-VL-3B/7B (Bai et al., 2025b) and Qwen3-VL-4B (Bai et al., 2025a) on 8,082 LLaVA-CoT math-and-logic examples (Xu et al., 2025) for one epoch with AdamW, learning rate 5×10^{-5}, and 8 H20 GPUs. All methods generate four responses per question; ConsistRoll allocates two per view with $\lambda = 0.5$. For the default $\pm 90^\circ$ rotation, validity checks admit 6,120 examples (75.7%) to paired training, while rejected examples retain ordinary GRPO. We compare Base, SFT, GRPO, Hint-GRPO, Share-GRPO, and PAPO under matched prompts, decoding, and LMMs-Eval evaluation (Zhang et al., 2025b), as shown in Appendix A.

Table 1 reports benchmark scores and domain means, with Math averaging nine math/logic scores, General averaging four top-level benchmarks, and Hallu. averaging three HallusionBench accuracies. Appendix D.1 retains component scores and exact averaging definitions, and code will be available.

Table 1: Benchmark scores and domain means across three backbones; shaded rows add ConsistRoll to the preceding optimizer. Complete scores and averaging definitions are in Appendix D.1.

	Mathematical and logical reasoning					General understanding					Hallu.
Method	Math Verse	Math Vision	Math Vista	Logic Vista	**Math**	GQA	MMB	MMStar	LLaVA -W	**Gen.**	**Avg.**
Qwen2.5-VL-3B	29.8	23.0	63.6	37.5	42.1	60.0	77.1	55.8	65.6	64.6	40.3
SFT	28.1	14.1	60.7	28.8	37.5	58.8	76.9	51.4	62.0	62.3	35.4
GRPO	31.4	21.1	64.6	36.8	43.1	60.5	77.9	57.0	64.4	65.0	40.8
+ ConsistRoll	32.4	24.7	64.4	36.6	43.5	60.4	77.1	56.0	64.6	64.5	41.3
Hint-GRPO	32.0	21.7	65.2	36.4	43.3	60.3	77.1	56.8	63.7	64.5	41.4
+ ConsistRoll	32.3	22.4	65.2	36.6	45.2	60.3	77.7	57.5	66.9	65.6	41.8
Qwen2.5-VL-7B	52.4	26.6	68.6	42.0	50.3	60.8	82.8	62.1	75.8	70.4	45.0
SFT	34.3	13.2	64.7	27.7	41.5	59.0	81.7	59.5	68.7	67.2	29.1
GRPO	53.6	27.6	68.0	42.0	50.6	61.0	83.2	61.5	68.3	68.5	46.3
+ ConsistRoll	54.2	28.9	69.2	45.8	52.8	60.8	83.2	63.2	76.1	70.8	47.3
Hint-GRPO	53.6	28.6	68.5	43.8	50.6	60.9	83.0	62.6	71.6	69.5	41.6
+ ConsistRoll	54.4	30.6	69.0	46.7	53.7	61.1	83.3	63.4	76.3	71.0	47.1
Qwen3-VL-4B	46.8	51.6	73.7	53.2	56.1	62.5	85.2	69.8	81.2	74.7	49.7
SFT	35.6	31.4	68.2	39.5	45.4	60.2	83.5	64.1	74.3	70.5	35.2
GRPO	47.2	52.0	74.3	53.5	57.1	62.8	85.5	69.0	75.4	73.2	50.7
+ ConsistRoll	48.5	54.3	75.6	55.4	58.6	62.6	85.7	70.8	81.9	75.3	51.8
Hint-GRPO	47.0	53.1	74.8	54.0	57.3	62.6	85.4	70.0	77.1	73.8	45.9
+ ConsistRoll	49.1	55.4	75.1	56.2	59.3	62.9	85.8	71.2	82.4	75.6	51.6

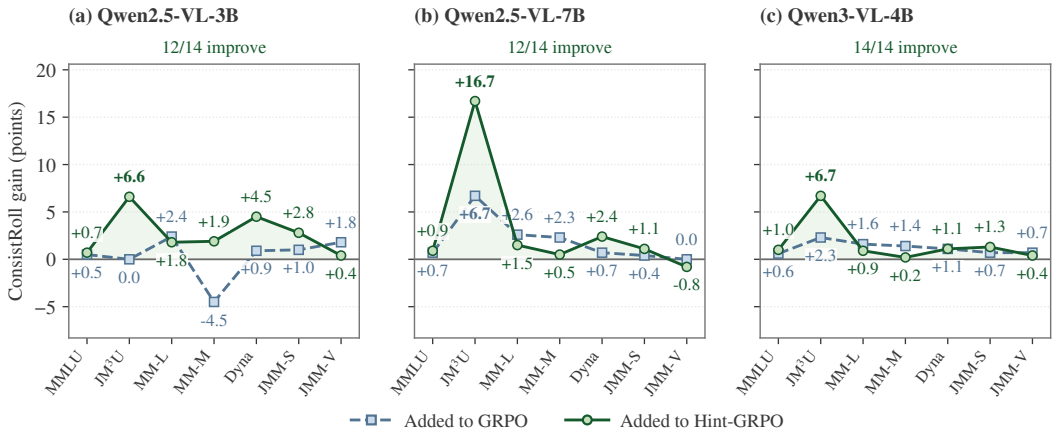


Figure 5: All 42 reasoning-score changes across three backbones and two optimizers, using shared axes and a zero-gain reference. Lines connect discrete diagnostics, absolute scores are in § D.1.

4.2 Task Performance

Gains across backbones and training methods. Table 1 compares six settings for each backbone. SFT lowers the Math average on all three models, while ConsistRoll improves it over both GRPO and Hint-GRPO. On 7B with GRPO, Math rises from 50.6 to 52.8 and General from 68.5 to 70.8. The individual scores locate the gains: LogicVista improves by 3.8 points and LLaVA-W by 7.8, while GQA and MMBench barely change. On LLaVA-W, much of the gain recovers the drop from the Base score of 75.8 to GRPO's 68.3. The 4B model improves in all three domains under both optimizers; 3B is less consistent, with a 0.5-point General decline under GRPO.

What the domain means conceal. Across seven reasoning diagnostics and six backbone–optimizer pairs, ConsistRoll improves 38 of the 42 displayed scores (Figure 5). Gains are broad on 4B, but not universal elsewhere: 3B with GRPO loses 4.5 points on MMStar-Math, and 7B with Hint-GRPO loses 0.8 on JMMMU-Vision. The largest gains occur on JM^{3}U-Math. Reporting these components alongside the domain means makes both the benefit and its task dependence explicit. Moreover, under matched 7B training, ConsistRoll leads on all five metrics in Table 2, exceeding Share-GRPO by 1.8 average points, mainly through LogicVista and LLaVA-W gains.

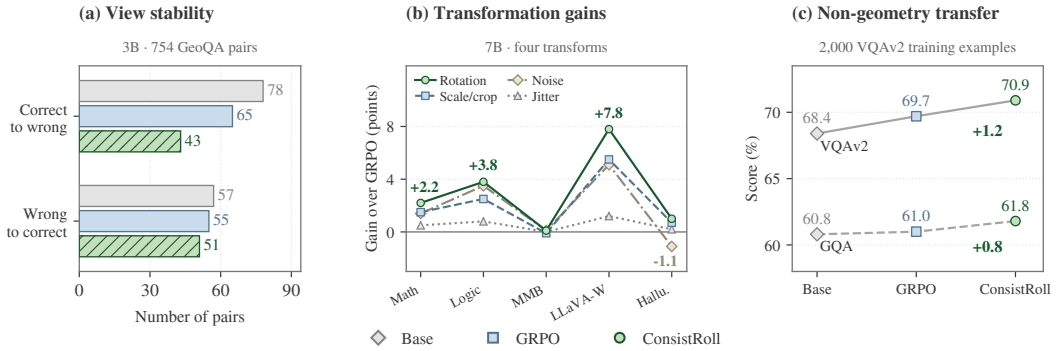


Figure 6: (a) Correctness changes on 754 GeoQA pairs. (b) Twenty gains across transformations and tasks. (c) VQAv2-trained transfer, with gains over GRPO. Protocols are in Appendices D.6 and D.4.

Table 2: Matched 7B comparison. Reported ± values reproduce two-seed mean±SD from Table 4(a).

Method	MathVerse	MathVista	LogicVista	LLaVA-W	Hallu.	Avg.
GRPO	53.6 ± 0.5	68.0 ± 0.4	42.0 ± 0.6	68.3 ± 0.7	46.3	55.64
Share-GRPO (Yao et al., 2025)	53.9	68.4	43.5	71.1	46.9	56.76
PAPO (Wang et al., 2026)	53.8	68.5	43.1	70.5	46.8	56.54
ConsistRoll	54.2 ± 0.3	69.2 ± 0.3	45.8 ± 0.4	76.1 ± 0.5	47.3	58.52

4.3 Component Ablations

Table 3 separates visual exposure, pooling, and paired credit at the same response budget. Splitting a same-view group does not help, while independent augmentation, pooling alone, and pairing alone gain 0.19, 0.44, and 1.09 points. Combining pooling and paired correctness gives the largest gain, 2.55, consistent with their distinct contributions in Eq. (7). The score is a matched-control composite, whose complete definition and results appear in the Appendix D.2.

Table 3: 7B ablations, four responses. GRPO uses four originals; other settings use a $2+2$ split, with identical views for the same-view control. Score and unrounded gains follow Appendix D.2.

Setting	**Norm.**	**Pair**	**Score**	**Δ**
GRPO	One group	–	54.91	–
Same-view split	Separate	–	54.80	−0.11
Independent DA	Per view	–	55.10	+0.19
Pooled DA	Pooled	–	55.35	+0.44
Pair only	Per view	✓	56.00	+1.09
ConsistRoll	Pooled	✓	57.45	+2.55

4.4 Reliability Across Views and Tasks

Fewer view-induced failures. On 754 GeoQA pairs (Chen et al., 2021), ConsistRoll reduces correct-to-wrong changes from 65 to 43, while wrong-to-correct changes decrease only slightly, from 55 to 51 (Figure 6 a). Overall correctness-state consistency rises from 84.08% to 87.53%. Showing both directions matters: consistency alone also counts pairs that remain wrong. These paired outcomes provide direct evidence of reduced view sensitivity, but do not estimate the per-question repeated-sampling probabilities in Section 2.

Transfer depends on the transformation. The transformation profiles show where transfer occurs (Figure 6 b). Rotation improves LogicVista and LLaVA-W by 3.8 and 7.8 points, versus 2.5 and 5.5 for scale/crop. Gaussian noise still helps reasoning but reduces HallusionBench by 1.1 points; color jitter has smaller effects throughout. MMBench changes by at most 0.1 points under every transformation. After training on 2,000 VQAv2 examples, ConsistRoll gains 1.2 points on VQAv2 and 0.8 on GQA over GRPO (Figure 6 c). These results support choosing transformations that preserve the evidence for the task at hand.

What the real cases reveal. The qualitative cases expose errors hidden by aggregate accuracy. In a tangent-geometry example, GRPO incorrectly applies a same-arc rule and returns 36°; ConsistRoll uses the perpendicular radius to obtain the reference answer, 27°. In an arithmetic case, both return 61, but GRPO includes the invalid step $28+9=35$, while ConsistRoll computes 37. These examples illustrate the difference between correct final answers and valid intermediate reasoning, they do not establish a general reasoning guarantee (Appendix Table 17).

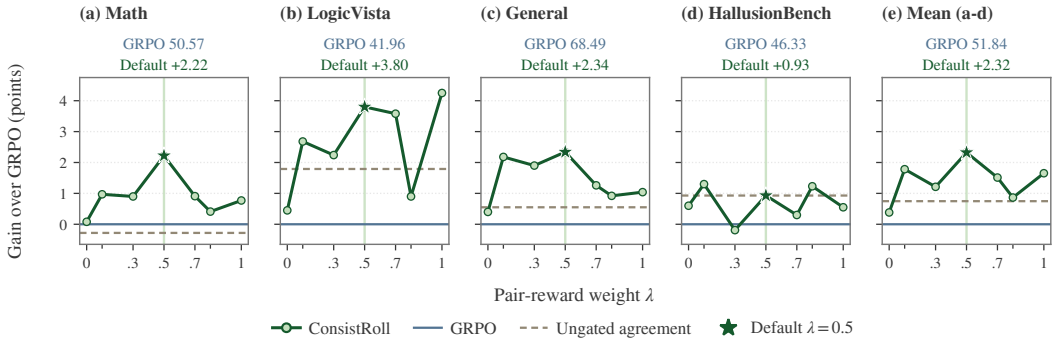


Figure 8: Seven pair-reward weights on 7B, $n = 4$. Stars mark $\lambda = 0.5$; dashed lines show ungated agreement. Panel (e) averages (a–d), counting LogicVista both separately and within Math, it differs from the Overall in Appendix Table 13.

Coupling beyond geometry. With the same verified scale/crop views, independent DA gains only 0.6/0.5/0.7 points on FactVQA ACC/sACC/F1, compared with ConsistRoll's 4.4/4.8/4.4, as shown in Figure 7, supporting paired outcomes over visual exposure alone in this OCR-heavy setting. Both methods retain the same answer-preservation checks, so the comparison isolates how valid views are rewarded.

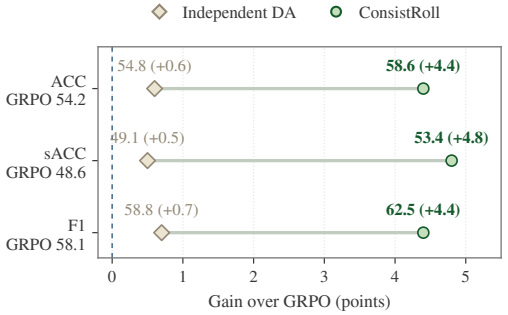


Figure 7: FactVQA with matched verified scale/crop. Labels show scores and gains over GRPO; full results are in Table 12 (c).

4.5 Hyperparameter Sensitivity

The default $\lambda = 0.5$ maximizes the mean of panels (a–d), reaching 54.16 versus 51.84 for GRPO and 52.59 for ungated agreement (Figure 8). It also maximizes the original three-domain Overall (56.96 vs. 55.13; Appendix Table 13). HallusionBench peaks at 0.1 and LogicVista at 1.0. A larger bonus is therefore not uniformly better. The full sweep also records a 1.36-point SciBench-Math decline at the default (Appendix D.5). We use 0.5 for its aggregate performance, with these task-specific trade-offs.

4.6 Response Budget, Validity, and Efficiency

Table 4: (a) Two-seed mean±SD (42 and 3407). (b) 7B with four or eight total responses

(a) Two-seed results across five benchmarks

Model	Method	Math Verse	Math Vista	Logic Vista	MMStar Math	LLaVA -W	Avg.
3B	GRPO	31.4 ± 0.6	64.6 ± 0.5	36.8 ± 0.8	63.0 ± 1.5	64.4 ± 0.9	52.0 ± 0.7
3B	ConsistRoll	32.3 ± 0.4	65.5 ± 0.3	37.0 ± 0.5	64.0 ± 0.7	67.2 ± 0.6	53.2 ± 0.4
7B	GRPO	53.6 ± 0.5	68.0 ± 0.4	42.0 ± 0.6	65.5 ± 1.2	68.3 ± 0.7	59.5 ± 0.6
7B	ConsistRoll	54.2 ± 0.3	69.2 ± 0.3	45.8 ± 0.4	67.8 ± 0.7	76.1 ± 0.5	62.6 ± 0.3

(b) Four versus eight generated responses

Method	n	Math	LogicVista	LLaVA-W	Hallu.	Avg.
GRPO	4	50.6	42.0	68.3	46.3	51.8
ConsistRoll	4	52.8	45.8	76.1	47.3	55.5
GRPO	8	51.9	43.5	70.1	46.9	53.1
ConsistRoll	8	**54.1**	**47.6**	**78.2**	**48.5**	**57.1**

Gains across seeds and response budgets. The repeated-run results improve on every displayed metric for both model sizes (Table 4a), with average gains of 1.2 points on 3B and 3.1 on 7B. The largest 7B changes are on LogicVista and LLaVA-W, and the reported standard deviations are smaller under ConsistRoll, although two seeds do not establish statistical significance. Doubling responses from four to eight improves all four budget-study metrics for both methods (Table 4b). ConsistRoll with four responses still exceeds GRPO with eight on each metric. This supports effective use of the response budget; it is not a direct measurement of the joint-success estimator variance.

Table 5: 7B validity and cost. (a) Paired-branch retention. (b) Audited precision and control-study scores. (c) Measurements on 8 H20 GPUs, $n = 4$, relative to GRPO.

(a) Paired-branch retention

Category	Total	Kept	Rate
Geometry	3,410	3,180	93.3%
Symbolic math	1,620	1,540	95.1%
Logic	1,290	980	76.0%
Chart/table	950	210	22.1%
OCR	812	210	25.9%
Total	8,082	6,120	75.7%

(b) Pair validity

Screen	Prec.	Math	Logic	Gen. cmp.	Score
Rules	97.3%	52.79	45.76	72.32	57.45
+ Verifier	99.2%	52.94	45.91	72.47	57.60

(c) System cost

Metric	GRPO	Ours	Change
Memory (GB/GPU)	68.4	69.1	+1.0%
Throughput (tokens/s)	1,240	1,215	−2.0%
Time/epoch (h)	14.2	14.5	+2.1%

The validity check matters. The image-aware verifier improves audited precision from 97.3% to 99.2%, with small gains across Math, LogicVista, and the general composite (Table 5a). The category counts explain its scope: most geometry and symbolic examples enter the paired branch, while most chart/table and OCR examples do not (panel b). Only 7 of 200 separately audited rejected questions were actually rotation-invariant. Rejected questions still train with ordinary GRPO, so the filter changes their training mode rather than removing them.

The extra view has a modest measured cost. Table 5c shows the full trade-off: memory increases by 1.0%, throughput decreases by 2.0%, and epoch time increases by 2.1%, despite unchanged autoregressive response counts. Together with the budget and seed results, this supports practical gains at small additional system cost. Appendices C and D.3 retain category counts, per-metric results, and measurement settings.

5 Related Work

Multimodal RLVR extends verifiable-reward training to visual reasoning (Shao et al., 2024; Liu et al., 2025c). Share-GRPO shares trajectories across question variants, while NoisyRollout uses visual augmentation to diversify rollouts (Yao et al., 2025; Liu et al., 2025a). Perception-aware methods such as PAPO strengthen dependence on visual evidence (Wang et al., 2026). Consistency learning encourages stable predictions under perturbations (Tarvainen & Valpola, 2017; Xie et al., 2020), but agreement alone can preserve a wrong answer. ConsistRoll brings these ideas into outcome-based RL: it couples verified answer-preserving views, rewards paired correctness, and normalizes both views together. This targets reliable success across equivalent visual presentations without requiring preference labels or matching full output distributions. Appendix B provides the full discussion.

6 Conclusion

A harmless change of view should preserve reliable success. Pointwise RLVR can assign the same mean reward to policies with very different weakest-view accuracy. ConsistRoll addresses this ambiguity through paired co-correctness: for a fixed verified pair, mean and joint success identify weakest-view reliability, and their combined target increases with either view's correctness while favoring balance at a fixed mean. Finite response-local rewards recover its raw gradient in expectation; pooled normalization and the clipped GRPO update then implement a practical surrogate. Experiments support improved reasoning and cross-view stability across backbones, transformations, and domains. These results establish a useful inductive bias under verified answer preservation, while leaving per-step accuracy guarantees and robustness to verifier errors outside the theoretical claims.

Limitations

ConsistRoll relies on answer-preserving transformations and reliable correctness feedback, so its benefits depend on the availability and quality of valid view pairs. Although our experiments cover multiple backbones and benchmarks, broader transformation families and open-ended tasks require further evaluation. Our theoretical results characterize fixed verified pairs, while guarantees for the full GRPO training dynamics remain an open question. (*We provide the AI Use Statement in §H*)

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:2511.21631*, 2025a.

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

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xu, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pp. 513–523, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 702–703, 2020.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. CHiP: Cross-modal hierarchical direct preference optimization for multimodal LLMs. In *International Conference on Learning Representations*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4848–4857, 2025a.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, 2026a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026b.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

Yann LeCun. Generalization and network design strategies. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, and L. Steels (eds.), *Connectionism in Perspective*. Elsevier, 1989.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2025b.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025c.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL*, 2024.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. In *International Conference on Learning Representations*, volume 2026, pp. 26868–26898, 2026.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6256–6268, 2020.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Technical Appendices and Supplements

This appendix provides reproducibility details, extended metric breakdowns, cross-view stability definitions, limitations, and full proofs. The main paper presents the core comparisons; this appendix retains full score breakdowns and the controls needed to inspect those conclusions.

A Detailed Experimental Settings

A.1 Training Setup

We use Qwen2.5-VL-3B/7B (Bai et al., 2025b) and Qwen3-VL-4B (Bai et al., 2025a). Unless a table states otherwise, every method is trained for one epoch on the same 8,082 LLaVA-CoT math-and-logic samples (Xu et al., 2025), with AdamW, learning rate 5×10^{-5}, total rollout count $n = 4$, and response format `<think>...</think><answer>...</answer>`. The default pair-reward weight is $\lambda = 0.5$. ConsistRoll allocates $M = n/2$ responses to each view; rejected transformation pairs use n original-view responses and standard GRPO. Figure 3 is generated from illustrative probabilities and exact equations by `scripts/plot_geometry_principle.py`; it contains no measured model outputs. The experimental summary plots are generated from the manuscript tables by `scripts/plot_experiment_figures.py`.

Transformation parameters. The primary rotation chooses $+90°$ or $-90°$. The additional ablations use Gaussian noise $\varepsilon \sim \mathcal{N}(0, 0.05^2)$, color jitter with brightness/contrast/saturation each varied within ± 0.2, or scaling followed by a random crop with scale factor 0.8. Each transformed image receives exactly one operation: transforms are not composed, and horizontal flipping is not used. Scale/crop enters the paired branch only when the image-aware verifier confirms that answer-relevant content remains visible.

A.2 Benchmarks and Metrics

Mathematical and logical evaluation uses MathVerse, MathVision, MathVista (Lu et al., 2023), LogicVista, MMLU-Pro Math, JM3U-Math, MMStar, DynaMath, SciBench-Math, and JMMMU-Pro. General-purpose evaluation uses GQA (Hudson & Manning, 2019), MMBench (Liu et al., 2025b), MMStar, and LLaVA-in-the-Wild. HallusionBench (Guan et al., 2024) supplies answer-level, figure-level, and question-pair accuracies. We also use GeoQA (Chen et al., 2021) for paired stability, VQAv2 and GQA for non-geometry transfer, and the OCR-heavy FactVQA evaluation for layout-robust factual extraction. FactVQA experiments use validity-filtered scale/crop; they do not assume that rotation or arbitrary document transformations preserve answers.

A.3 Baselines

Base is the unmodified checkpoint; **SFT** uses the same data and output format; **GRPO** uses standard pointwise verifiable rewards (Shao et al., 2024); and **Hint-GRPO** adds adaptive hints (Huang et al., 2025a). We additionally reproduce Share-GRPO (Yao et al., 2025) and PAPO (Wang et al., 2026) with the same Qwen2.5-VL-7B backbone, data, rollout budget, and evaluation.

A.4 Compute and System Cost

The fixed rollout count refers specifically to autoregressive generation. ConsistRoll still processes a second image, increasing visual-encoder/prefill work and KV-cache memory. On 8 H20 GPUs, peak memory rises from 68.4 to 69.1 GB/GPU, throughput changes from 1,240 to 1,215 tokens/s, and an epoch from 14.2 to 14.5 hours (Table 11). Thus, ConsistRoll keeps the generated-response count unchanged and adds approximately 1–2% system overhead in our implementation; it is not computationally free.

A.5 Definition of the Ungated Control

The "consistency-only" row in Table 13 uses only $\mathbf{1}[\hat{a}(o_0) = \hat{a}(o_1)]$ without correctness gating. It can reward a constant wrong answer and is therefore a deliberately degenerate control. The full

method instead uses $b_i = c_{0,i}c_{1,i}$ under the canonical verifier. This comparison establishes the need for correctness gating; it does not claim that κ independently improves over a both-correct gate.

B Detailed Related Work

We discuss multimodal RLVR, perception-aware optimization, consistency learning, and the inductive biases underlying reliable visual grounding.

B.1 RLVR for multimodal reasoning

Reinforcement learning has become a central post-training paradigm for large language models and MLLMs, from RLHF with learned reward models (Ziegler et al., 2019; Ouyang et al., 2022) to RLVR, where rule-based answer checkers provide scalable supervision. GRPO (Shao et al., 2024) removes the critic by normalizing rewards within a sampled group and has been widely adopted in reasoning models such as DeepSeek-R1 (Guo et al., 2025) and Kimi k1.5 (Kimi Team, 2025). Visual-RFT (Liu et al., 2025c), Vision-R1 (Huang et al., 2025b), Reason-RFT (Tan et al., 2025), R1-VL (Zhang et al., 2025a), VLM-R1 (Shen et al., 2025), and Perception-R1 (Xiao et al., 2025) extend this paradigm to visual reasoning through task verifiers, cold-start reasoning data, staged optimization, or perception rewards. Share-GRPO (Yao et al., 2025) expands the question space, shares discovered trajectories, and estimates advantages hierarchically across and within question variants to mitigate sparse rewards and advantage vanishing. NoisyRollout (Liu et al., 2025a) uses augmented visual inputs to increase rollout diversity. ConsistRoll instead couples verified answer-preserving *visual* views through a paired reward and shared normalization, targeting view-fragile correctness. At $\lambda = 0$, ConsistRoll reduces to pooled cross-view data augmentation; the full method adds paired co-correctness credit. Matched reproductions are reported in Table 2.

B.2 Perception-aware policy optimization

PAPO (Wang et al., 2026) introduces an implicit-perception KL objective using weakened visual evidence, while VPPO (Huang et al., 2026a) emphasizes token-level perception under masked inputs. These methods ask whether predictions depend on useful visual evidence. ConsistRoll asks a complementary question: whether the policy remains correct when the same evidence is presented in another verified valid form. Our experiments show that ConsistRoll outperforms reproduced PAPO under matched settings (Table 2); this comparison does not establish the effect of combining the two methods.

B.3 Augmentation, consistency, and invariance

Data augmentation is a standard mechanism for injecting invariances into vision systems, with classical and automated policies such as AutoAugment, RandAugment, and AugMix improving robustness by exposing models to label-preserving transformations (Cubuk et al., 2019; 2020; Hendrycks et al., 2020). Prediction consistency across perturbations has also been widely studied in semi-supervised and robust learning, including Mean Teacher, Virtual Adversarial Training, and UDA (Tarvainen & Valpola, 2017; Miyato et al., 2018; Xie et al., 2020). These methods typically impose consistency through supervised or semi-supervised probability-matching losses. ConsistRoll instead operates in a language-generation RL setting, where answers are sampled, parsed, and scored by verifiable rewards. A consistency-only objective admits degenerate policies that repeat the same wrong answer. We therefore use a correctness-gated pair reward. Under our canonical exact-match verifier this event is exactly paired co-correctness, so κ is not an independent empirical source of gain; the cross-view effect comes from optimizing the joint success event and normalizing the two views together.

B.4 Inductive biases, hallucination, and spurious grounding

The broader motivation of ConsistRoll follows the principle that models generalize better when their inductive biases match task symmetries. Classical architectures encode such biases structurally, including translation equivariance in CNNs (LeCun, 1989), permutation invariance in set and graph models (Zaheer et al., 2017), and symmetry constraints in geometric deep learning (Cohen & Welling, 2016; Weiler & Cesa, 2019; Bronstein et al., 2021). Modern MLLMs, however, are typically adapted

through post-training rather than architectural redesign. ConsistRoll therefore encodes a correctness-level invariance bias algorithmically through the RL objective; it does not require equality of full language-output distributions. This is particularly relevant for hallucination and spurious visual grounding, which remain central reliability failures in MLLMs (Li et al., 2023; Wang et al., 2023; Guan et al., 2024). Related preference-optimization methods reduce hallucination through human, model-generated, or synthetic preference pairs (Sun et al., 2024; Yu et al., 2024; Wang et al., 2024; Xie et al., 2024; Fu et al., 2025). In contrast, ConsistRoll targets tasks with verifiable answers and no preference annotations, penalizing view-specific shortcuts through paired, semantically-invariant visual transformations.

B.5 Method Feature Comparison

Table 6: Method-level comparison with representative RLVR-based MLLM reasoning methods. ConsistRoll changes the comparison unit by coupling verified answer-preserving visual views, applying a paired co-correctness reward, and keeping the generated-response count fixed.

Method	RLVR Training	Variant Source	Cross-Variant Credit	Answer-Preserving Visual Views	Paired Co-Correctness	Fixed Generated Responses	Modified Component(s)
GRPO (Shao et al., 2024)	✓	–	✗	✗	✗	✓	group-normalized outcome reward
Visual-RFT (Liu et al., 2025c)	✓	task verifier	✗	✗	✗	✓	task-specific verifiable rewards
Vision-R1 (Huang et al., 2025b)	✓	cold-start CoT	✗	✗	✗	✓	cold-start data and training schedule
Reason-RFT (Tan et al., 2025)	✓	curated CoT data	✗	✗	✗	✓	two-stage SFT–RL training
R1-VL (Zhang et al., 2025a)	✓	reasoning steps	✗	✗	✗	✓	step-wise reward design
VLM-R1 (Shen et al., 2025)	✓	task ground truth	✗	✗	✗	✓	rule-based rewards and data filtering
Perception-R1 (Xiao et al., 2025)	✓	visual annotations	✗	✗	✗	✓	perception reward with LLM judge
NoisyRollout (Liu et al., 2025a)	✓	clean/noisy images	△	△	✗	✓	rollout-level data augmentation
Share-GRPO (Yao et al., 2025)	✓	question variants	✓	✗	✗	△	trajectory and reward sharing
Hint-GRPO (Huang et al., 2025a)	✓	adaptive hints	✗	✗	✗	△	hint-based reward and text-bias calibration
ConsistRoll (ours)	✓	original / transformed visual views	✓	✓	✓	✓	rollout grouping and pair reward

Note: ✓ denotes explicit use, ✗ no use, and △ partial or indirect use. "Fixed generated responses" refers only to the autoregressive response count; ConsistRoll still incurs the small second-image prefill cost reported in Appendix A.4.

C Answer-Preserving Subset Construction

The paired branch is enabled only after two safeguards. First, a conservative question filter removes explicit view-dependent terms (left/right, top/bottom, front/behind, compass directions, clockwise/counterclockwise, and ordering language). Second, Qwen3.5-35B-A3B receives the original image, transformed image, question, and reference answer, and rejects the pair whenever the same answer is no longer visibly supported. Common failures include coordinate axes, oriented OCR labels, arrows, tables, and layout-dependent reading order.

Table 7 reports the final category breakdown. The pipeline retains 6,120 samples for coupling and routes 1,962 to standard GRPO. An independent GPT-5.5 audit is used to check transformation validity, while Qwen3.5-35B-A3B is the online image-aware verifier. The audited kept-set precision increases from 97.3% with the rule alone to 99.2% with the verifier. In a separate conservative-filter check, only 7 of 200 filtered-out samples (3.5%) were actually rotation-invariant, showing that the filter mostly sacrifices recall rather than admitting unsafe pairs.

Table 7: Construction and audit of the answer-preserving subset. Rejected samples are not discarded; they use standard single-view GRPO. Panel (b) retains the fixed control-study composite solely for a matched comparison of the two validity pipelines.

(a) Category breakdown

Category	Total	Kept	Rate	Removed	Main exclusion
Geometry	3,410	3,180	93.3%	230	Direction-sensitive question
Symbolic math	1,620	1,540	95.1%	80	Orientation-sensitive notation
Logical reasoning	1,290	980	76.0%	310	Layout/order dependence
Chart/table	950	210	22.1%	740	Axes, labels, layout
OCR-containing	812	210	25.9%	602	Readability/reading order
Total	**8,082**	**6,120**	**75.7%**	**1,962**	–

(b) Image-aware verifier audit and downstream score

Validity pipeline	Precision	Math	Logic	Gen. comp.	Composite
Rule-based filter	97.3%	52.79	45.76	72.32	57.45
Rule + Qwen3.5-35B-A3B	**99.2%**	52.94	45.91	72.47	57.60

D Extended Experimental Results

The tables below provide the full benchmark, control, seed, and transfer results underlying Section 4. Derived changes and duplicated summary plots are omitted.

D.1 Full Benchmark Results

Tables 8 and 9 retain every benchmark score for all three backbones, including base and SFT checkpoints. Math averages the first nine math/logic columns in Table 8, excluding JMMMU-Pro. General averages GQA, MMBench, MMStar, and LLaVA-W overall; its sub-scores are not counted again. Hallu. averages aAcc, fAcc, and qAcc. The summary preserves the reported domain averages, which may reflect greater precision than the displayed benchmark scores.

D.2 Mechanism Controls

Table 10 holds the total response budget fixed. Its control-study score averages Math, HallusionBench, and a general composite that includes GQA, MMBench, MMStar, and the four LLaVA-W columns. This inherited composite weights LLaVA-W more heavily than the main General average; it is used only for matched comparisons within this control study, not as an overall benchmark ranking. Table 3 summarizes the same source-reported gains before display rounding.

D.3 Run Variation, Budget, and Cost

Table 11 reports the two-seed results, the total-response budget comparison, and the hardware measurements. Its panels use their explicitly listed metric sets; these averages are not interchangeable with the three-domain summary.

D.4 Transformation and Domain Transfer

Table 12 reports all component scores, including the HallusionBench decrease under Gaussian noise. Transfer uses a task-appropriate valid transformation rather than assuming universal rotation invariance.

D.5 Full Hyperparameter Ablation

Table 13 summarizes the sweep; Tables 14 and 15 retain the component scores, including SciBench-Math. The duplicated default-method row has been removed: $\lambda = 0.5$ is ConsistRoll.

Table 8: Results on mathematical, logical, and part of general reasoning benchmarks across three backbones. All reported benchmark scores are retained; redundant changes relative to the base checkpoint are omitted.

Methods	Math-Verse	Math-Vision	Math-Vista	Logic-Vista	MMLU Math	JM^{3}U-Math	MMStar		Dyna-Math	Math Avg.	JMMMU Pro	
							Logical	Math			Std.	Vision
Qwen2.5-VL-3B	29.8	23.0	63.6	37.5	44.5	33.3	54.1	61.0	32.4	42.1	42.6	38.8
+ SFT	28.1	14.1	60.7	28.8	33.3	40.0	55.2	43.2	33.8	37.5	35.4	34.5
+ GRPO (Shao et al., 2024)	31.4	21.1	64.6	36.8	46.7	36.7	53.7	63.0	33.6	43.1	42.2	38.1
+ ConsistRoll (Ours)	32.4	24.7	64.4	36.6	47.2	36.7	56.1	58.5	34.5	43.5	43.2	39.9
+ Hint-GRPO (Huang et al., 2025a)	32.0	21.7	65.2	36.4	46.9	36.7	56.7	61.8	32.3	43.3	39.5	38.3
+ ConsistRoll (Ours)	32.3	22.4	65.2	36.6	47.6	43.3	58.5	63.7	36.8	45.2	42.3	38.7
Qwen2.5-VL-7B	52.4	26.6	68.6	42.0	51.0	36.7	61.0	66.3	47.9	50.3	46.1	45.2
+ SFT	34.3	13.2	64.7	27.7	37.8	33.3	58.9	59.8	43.9	41.5	41.7	42.7
+ GRPO (Shao et al., 2024)	53.6	27.6	68.0	42.0	52.5	33.3	60.8	65.5	51.9	50.6	46.3	46.4
+ ConsistRoll (Ours)	54.2	28.9	69.2	45.8	53.2	40.0	63.4	67.8	52.6	52.8	46.7	46.4
+ Hint-GRPO (Huang et al., 2025a)	53.6	28.6	68.5	43.8	52.8	30.0	63.7	66.4	48.0	50.6	46.2	46.3
+ ConsistRoll (Ours)	54.4	30.6	69.0	46.7	53.7	46.7	65.2	66.9	50.4	53.7	47.3	45.5
Qwen3-VL-4B	46.8	51.6	73.7	53.2	48.2	38.5	69.8	65.3	58.1	56.1	48.0	46.5
+ SFT	35.6	31.4	68.2	39.5	36.8	35.0	64.2	49.5	48.2	45.4	42.5	41.2
+ GRPO (Shao et al., 2024)	47.2	52.0	74.3	53.5	50.6	40.0	70.2	66.5	60.1	57.1	48.5	47.2
+ ConsistRoll (Ours)	48.5	54.3	75.6	55.4	51.2	42.3	71.8	67.9	61.2	58.6	49.2	47.9
+ Hint-GRPO (Huang et al., 2025a)	47.0	53.1	74.8	54.0	50.8	38.3	71.5	67.1	59.4	57.3	48.4	47.0
+ ConsistRoll (Ours)	49.1	55.4	75.1	56.2	51.8	45.0	72.4	67.3	60.5	59.3	49.7	47.4

Table 9: Results on general and hallucination benchmarks. General Avg. is the unweighted mean of the four top-level benchmarks (GQA, MMBench, MMStar, and LLaVA-W all); the three LLaVA-W sub-scores are displayed but excluded from the average.

Methods	GQA	MMB	MMStar	LLaVA in the wild				General Avg.	HallusionBench			Hallu. Avg.
			Avg.	all	complex	conv	detail		aAcc	fAcc	qAcc	
Qwen2.5-VL-3B	60.0	77.1	55.8	65.6	65.4	84.9	45.3	64.6	57.7	33.0	30.1	40.3
+ SFT	58.8	76.9	51.4	62.0	56.2	78.6	51.9	62.3	51.2	28.9	26.1	35.4
+ GRPO (Shao et al., 2024)	60.5	77.9	57.0	64.4	67.9	71.9	44.2	65.0	57.5	33.8	31.2	40.8
+ ConsistRoll (Ours)	60.4	77.1	56.0	64.6	61.3	86.8	46.5	64.5	58.6	33.2	32.0	41.3
+ Hint-GRPO (Huang et al., 2025a)	60.3	77.1	56.8	63.7	63.4	82.4	42.2	64.5	57.7	34.1	32.3	41.4
+ ConsistRoll (Ours)	60.3	77.7	57.5	66.9	67.6	81.8	48.6	65.6	58.4	34.7	32.3	41.8
Qwen2.5-VL-7B	60.8	82.8	62.1	75.8	85.9	83.0	51.9	70.4	62.7	36.4	36.0	45.0
+ SFT	59.0	81.7	59.5	68.7	72.0	81.4	48.1	67.2	44.5	24.6	18.2	29.1
+ GRPO (Shao et al., 2024)	61.0	83.2	61.5	68.3	72.7	78.8	49.3	68.5	63.4	38.4	37.1	46.3
+ ConsistRoll (Ours)	60.8	83.2	63.2	76.1	82.4	85.6	54.9	70.8	63.9	39.6	38.2	47.3
+ Hint-GRPO (Huang et al., 2025a)	60.9	83.0	62.6	71.6	87.9	73.9	44.1	69.5	55.6	32.5	36.6	41.6
+ ConsistRoll (Ours)	61.1	83.3	63.4	76.3	86.5	83.7	52.2	71.0	63.8	39.6	37.8	47.1
Qwen3-VL-4B (Instruct)	62.5	85.2	69.8	81.2	89.4	87.5	55.3	74.7	67.5	41.2	40.4	49.7
+ SFT	60.2	83.5	64.1	74.3	76.8	83.0	51.5	70.5	51.3	29.8	24.5	35.2
+ GRPO (Shao et al., 2024)	62.8	85.5	69.0	75.4	78.2	82.3	53.0	73.2	68.2	42.5	41.3	50.7
+ ConsistRoll (Ours)	62.6	85.7	70.8	81.9	88.3	89.2	58.4	75.3	68.9	43.8	42.6	51.8
+ Hint-GRPO (Huang et al., 2025a)	62.6	85.4	70.0	77.1	91.0	78.5	48.2	73.8	61.2	36.8	39.7	45.9
+ ConsistRoll (Ours)	62.9	85.8	71.2	82.4	90.2	88.0	56.1	75.6	68.8	44.0	42.1	51.6

D.6 Cross-View Stability Metrics

For each original–rotation pair, define correctness indicators $c_i^{(0)}$ and $c_i^{(1)}$. With N pairs,

$$
\begin{aligned}
D_{\mathrm{rot}} &= \sum_{i=1}^{N} c_i^{(0)}(1-c_i^{(1)}), \qquad I_{\mathrm{rot}} = \sum_{i=1}^{N}(1-c_i^{(0)})c_i^{(1)}, \\
F_{\mathrm{rot}} &= D_{\mathrm{rot}} + I_{\mathrm{rot}}, \qquad C_{\mathrm{rot}} = 1 - \frac{F_{\mathrm{rot}}}{N}, \\
\hat{p}_0 - \hat{p}_1 &= \frac{D_{\mathrm{rot}} - I_{\mathrm{rot}}}{N}, \qquad \frac{F_{\mathrm{rot}}}{N} \geq |\hat{p}_0 - \hat{p}_1|.
\end{aligned} \tag{8}
$$

D_{rot} counts correct-to-wrong degradation, I_{rot} wrong-to-correct improvement, and F_{rot} any correctness-state change. Here $\hat{p}_v = N^{-1}\sum_i c_i^{(v)}$ is a corpus-level accuracy, not the repeated-

Table 10: Factorial control of rollout allocation, augmentation, pair reward, and normalization on Qwen2.5-VL-7B. Every row generates exactly $n = 4$ total responses. The control-study Composite is the fixed mean of Math Avg., the seven-column general composite defined in Appendix D.2, and Hallu. Avg.; it is retained only for matched within-table comparison. Δ is relative to GRPO and is computed before display rounding.

Reward/group design	Rollout allocation	Composite	Δ
GRPO	4 original, one group	54.91	–
Same-view half-rollout GRPO	2+2 original, two groups	54.80	-0.11
Independent DA-GRPO	2 original + 2 transformed, view-local	55.10	$+0.19$
Pooled cross-view DA ($\lambda = 0$)	2 original + 2 transformed, pooled	55.35	$+0.44$
Pair reward, no shared normalization	2 original + 2 transformed, view-local	56.00	$+1.09$
ConsistRoll	2 original + 2 transformed, pooled	**57.45**	$+2.55$

Table 11: Reliability, rollout scaling, and measured cost. Panel (a) reports mean$\pm$standard deviation over two full training runs (seeds 42 and 3407). Panel (b) uses Qwen2.5-VL-7B and reports the mean of its four displayed metrics. Every n is the *total* number of generated responses.

(a) Per-metric repeated-run results

Backbone	Method	MathVerse	MathVista	LogicVista	MMStar-Math	LLaVA-W	Avg.
Qwen2.5-VL-3B	GRPO	31.4 ± 0.6	64.6 ± 0.5	36.8 ± 0.8	63.0 ± 1.5	64.4 ± 0.9	52.0 ± 0.7
	ConsistRoll	32.3 ± 0.4	65.5 ± 0.3	37.0 ± 0.5	64.0 ± 0.7	67.2 ± 0.6	53.2 ± 0.4
Qwen2.5-VL-7B	GRPO	53.6 ± 0.5	68.0 ± 0.4	42.0 ± 0.6	65.5 ± 1.2	68.3 ± 0.7	59.5 ± 0.6
	ConsistRoll	54.2 ± 0.3	69.2 ± 0.3	45.8 ± 0.4	67.8 ± 0.7	76.1 ± 0.5	62.6 ± 0.3

(b) Rollout-budget scaling

Method	n	Math Avg.	LogicVista	LLaVA-W	Hallu. Avg.	4-metric Avg.
GRPO	4	50.6	42.0	68.3	46.3	51.8
ConsistRoll	4	52.8	45.8	76.1	47.3	55.5
GRPO	8	51.9	43.5	70.1	46.9	53.1
ConsistRoll	8	**54.1**	**47.6**	**78.2**	**48.5**	**57.1**

(c) Measured system cost on 8$\times$H20, Qwen2.5-VL-7B, $n = 4$

Method	Peak memory/GPU	Throughput	Time/epoch
GRPO	68.4 GB	1,240 tokens/s	14.2 h
ConsistRoll	69.1 GB ($+1.0\%$)	1,215 tokens/s (-2.0%)	14.5 h ($+2.1\%$)

sampling probability $p_v(Z;\theta)$ of one fixed problem. Opposing flips can cancel in $\hat{p}_0 - \hat{p}_1$ while remaining visible in F_{rot}. Thus the table does not directly estimate the per-problem gaps or the nonlinear identification formula in Eq. (1). Relative to GRPO, ConsistRoll reduces D_{rot} from 65 to 43 and F_{rot} from 120 to 94, improving C_{rot} by 3.45 points.

Table 12: Transfer beyond the default rotation. Panel (a) uses Qwen2.5-VL-7B, the 8,082-sample training set, $n = 4$, and $\lambda = 0.5$; each image receives one verified transformation. Panel (b) trains on 2,000 VQAv2 samples with a mild perturbation. Panel (c) uses validity-filtered scale/crop on the OCR-heavy FactVQA benchmark, not arbitrary document rotation.

(a) Transformation families on the 8,082-sample training set

Method/transform	Math	Logic	MMB	LLaVA-W	Hallu.	Avg.
GRPO	50.6	42.0	83.1	68.3	46.3	58.0
ConsistRoll, rotation	52.8	45.8	83.2	76.1	47.3	61.0
ConsistRoll, Gaussian noise	52.0	45.5	83.2	73.4	45.2	59.9
ConsistRoll, color jitter	51.1	42.8	83.1	69.5	46.5	58.6
ConsistRoll, scale/crop	52.1	44.5	83.0	73.8	47.0	60.1

(b) Non-geometry training

Method	VQAv2	GQA
Base	68.4	60.8
GRPO	69.7	61.0
ConsistRoll	**70.9**	**61.8**

(c) OCR-heavy FactVQA

Method	ACC	sACC	F1	Avg.
GRPO	54.2	48.6	58.1	53.6
Independent DA-GRPO	54.8	49.1	58.8	54.2
ConsistRoll	**58.6**	**53.4**	**62.5**	**58.1**

Table 13: Ablation of the pair-reward weight λ on Qwen2.5-VL-7B. General Average is the mean of GQA, MMBench, MMStar, and LLaVA-W all; Overall is the mean of Math Average, General Average, and Hallu. Average. "Consistency-only" is an ungated agreement reward, and $\lambda = 0$ removes the pair reward. Δ is relative to GRPO.

Settings	Math Average	Logic-Vista	General Average	HallusionBench			Hallu. Average	Overall	Δ
				aAcc	fAcc	qAcc			
GRPO (Shao et al., 2024)	50.57	41.96	68.49	63.41	38.44	37.14	46.33	55.13	–
Consistency-only (ungated)	50.29	43.75	69.04	63.93	39.60	38.24	47.26	55.53	+0.40
Pooled cross-view DA ($\lambda = 0$)	50.65	42.41	68.89	64.04	38.73	38.02	46.93	55.49	+0.36
$\lambda = 0.1$	51.54	44.64	70.67	64.46	39.31	**39.12**	**47.63**	56.61	+1.48
$\lambda = 0.3$	51.47	44.20	70.39	63.20	37.86	37.36	46.14	56.00	+0.87
$\lambda = 0.5$ (default)	**52.79**	45.76	**70.83**	63.93	**39.60**	38.24	47.26	**56.96**	**+1.83**
$\lambda = 0.7$	51.48	45.54	69.75	63.51	38.15	38.24	46.63	55.95	+0.82
$\lambda = 0.8$	50.98	42.86	69.41	**64.77**	39.02	38.90	47.56	55.98	+0.85
$\lambda = 1.0$	51.34	**46.21**	69.53	64.04	39.02	37.58	46.88	55.91	+0.79

Table 14: Full λ ablation results on math and reasoning benchmarks from ConsistRoll. All results are reported on Qwen2.5-VL-7B.

Setting	Math-Verse	Math-Vision	Math-Vista	MMLU-ProMath	JM^{3}U-Math	Dyna-Math	Logic-Vista	MMStar		SciBench Math
								Logical	Math	
Qwen2.5-VL-7B	52.41	26.64	68.60	51.00	36.67	47.86	41.96	60.97	66.31	45.58
GRPO	53.58	27.63	68.00	52.48	33.33	51.92	41.96	60.80	65.47	49.66
$\lambda = 0.0$	52.10	27.63	68.10	53.11	30.00	52.08	42.41	64.22	66.22	51.70
$\lambda = 0.1$	52.17	28.95	68.10	52.11	36.67	50.88	44.64	62.51	67.80	51.02
$\lambda = 0.3$	51.56	29.93	68.80	54.33	33.33	51.88	44.20	62.82	66.41	46.94
$\lambda = 0.5$	54.25	28.89	69.20	53.22	40.00	52.61	45.76	63.37	67.78	48.30
$\lambda = 0.7$	50.37	28.95	69.00	52.18	36.67	51.76	45.54	63.62	65.26	48.30
$\lambda = 0.8$	50.98	28.62	69.10	50.85	36.67	50.90	42.86	61.88	67.00	46.94
$\lambda = 1.0$	51.36	30.92	67.90	52.48	33.33	52.02	46.21	61.66	66.14	45.58
Consistency-only	50.52	26.97	68.40	51.96	33.33	50.70	43.75	62.51	64.51	46.26

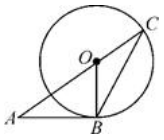
Table 15: Full λ ablation results on general and hallucination benchmarks from ConsistRoll. All results are reported on Qwen2.5-VL-7B. A-Acc, F-Acc, and Q-Acc denote the answer-level, figure-level, and question-pair accuracies of HallusionBench.

Setting	GQA	MMB Acc.	MMStar Avg.	LLaVA in the Wild				JMMMU-Pro		HallusionBench		
				All	Complex	Conv	Detail	Standard	Vision	A-Acc	F-Acc	Q-Acc
Qwen2.5-VL-7B	60.81	82.82	62.09	75.80	85.90	83.00	51.90	46.06	45.15	62.67	36.42	36.04
GRPO	61.04	83.16	61.45	68.30	72.70	78.80	49.30	46.29	46.44	63.41	38.44	37.14
$\lambda = 0.0$	60.88	82.82	62.14	69.70	75.00	79.20	49.60	43.56	43.86	64.04	38.73	38.02
$\lambda = 0.1$	60.88	83.76	63.33	74.70	76.80	78.10	51.50	46.36	44.24	64.46	39.31	39.12
$\lambda = 0.3$	60.76	82.91	62.47	75.40	86.30	81.80	51.10	45.61	45.38	63.20	37.86	37.36
$\lambda = 0.5$	60.83	83.16	63.24	76.10	82.40	85.60	54.90	46.67	46.44	63.93	39.60	38.24
$\lambda = 0.7$	60.94	83.33	62.43	72.30	75.50	85.70	51.10	46.82	46.89	63.51	38.15	38.24
$\lambda = 0.8$	60.81	83.16	62.57	71.10	79.80	74.60	53.00	47.65	46.74	64.77	39.02	38.90
$\lambda = 1.0$	61.11	82.99	62.60	71.40	75.20	81.40	53.30	46.59	47.12	64.04	39.02	37.58
Consistency-only	60.87	82.73	62.77	69.80	77.40	75.80	47.40	45.76	46.06	63.93	39.60	38.24

Table 16: Cross-view stability on 754 GeoQA pairs, Qwen2.5-VL-3B. D and I count opposite correctness changes; $F = D + I$ and $C = 100(1 - F/754)$.

Method	D_{rot}	I_{rot}	F_{rot}	C_{rot} (%)
Base	78	57	135	82.10
GRPO	65	55	120	84.08
ConsistRoll	43	51	94	87.53

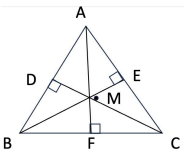
D.7 Qualitative Cases



Question: As shown in the figure, if angle A = 36.0, then angle C is equal to ()
Choices:
A:36° B:54° C:60° **D:27°**

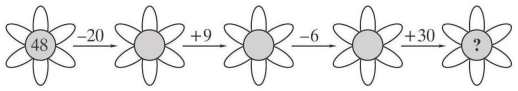
Key reasoning of GRPO:
1. Identify the given angle: angle A = 36°. 2. Note that angle A and angle C are circumference angles subtended by the same arc. 3. By the same-arc theorem, equal arcs subtend equal circumference angles. Therefore, angle C = angle A = 36°.

Key reasoning of ConsistRoll:
1. Identify the tangent property: line AB is tangent at B, so radius OB is perpendicular to AB and angle OBA = 90°. 2. Find angle AOB: angle AOB = 180° - 90° - 36° = 54°. 3. Since OB and OC are both radii, triangle OBC is isosceles, so angle OCB equals angle OBC. 4. By the exterior angle property, angle AOB equals twice angle OCB. Therefore, angle OCB = 27°.



Question: As shown in the figure, if BE ⊥ AC, CD ⊥ AB, AF ⊥ BC, then is M the incenter of triangle ABC?
Please answer yes or no.

Prediction of GRPO: Yes.
Prediction of ConsistRoll: No.



Question: Which number do you have to write in the last daisy?

Key reasoning of GRPO:
1. Start at 48. 2. Deduct 20 from 48: (48 - 20 = 28). 3. Adjust the weight and add 9: (28 + 9 = 35). 4. Subtract 6 from the daisy and add 30 to get 61: (35 - 6 + 30 = 61). The number is (61)

Key reasoning of ConsistRoll:
1. Start with the number 48. 2. Subtract 20 from 48: (48 - 20 = 28). 3. Add 9 to 28: (28 + 9 = 37). 4. Subtract 6 from 37: (37 - 6 = 31). 5. Add 30 to 31: (31 + 30 = 61). The number is (61)

Table 17: Qualitative case studies comparing GRPO and our proposed ConsistRoll.

E Limitations

Transformation validity. Answer preservation is necessary and is not universal. Rotations can invalidate axes, arrows, OCR, tables, spatial language, fine-grained layout, and GUI semantics; crop/scale can hide evidence. ConsistRoll must use a task-appropriate transformation plus an image-aware validity check, and uncertain samples must fall back to standard GRPO. The FactVQA result applies specifically to verified scale/crop and should not be generalized to arbitrary document transformations.

Verifier and parser dependence. Parser errors, ambiguous answer formats, or noisy labels affect both base correctness and the pair event. With a unique canonical answer, κ is mathematically redundant once both responses are correct. A semantic setting in which correctness and agreement genuinely diverge would require a broader verifier and is outside the evidence provided here.

Compute and rollout allocation. The method keeps autoregressive response generation fixed but processes two images, causing the measured prefill/memory overhead in Appendix A.4. It also halves responses per view. At small n, binary rewards can make each view-local sample all-correct or all-wrong; $n = 8$ reduces but does not eliminate this variance. Proposition G.5 proves that the all-to-all cross-view estimator has weakly smaller variance than index-wise pairing, with equality only in degenerate cases. We retain the index-wise construction because it supplies an immediate response-local partner; realizing the all-to-all gain inside GRPO requires a different local-reward dependence and remains future work.

Scope and statistical evidence. Most training data are math/logic examples, the largest general-purpose improvement is concentrated on LLaVA-W, and the 3B model exhibits smaller and more variable subset effects. We report two seeds rather than a large-scale significance study. These results support a useful inductive bias under verified transformations, not uniform improvement on every benchmark or model scale.

F Broader Impacts

ConsistRoll may make MLLMs less sensitive to superficial visual changes in diagrams, educational content, scientific figures, and hallucination-sensitive VQA. It requires no extra human annotation or

architectural modification. However, stable predictions can still be wrong when the verifier fails, the transformation changes semantics, or deployment shifts the input distribution. Cross-view stability should not be treated as sufficient evidence for correctness in medical, legal, surveillance, or safety-critical uses; those settings require domain validation, uncertainty estimation, human oversight, and appropriate privacy and safety controls.

G Complete Proofs and Supporting Analysis

The appendix follows the same three levels as Sections 2–3: the desired long-run behavior, its estimate from finitely many generated responses, and the practical GRPO update. The long-run reliability target is simply an expectation over responses the model could generate repeatedly; it has nothing to do with a human population. Section G.1 first identifies the missing statistic, then proves what this target prefers, and finally shows that finite response-local rewards recover its gradient on average. Section G.2 proves the K-view extension, Section G.4 analyzes finite sampled pairs and same-group centering, and Section G.5 marks the transition to normalized/clipped GRPO. These are deliberately separate claims: a desirable long-run target, an unbiased raw estimator, and a practical optimizer are not the same mathematical object.

All expectations are conditional on a fixed verified pair $Z = (x_0, x_1, y^*)$ and a deterministic parser $\hat{a}$ unless an outer expectation over $Z \sim \mathcal{D}_{\text{pair}}$ is shown; only policy responses are random after conditioning. The dataset-level target is $\mathcal{R}_\lambda(\theta) = \mathbb{E}_Z[\mathcal{J}_\lambda(p_0(Z;\theta), p_1(Z;\theta))]$. The product $p_0(Z)p_1(Z)$ is therefore formed *within the same pair* before averaging; in general it is not $\mathbb{E}_Z[p_0(Z)]\mathbb{E}_Z[p_1(Z)]$. This conditioning order is required for identifiability and is preserved by constructing paired rollouts from the same training example. The exact results use five assumptions:

1. **Answer preservation.** Both views have the same unique canonical target, and the parser applies the same correctness rule to both.
2. **Conditional independent sampling.** Given (x_0, x_1) and θ, responses are independent across views and indices. This is not an independence claim about the visual inputs.
3. **Differentiability for policy gradients.** π_θ is differentiable, its response support is locally independent of θ, and differentiation can be exchanged with summation or expectation. These conditions justify the usual log-derivative identity.
4. **Fixed verifier rewards during backpropagation.** Correctness and pair rewards are sampled coefficients in the policy loss; gradients do not pass through the parser or verifier.
5. **Regular outer expectation.** The verified-pair distribution is fixed with respect to θ. In a neighborhood of θ, the pair-level gradient exists almost surely and is dominated by an integrable envelope, so $\nabla_\theta \mathbb{E}_Z[\mathcal{J}_\lambda(Z;\theta)] = \mathbb{E}_Z[\nabla_\theta \mathcal{J}_\lambda(Z;\theta)]$.

Here n is an even total rollout budget, $M = n/2$ is the number of responses per view, and $b_i = c_{0,i}c_{1,i}$. The agreement indicator κ is absent because canonical exact match makes it redundant whenever $b_i = 1$.

G.1 Exact Two-View Reliability and Local Credit

For view $v \in \{0, 1\}$, abbreviate $\pi_v(o) = \pi_\theta(o \mid x_v)$, define $c_v(o) = \mathbf{1}[\hat{a}(o) = y^*]$, and let $p_v = \sum_o \pi_v(o)c_v(o)$. Draw $o_0 \sim \pi_0$ and $o_1 \sim \pi_1$ independently, set $b = c_0(o_0)c_1(o_1)$, and write $m = (p_0 + p_1)/2$ and $d = p_0 - p_1$.

Axiomatic uniqueness of the binary signal.

Lemma G.1 (Admissible binary interaction)**.** *Every nonnegative binary interaction that is zero whenever either response is incorrect equals $\alpha c_0 c_1$ for a unique $\alpha \geq 0$. Under a shared pair rule, the corresponding permutation-invariant pairwise-additive K-view interaction equals $\alpha e_2(\boldsymbol{c})$.*

Proof. Consider the cone of nonnegative correctness-only bonuses satisfying the error-null axiom,

$$\mathcal{H} := \left\{ h : \{0,1\}^2 \to \mathbb{R}_{\geq 0} \;\middle|\; h(c_0, c_1) = 0 \text{ whenever } c_0 c_1 = 0 \right\}. \tag{9}$$

The axiom says that cross-view credit may be added only to a response pair for which neither member is incorrect; it is deliberately stronger than symmetry or agreement. The values of any $h \in \mathcal{H}$ at $(0,0)$, $(0,1)$, and $(1,0)$ are forced to zero. Writing $\alpha = h(1,1) \geq 0$ fixes the remaining value, and therefore

$$\mathcal{H} = \operatorname{cone}\{(c_0, c_1) \mapsto c_0 c_1\} = \{\alpha c_0 c_1 : \alpha \geq 0\}. \tag{10}$$

Thus $c_0 c_1$ is the unique extreme ray—not the unique imaginable notion of consistency. The restricted claim is precisely that every nonzero, nonnegative, correctness-gated *binary* bonus is the same signal up to a scale absorbed into λ.

The characterization lifts without adding a new axiom. For $K \geq 2$, consider permutation-invariant pairwise-additive interactions

$$\mathfrak{H}_K := \left\{ H(\boldsymbol{c}) = \sum_{1 \leq u < v \leq K} h(c_u, c_v) : h \in \mathcal{H} \right\}. \tag{11}$$

Equation (10) implies $\mathfrak{H}_K = \{\alpha e_2(\boldsymbol{c}) : \alpha \geq 0\}$, where $e_2(\boldsymbol{c}) = \sum_{u<v} c_u c_v$. Thus the second elementary symmetric polynomial is the unique safe interaction under nonnegativity, error-null gating, pairwise additivity, and a shared interaction rule across view pairs. This is the exact scope of the axiomatic claim; dropping pairwise additivity permits higher-order gates such as $\prod_v c_v$, and dropping error-null gating permits agreement-based rewards. □

Joint success, imbalance, and worst-view reliability.

Proof of Proposition 2.1. Conditional independence and elementary two-variable algebra give the exact response-distribution geometry in one derivation:

$$\begin{aligned} q := \mathbb{E}[b] &= \sum_{o_0, o_1} \pi_0(o_0)\pi_1(o_1)c_0(o_0)c_1(o_1), \\ &= \left(\sum_{o_0} \pi_0 c_0\right)\left(\sum_{o_1} \pi_1 c_1\right) = p_0 p_1, \\ q = m^2 - \frac{d^2}{4}, &\qquad p_{\min} := \min(p_0, p_1) = m - \frac{|d|}{2} = m - \sqrt{m^2 - q}. \end{aligned} \tag{12}$$

The sums range over complete response sequences; integrals give the identical result for a continuous response space. The probability constraints $0 \leq p_0, p_1 \leq 1$ are equivalent to $0 \leq m \leq 1$ and $|d| \leq 2\min(m, 1-m)$. On every feasible fixed-m slice, $p_{\min}$ is therefore a strictly increasing function of q until the balanced endpoint $q = m^2$. Thus the pair event is not merely correlated with the robustness target: together with m, it exactly identifies worst-view correctness. Finally, p_0 and p_1 are the roots of $z^2 - 2mz + q = 0$, so $\{p_0, p_1\} = \{m - \sqrt{m^2 - q}, m + \sqrt{m^2 - q}\}$. This proves both exact recovery claims. □

Accuracy compatibility and strict balance preference.

Proof of Theorem 2.2. Substituting Eq. (12) into the raw local-credit objective yields all claims in the theorem:

$$\begin{aligned} \mathcal{J}_\lambda &= m + \frac{\lambda}{2} p_0 p_1 = m + \frac{\lambda}{2} m^2 - \frac{\lambda}{8} d^2, \\ \frac{\partial \mathcal{J}_\lambda}{\partial p_v} &= \frac{1}{2} + \frac{\lambda}{2} p_{1-v} \geq \frac{1}{2}, \\ \mathcal{J}_\lambda(m, m) - \mathcal{J}_\lambda(p_0, p_1) &= \frac{\lambda d^2}{8}, \\ (p_0 - p_1)\left(\frac{\partial \mathcal{J}_\lambda}{\partial p_0} - \frac{\partial \mathcal{J}_\lambda}{\partial p_1}\right) &= -\frac{\lambda}{2}(p_0 - p_1)^2 \leq 0. \end{aligned} \tag{13}$$

The positive coordinate derivatives establish Pareto monotonicity: lowering either view's correctness while holding the other fixed decreases the criterion. The last line is the differential criterion for Schur-concavity of a symmetric function; it is strict when $\lambda > 0$ and $p_0 \neq p_1$. The fixed-mean

expression is a strictly decreasing function of d^2, so it strictly favors any nontrivial mean-preserving balancing transfer and is uniquely maximized at $d = 0$ for $\lambda > 0$. When $\lambda = 0$, that preference disappears exactly. Over $[0, 1]^2$, coordinate-wise monotonicity gives the unique probability-level maximizer $(1, 1)$. A constant wrong-answer policy has $p_0 = p_1 = 0$, so both its pointwise and pair terms vanish. These statements concern the probability square; they do not claim that every point is realizable by a finite neural parameterization. □

Scope: population identity and objective preference. The identification result conditions on one problem before taking products. Corpus-level summaries do not in general recover average weakest-view success. For example, an equal mixture of pairs with probabilities $(1, 0)$ and $(0.5, 0.5)$ has $\mathbb{E}[m] = 0.5$, $\mathbb{E}[q] = 0.125$, and $\mathbb{E}[p_{\min}] = 0.25$. A homogeneous collection with probabilities $(0.5+\sqrt{0.125}, 0.5-\sqrt{0.125})$ has the same first two summaries but $\mathbb{E}[p_{\min}] = 0.5-\sqrt{0.125} \approx 0.1464$. Nor can a single binary pair observation be substituted into the identity as a calibrated confidence certificate.

Coordinate monotonicity and fixed-mean balance preference also do not imply that every objective increase preserves mean or weakest-view accuracy. With $\lambda = 0.5$, moving from $(0.9, 0.3)$ to $(0.59, 0.59)$ raises $\mathcal{J}_\lambda$ from 0.6675 to 0.677025 while reducing the mean from 0.60 to 0.59. Moving from $(0.6, 0.6)$ to $(0.9, 0.4)$ raises the target from 0.69 to 0.74 while reducing the weakest-view success from 0.60 to 0.40. Thus improved q identifies improved $p_{\min}$ *at fixed* m; the target expresses an inductive preference, not an unconditional training guarantee. Shared-parameter gradient interactions and the practical GRPO surrogate impose further limits.

Finite-batch local-gradient estimator.

Proof of Proposition 3.1. Let $\psi_{v,i} = \nabla_\theta \log \pi_\theta(o_{v,i} \mid x_v)$. Score-function regularity, stopped reward, and conditional independence imply, for every index i,

$$\mathbb{E}[c_{v,i}\psi_{v,i}] = \nabla_\theta p_v, \qquad \mathbb{E}[\psi_{v,i}] = 0, \qquad \mathbb{E}[b_i\psi_{v,i}] = p_{1-v}\nabla_\theta p_v. \tag{14}$$

For example, the last identity follows by factoring the partner term: $\mathbb{E}[c_{1-v,i}]\,\mathbb{E}[c_{v,i}\psi_{v,i}] = p_{1-v}\nabla p_v$. This uses independence of the sampled responses, while the two correctness probabilities remain functions of the same shared parameter vector.

With $r_{v,i} = c_{v,i} + \lambda b_i$ and $g_M = (2M)^{-1}\sum_{i,v} r_{v,i}\psi_{v,i}$, linearity of expectation and Eq. (14) give the exact finite-batch bridge

$$\mathbb{E}[g_M \mid Z] = \sum_{v=0}^{1}\left(\frac{1}{2} + \frac{\lambda}{2}p_{1-v}\right)\nabla_\theta p_v = \nabla_\theta\left[\frac{p_0 + p_1}{2} + \frac{\lambda}{2}p_0 p_1\right] = \nabla_\theta \mathcal{J}_\lambda(Z;\theta). \tag{15}$$

Taking the outer expectation and invoking Assumption 5 yields

$$\mathbb{E}_{Z,o}[g_M(Z)] = \mathbb{E}_Z[\nabla_\theta \mathcal{J}_\lambda(Z;\theta)] = \nabla_\theta \mathbb{E}_Z[\mathcal{J}_\lambda(Z;\theta)] = \nabla_\theta \mathcal{R}_\lambda(\theta). \tag{16}$$

This proves Proposition 3.1 for every $M \geq 1$, rather than only for a single response pair. If $p_0 > p_1$, the coefficient on ∇p_1 exceeds the coefficient on ∇p_0 by $\lambda(p_0 - p_1)/2$. The statement concerns scalar reciprocal weights, not the alignment or norm of the shared-parameter gradients. □

Why the coefficient is $\lambda/2$, not λ. The implemented estimator associates $r_{v,i}$ only with the score of response (v, i) and averages all $2M$ local terms, yielding $(\lambda/2)\nabla(p_0 p_1)$. In contrast, first forming the scalar group statistic $(r_0 + r_1)/2$ for one pair and differentiating its full expectation would let that statistic multiply both response scores. Because its expectation is $m + \lambda p_0 p_1$, that different global-cross-credit construction would have coefficient λ. It is not the response-local GRPO loss used by ConsistRoll.

G.2 Why the Long-Run Target Favors Reliable Views

Corollary 2.3 states the multi-view certificate compactly; this section gives the complete geometric theorem and derives its local-gradient result from first principles. Let $K \geq 2$, let $p_k \in [0, 1]$ be the canonical-answer correctness probability on view k, and write $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$ and $m = K^{-1}\mathbf{1}^\top \boldsymbol{p}$. Define the orthogonal projector $\mathbf{P}_\perp = \mathbf{I} - K^{-1}\mathbf{1}\mathbf{1}^\top$, so $\mathbf{P}_\perp \boldsymbol{p} = \boldsymbol{p} - m\mathbf{1}$. For each

indexed K-tuple of conditionally independent responses, let $b_{uv,i} = c_{u,i}c_{v,i}$ and assign the local reward

$$r_{v,i}^{(K)} = c_{v,i} + \frac{\lambda}{K-1}\sum_{u\neq v} b_{uv,i}, \qquad g_{K,M} = \frac{1}{KM}\sum_{i=1}^{M}\sum_{v=1}^{K} r_{v,i}^{(K)}\psi_{v,i}. \tag{17}$$

The factor $1/(K-1)$ keeps the maximum pair bonus per response equal to λ as the number of views changes.

Theorem G.2 (K-view local credit and strong Schur robustness)**.** *Let $e_2(\boldsymbol{p}) = \sum_{u<v} p_u p_v$ and $Q_K = \binom{K}{2}^{-1} e_2(\boldsymbol{p})$. Under the assumptions stated at the beginning of this appendix, for every $M \geq 1$,*

$$\begin{aligned}
\mathcal{J}_\lambda^{(K)} &:= m + \frac{\lambda}{K(K-1)} e_2(\boldsymbol{p}) = m + \frac{\lambda}{2}m^2 - \frac{\lambda}{2K(K-1)}\left\|\mathbf{P}_\perp \boldsymbol{p}\right\|_2^2, \\
\mathbb{E}[g_{K,M}] &= \nabla_\theta \mathcal{J}_\lambda^{(K)} = \sum_{v=1}^{K}\left[\frac{1}{K} + \frac{\lambda}{K(K-1)}\sum_{u\neq v} p_u\right]\nabla_\theta p_v, \\
\frac{\partial \mathcal{J}_\lambda^{(K)}}{\partial p_v} &= \frac{1}{K} + \frac{\lambda}{K(K-1)}\sum_{u\neq v} p_u \geq \frac{1}{K}, \\
\nabla_{\boldsymbol{p}}^2 \mathcal{J}_\lambda^{(K)} &= \frac{\lambda}{K(K-1)}(\mathbf{1}\mathbf{1}^\top - \mathbf{I}), \\
\boldsymbol{z}^\top \nabla_{\boldsymbol{p}}^2 \mathcal{J}_\lambda^{(K)} \boldsymbol{z} &= -\frac{\lambda}{K(K-1)}\|\boldsymbol{z}\|_2^2 \quad \textit{for every } \mathbf{1}^\top \boldsymbol{z} = 0, \\
p_{\min} &\geq m - \sqrt{\frac{K-1}{K}}\left\|\mathbf{P}_\perp \boldsymbol{p}\right\|_2 = m - (K-1)\sqrt{m^2 - Q_K}.
\end{aligned} \tag{18}$$

Moreover, if $p_u > p_v$ and $0 \leq \varepsilon \leq (p_u - p_v)/2$, the mean-preserving transfer $\boldsymbol{p}' = \boldsymbol{p} - \varepsilon \boldsymbol{e}_u + \varepsilon \boldsymbol{e}_v$ obeys

$$\mathcal{J}_\lambda^{(K)}(\boldsymbol{p}') - \mathcal{J}_\lambda^{(K)}(\boldsymbol{p}) = \frac{\lambda\varepsilon}{K(K-1)}(p_u - p_v - \varepsilon) \geq 0. \tag{19}$$

Consequently, $\mathcal{J}_\lambda^{(K)}$ is Pareto-monotone for $\lambda \geq 0$ and, for $\lambda > 0$, strongly concave on each fixed-mean slice and strictly Schur-concave. Its unique fixed-mean maximizer is $p_1 = \cdots = p_K = m$, and its unique maximizer on the probability cube is $\mathbf{1}$.

Proof. We prove the algebraic representation, finite-rollout gradient identity, monotonicity, spectral curvature, finite balancing step, and weak-view certificate in that order.

Step 1: symmetric-polynomial and projection identities. Expanding the second elementary symmetric polynomial gives

$$2\sum_{u<v} p_u p_v = \left(\sum_{k=1}^{K} p_k\right)^2 - \sum_{k=1}^{K} p_k^2. \tag{20}$$

Because $\sum_k p_k = Km$ and $\sum_k p_k^2 = Km^2 + \sum_k (p_k - m)^2$, division by $K(K-1)$ gives

$$Q_K = m^2 - \frac{1}{K(K-1)}\sum_{k=1}^{K}(p_k - m)^2. \tag{21}$$

Since $\mathbf{P}_\perp$ is symmetric and idempotent, $\|\mathbf{P}_\perp \boldsymbol{p}\|_2^2 = \sum_k (p_k - m)^2$. Substitution into $m + (\lambda/2)Q_K$ proves the first line of Eq. (18). This is an exact identity, not a local quadratic approximation. It is also Maclaurin's second-order inequality $Q_K \leq m^2$ with the equality case made explicit.

Step 2: finite-rollout gradient identity. Conditional independence and the score-function identity imply

$$\mathbb{E}[c_{v,i}\psi_{v,i}] = \nabla p_v, \qquad \mathbb{E}[b_{uv,i}\psi_{v,i}] = p_u \nabla p_v. \tag{22}$$

Taking the expectation of Eq. (17), collecting the two directed contributions of each unordered pair, and using symmetry of the double sum yield

$$\begin{aligned}\mathbb{E}[g_{K,M}] &= \frac{1}{K}\sum_v \nabla p_v + \frac{\lambda}{K(K-1)}\sum_v \sum_{u\neq v} p_u \nabla p_v \\ &= \nabla_\theta \left[\frac{1}{K}\sum_v p_v + \frac{\lambda}{K(K-1)}\sum_{u<v} p_u p_v\right] = \nabla_\theta \left[m + \frac{\lambda}{2} Q_K\right].\end{aligned} \tag{23}$$

This proves the response-local gradient bridge without differentiating through the sampled reward. It also proves the coefficient expansion in Eq. (18).

Step 3: Pareto monotonicity and the global optimum. Differentiating the pairwise form gives

$$\partial_{p_v} \mathcal{J}_\lambda^{(K)} = \frac{1}{K} + \frac{\lambda}{K(K-1)}\sum_{u\neq v} p_u \geq \frac{1}{K} > 0. \tag{24}$$

Thus increasing any coordinate while holding the others fixed strictly increases the objective, including at $\lambda = 0$. Repeated coordinate increases show that $\mathbf{1}$ is the unique maximizer on $[0,1]^K$.

Step 4: Hessian spectrum and strong fixed-mean concavity. The linear mean term has zero Hessian. Every monomial $p_u p_v$ contributes one to entries (u,v) and (v,u) and zero to the diagonal, hence

$$\nabla^2_{\boldsymbol{p}} \mathcal{J}_\lambda^{(K)} = \frac{\lambda}{K(K-1)}(\mathbf{1}\mathbf{1}^\top - \mathbf{I}). \tag{25}$$

The vector $\mathbf{1}$ is an eigenvector with eigenvalue λ/K, whereas every $\boldsymbol{z} \perp \mathbf{1}$ is an eigenvector with eigenvalue $-\lambda/[K(K-1)]$. Therefore the objective is not globally concave when $\lambda > 0$; rather, it is $\lambda/[K(K-1)]$-strongly concave on the tangent space of every fixed-mean hyperplane. This distinction is why the theorem states a fixed-mean robustness result instead of an unrestricted concavity claim.

Step 5: finite balancing transfers and Schur order. For two coordinates $u \neq v$, all common terms cancel:

$$\partial_{p_u} \mathcal{J}_\lambda^{(K)} - \partial_{p_v} \mathcal{J}_\lambda^{(K)} = \frac{\lambda}{K(K-1)}(p_v - p_u). \tag{26}$$

Multiplication by $p_u - p_v$ yields the differential Schur criterion

$$(p_u - p_v)\Big(\partial_{p_u} \mathcal{J}_\lambda^{(K)} - \partial_{p_v} \mathcal{J}_\lambda^{(K)}\Big) = -\frac{\lambda}{K(K-1)}(p_u - p_v)^2 \leq 0. \tag{27}$$

For completeness, the finite version follows without invoking the criterion. The transfer keeps $\sum_k p_k$ fixed, so only e_2 changes. Using $2e_2(\boldsymbol{p}) = (\sum_k p_k)^2 - \sum_k p_k^2$,

$$e_2(\boldsymbol{p}') - e_2(\boldsymbol{p}) = -\frac{1}{2}\big[(p_u - \varepsilon)^2 + (p_v + \varepsilon)^2 - p_u^2 - p_v^2\big] = \varepsilon(p_u - p_v - \varepsilon). \tag{28}$$

Multiplying by $\lambda/[K(K-1)]$ proves Eq. (19). For $\lambda > 0$ and $0 < \varepsilon \leq (p_u - p_v)/2$, the gain is strict. Finite compositions of such T-transforms generate the majorization order, establishing strict Schur-concavity. The projection identity gives the same conclusion geometrically: on a fixed-mean hyperplane, maximizing $\mathcal{J}$ is exactly minimizing $\|\mathbf{P}_\perp \boldsymbol{p}\|_2^2$, whose unique minimum is zero at $\boldsymbol{p} = m\mathbf{1}$.

Step 6: sharp weakest-view certificate. Choose an index v_* with $p_{v_*} = p_{\min}$ and put $a = m - p_{\min} \geq 0$. Let $z_k = p_k - m$, so $z_{v_*} = -a$ and $\sum_{k\neq v_*} z_k = a$. Cauchy–Schwarz gives

$$a^2 = \left(\sum_{k\neq v_*} z_k\right)^2 \leq (K-1)\sum_{k\neq v_*} z_k^2, \qquad \|\boldsymbol{z}\|_2^2 \geq a^2 + \frac{a^2}{K-1} = \frac{K}{K-1}a^2. \tag{29}$$

Rearranging yields $p_{\min} \geq m - \sqrt{(K-1)/K}\,\|\mathbf{P}_\perp \boldsymbol{p}\|_2$. Equation (21) implies $\|\mathbf{P}_\perp \boldsymbol{p}\|_2^2 = K(K-1)(m^2 - Q_K)$, giving the second form in the theorem. Equality in Cauchy–Schwarz holds exactly when all z_k, $k \neq v_*$, are equal, so the certificate is sharp. □

For $K = 2$, $Q_2 = p_0 p_1$ and $\|\mathbf{P}_\perp \boldsymbol{p}\|_2^2 = (p_0 - p_1)^2/2$, so Theorem G.2 reduces exactly to Eq. (2). The weak-view certificate is then an equality, recovering Eq. (1). For $K > 2$, (m, Q_K) controls a sharp lower bound and the complete dispersion, but does not identify the minimum coordinate itself; no stronger identification claim is made.

G.3 Why Agreement Alone Is Insufficient

Let $\kappa(o_0, o_1) = \mathbf{1}[\hat{a}(o_0) = \hat{a}(o_1)]$. A policy that always emits one fixed answer obtains $\kappa = 1$ on every pair, including when that answer differs from y^*. Agreement alone therefore admits a wrong constant-answer optimum. ConsistRoll instead uses $b = c_0 c_1$, which is one only when both responses are correct. Under a unique canonical target and exact-match parsing, $b\kappa = b$; multiplying by κ would leave every reward unchanged. Thus κ is useful for defining an ungated control, but it is neither an independent bonus nor an independent source of gain in the canonical setting.

G.4 How Finite Rollouts Change the Signal

Let $n = 2M$ and $\mathcal{G} = \{(v, i) : v \in \{0, 1\}, i \in [M]\}$. Each b_i is Bernoulli with parameter $q = p_0 p_1$, and distinct index pairs are independent.

Proposition G.3 (Exact finite-group effects)**.** *Let $\widehat{q}_M = M^{-1} \sum_i b_i$ and $g^\circ_M = (2M)^{-1} \sum_{v,i} (r_{v,i} - \bar{r}_\mathcal{G})\psi_{v,i}$. Then, before random standardization and clipping,*

$$\mathbb{E}[\widehat{q}_M] = q, \qquad \mathrm{Var}(\widehat{q}_M) = \frac{q(1-q)}{M}, \qquad \mathbb{E}[g^\circ_M] = \left(1 - \frac{1}{n}\right) \nabla m + \frac{\lambda}{2} \left(1 - \frac{2}{n}\right) \nabla q. \tag{30}$$

Proposition G.4 (Exact moments and Bernstein concentration)**.** *For $\widehat{q}_M = M^{-1} \sum_i b_i$ and every $\varepsilon > 0$,*

$$\mathbb{E}[\widehat{q}_M] = q, \qquad \mathrm{Var}(\widehat{q}_M) = \frac{q(1-q)}{M}, \qquad \Pr(|\widehat{q}_M - q| \geq \varepsilon) \leq 2 \exp\left[-\frac{M\varepsilon^2}{2q(1-q) + 2\varepsilon/3} \right]. \tag{31}$$

Proof. Linearity gives $\mathbb{E}[\widehat{q}_M] = M^{-1} \sum_i q = q$. Independence across indices eliminates all off-diagonal covariance terms, hence

$$\mathrm{Var}(\widehat{q}_M) = \frac{1}{M^2} \sum_{i=1}^{M} \mathrm{Var}(b_i) = \frac{q(1-q)}{M}. \tag{32}$$

For concentration, put $Z_i = b_i - q$. Then the Z_i are independent and centered, $|Z_i| \leq 1$, and $\sum_i \mathbb{E}[Z_i^2] = Mq(1-q)$. Applying the two-sided Bernstein inequality to $\sum_i Z_i$ at threshold $M\varepsilon$ gives

$$\Pr\left(\left| \sum_i Z_i \right| \geq M\varepsilon \right) \leq 2 \exp\left[-\frac{M^2\varepsilon^2}{2Mq(1-q) + 2M\varepsilon/3} \right], \tag{33}$$

which simplifies to Eq. (31). No asymptotic approximation is used. □

This proves the first two claims of Proposition G.3. Holding n fixed leaves only $M = n/2$ indexed pair events; with confidence $1 - \delta$, it is sufficient that $M \geq [2q(1-q) + 2\varepsilon/3]\varepsilon^{-2} \log(2/\delta)$ to achieve error at most ε. Index pairing requires no correspondence between reasoning traces.

The same $2M$ Bernoulli outcomes also admit an all-pairs estimator. Its exact variance clarifies that the index-wise construction is simple and unbiased, but not statistically efficient among estimators built from the same correctness bits.

Proposition G.5 (Indexed versus all-pairs estimation)**.** *Let*

$$\widehat{q}_M^{\mathrm{idx}} = \frac{1}{M}\sum_{i=1}^{M} c_{0,i}c_{1,i}, \qquad \widehat{q}_M^{\mathrm{all}} = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} c_{0,i}c_{1,j} = \bar{c}_0\bar{c}_1. \tag{34}$$

Both estimators are unbiased for $q = p_0p_1$, and

$$\begin{aligned} \mathrm{Var}(\widehat{q}_M^{\mathrm{all}}) &= \frac{p_0^2p_1(1-p_1) + p_1^2p_0(1-p_0)}{M} + \frac{p_0(1-p_0)p_1(1-p_1)}{M^2}, \\ \mathrm{Var}(\widehat{q}_M^{\mathrm{idx}}) - \mathrm{Var}(\widehat{q}_M^{\mathrm{all}}) &= \frac{q(1-p_0)(1-p_1)(M-1)}{M^2} \geq 0. \end{aligned} \tag{35}$$

The inequality is strict for $M > 1$ whenever $0 < p_0, p_1 < 1$.

Proof. Because the two view-wise samples are independent, $\bar{c}_0$ and $\bar{c}_1$ are independent and $\mathbb{E}[\bar{c}_v] = p_v$. Therefore $\mathbb{E}[\widehat{q}_M^{\mathrm{all}}] = p_0p_1 = q$. Moreover,

$$\mathbb{E}[\bar{c}_v^2] = p_v^2 + \frac{p_v(1-p_v)}{M}. \tag{36}$$

Using independence again, $\mathrm{Var}(\bar{c}_0\bar{c}_1) = \mathbb{E}[\bar{c}_0^2]\mathbb{E}[\bar{c}_1^2] - p_0^2p_1^2$; expanding the product proves the first line of Eq. (35). Subtracting it from $q(1-q)/M$ and factoring gives

$$\frac{p_0p_1}{M}(1-p_0)(1-p_1) - \frac{p_0p_1}{M^2}(1-p_0)(1-p_1) = \frac{q(1-p_0)(1-p_1)(M-1)}{M^2}, \tag{37}$$

which proves the second line and its equality cases. □

ConsistRoll uses $\widehat{q}_M^{\mathrm{idx}}$ because the same index immediately supplies a response-local pair reward. Proposition G.5 does not change the estimator analyzed in the main method; it proves the variance tradeoff behind the all-to-all extension noted in the limitations. An all-pairs local-credit rule would require a correspondingly different within-group reward dependence, even though it can target the same expected polynomial.

For reference, independently augmented GRPO normalizes each view separately, whereas ConsistRoll uses the pair reward and a shared group:

$$A_{v,i}^{\mathrm{DA}} = \frac{c_{v,i} - \bar{c}_v}{s_v + \delta}, \qquad b_i = c_{0,i}c_{1,i}, \quad r_{v,i} = c_{v,i} + \lambda b_i, \quad A_{v,i}^{\mathrm{CR}} = \frac{r_{v,i} - \bar{r}_{\mathcal{G}}}{s_{\mathcal{G}} + \delta}. \tag{38}$$

Here $\bar{c}_v = M^{-1}\sum_i c_{v,i}$, s_v is the corresponding within-view standard deviation, and $\bar{r}_{\mathcal{G}}, s_{\mathcal{G}}$ are computed over all n responses.

Define $\bar{c} = (\bar{c}_0 + \bar{c}_1)/2$ and $\bar{b} = M^{-1}\sum_i b_i$. Since the same b_i is assigned to both views, the pooled mean and centered numerator satisfy one exact sample identity:

$$\begin{aligned} \bar{r}_{\mathcal{G}} &= \frac{1}{2M}\sum_{v,i}(c_{v,i} + \lambda b_i) = \bar{c} + \lambda\bar{b}, \qquad \mathbb{E}[\bar{r}_{\mathcal{G}}] = m + \lambda p_0p_1, \\ r_{v,i} - \bar{r}_{\mathcal{G}} &= (c_{v,i} - \bar{c}_v) + (\bar{c}_v - \bar{c}) + \lambda(b_i - \bar{b}). \end{aligned} \tag{39}$$

The second line follows by adding and subtracting $\bar{c}_v$. It separates within-view correctness, the pooled view baseline, and paired co-correctness. The three components are generally correlated, and all are divided by the same random $s_{\mathcal{G}} + \delta$; consequently the shared variance is not the sum of three independent component variances. With $\lambda = 0$, joint normalization retains the pooled mean and scale, whereas separate normalization reduces to independent DA. At fixed normalization, comparing $\lambda > 0$ against $\lambda = 0$ isolates paired co-correctness, not κ.

The coefficient difference in Eq. (39) is intentional. The expectation of the group-mean statistic has pair coefficient λ, but the raw response-local policy objective in Eq. (15) has coefficient $\lambda/2$. Treating $\nabla\mathbb{E}[\bar{r}_{\mathcal{G}}]$ as the implemented policy gradient would incorrectly give each response the entire group's cross-credit.

G.5 Where the Exact Theory Ends

After constructing $\widehat{A}_{v,i}$, ConsistRoll uses the ordinary token-level GRPO surrogate. Put $h_{v,i,t} = (x_v, o_{v,i,<t})$ and define

$$\rho_{v,i,t}(\theta) := \frac{\pi_\theta(o_{v,i,t} \mid h_{v,i,t})}{\pi_{\theta_{\text{old}}}(o_{v,i,t} \mid h_{v,i,t})}, \qquad \mathcal{K}_{v,i,t}(\theta) := D_{\text{KL}}(\pi_\theta(\cdot \mid h_{v,i,t}) \,\|\, \pi_{\text{ref}}(\cdot \mid h_{v,i,t})). \tag{40}$$

Let $P_{\text{old}}(\mathcal{O}_\mathcal{G} \mid Z) = \prod_{(v,i)\in\mathcal{G}} \pi_{\theta_{\text{old}}}(o_{v,i} \mid x_v)$ and $\ell_\epsilon(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A\}$. The complete maximization objective is

$$\begin{aligned} \mathcal{J}_{\text{CR-GRPO}}(\theta) &= \mathbb{E}_{\substack{Z\sim\mathcal{D}_{\text{pair}},\\ \mathcal{O}_\mathcal{G}\sim P_{\text{old}}(\cdot|Z)}} \left[\frac{1}{2M} \sum_{(v,i)\in\mathcal{G}} \frac{1}{|o_{v,i}|} \sum_{t=1}^{|o_{v,i}|} \left(\ell_\epsilon(\rho_{v,i,t}, \widehat{A}_{v,i}) - \beta \mathcal{K}_{v,i,t} \right) \right], \\ \widehat{A}_{v,i} &= \frac{c_{v,i} + \lambda c_{0,i} c_{1,i} - \bar{r}_\mathcal{G}}{s_\mathcal{G} + \delta}. \end{aligned} \tag{41}$$

This objective leaves the model architecture and optimizer unchanged, but it is not algebraically identical to the raw on-policy response-level score in Proposition 3.1: advantages are standardized on the same sampled group, response scores are replaced by token-averaged old-policy ratios, those ratios are clipped, and the KL regularizer adds a separate gradient.

Proposition G.6 (First-order form of the practical surrogate)**.** *Assume $\epsilon > 0$, hold $\widehat{A}_{v,i}$ fixed during differentiation, and evaluate Eq.* (41) *at $\theta = \theta_{\text{old}}$. With $\bar{\psi}^{\text{tok}}_{v,i} = |o_{v,i}|^{-1} \sum_t \nabla_\theta \log \pi_\theta(o_{v,i,t} \mid h_{v,i,t})|_{\theta_{\text{old}}}$,*

$$\nabla_\theta \mathcal{J}_{\text{CR-GRPO}}|_{\theta_{\text{old}}} = \mathbb{E}\left[\frac{1}{2M} \sum_{(v,i)\in\mathcal{G}} \widehat{A}_{v,i} \bar{\psi}^{\text{tok}}_{v,i} - \frac{\beta}{2M} \sum_{(v,i)\in\mathcal{G}} \frac{1}{|o_{v,i}|} \sum_t \nabla_\theta \mathcal{K}_{v,i,t}|_{\theta_{\text{old}}} \right]. \tag{42}$$

Proof. At $\theta = \theta_{\text{old}}$, every ratio equals one, which lies in the interior of $[1-\epsilon, 1+\epsilon]$. Hence the unclipped and clipped branches agree on a neighborhood and $\partial_\rho \ell_\epsilon(1, A) = A$. Moreover, $\nabla_\theta \rho_{v,i,t}|_{\theta_{\text{old}}} = \nabla_\theta \log \pi_\theta(o_{v,i,t} \mid h_{v,i,t})|_{\theta_{\text{old}}}$. Differentiating the frozen sampling expectation, averaging tokens, and retaining the KL derivative gives Eq. (42). □

The proposition identifies the optimizer's local direction exactly; it does *not* make the standardized advantage an unbiased estimator of $\nabla\mathcal{R}_\lambda$.

The boundary already appears with mean centering alone, and the pair-dependent reward makes the distortion nonuniform. Omit the shared scale and define $\bar{\psi} = n^{-1}\sum_{v,i} \psi_{v,i}$ and $g^\circ_M = n^{-1}\sum_{v,i}(r_{v,i} - \bar{r}_\mathcal{G})\psi_{v,i}$. Because $\bar{r}_\mathcal{G} = \bar{c} + \lambda\bar{b}$, exact sample algebra first gives

$$\begin{aligned} g^\circ_M &= g_M - \bar{r}_\mathcal{G}\bar{\psi}, \\ \mathbb{E}[\bar{c}\,\bar{\psi}] &= \frac{1}{n^2}\sum_{i,v} \mathbb{E}[c_{v,i}\psi_{v,i}] = \frac{1}{n}\nabla m, \\ \mathbb{E}[\bar{b}\,\bar{\psi}] &= \frac{1}{Mn}\sum_{i,v} \mathbb{E}[b_i\psi_{v,i}] = \frac{1}{n}\nabla q. \end{aligned} \tag{43}$$

For the second and third lines, expanding the products creates double sums over reward and score indices. Every cross-index term vanishes because the relevant score has zero mean and is independent of the other index. The surviving same-response terms use $\mathbb{E}[c_{v,i}\psi_{v,i}] = \nabla p_v$; the surviving same-pair terms use $\sum_v \mathbb{E}[b_i\psi_{v,i}] = \nabla q$. Since $n = 2M$, the displayed coefficients follow exactly.

Combining Eq. (43) with $\mathbb{E}[g_M] = \nabla m + (\lambda/2)\nabla q$ proves

$$\mathbb{E}[g^\circ_M] = \left(1 - \frac{1}{n}\right)\nabla m + \frac{\lambda}{2}\left(1 - \frac{2}{n}\right)\nabla q. \tag{44}$$

This proves the third claim of Proposition G.3. The same-sample baseline attenuates the task component by $1 - 1/n$ and the pair component by the smaller factor $1 - 2/n$; for $n = 2$ it removes

pair credit exactly. At $\lambda = 0$, Eq. (44) reduces to $(1 - 1/n)\nabla m$, the simpler counterexample obtained from pointwise correctness. Thus mean centering alone is already not an unbiased estimator of the raw objective, although its bias vanishes as $n \to \infty$ under this fixed-pair model.

The attenuation is caused by reusing the current pair inside its own baseline, not by centering per se. This can be isolated exactly. For $M \geq 2$, define the leave-one-pair-out statistic

$$\bar{r}_{-i} := \frac{1}{2(M-1)} \sum_{j \neq i}^{} \sum_{u=0}^{1} r_{u,j}, \qquad g_M^{\mathrm{LOPO}} := \frac{1}{2M} \sum_{i=1}^{M} \sum_{v=0}^{1} (r_{v,i} - \bar{r}_{-i}) \psi_{v,i}. \tag{45}$$

For every (v, i), $\bar{r}_{-i}$ depends only on response pairs with index $j \neq i$ and is therefore independent of $\psi_{v,i}$. Since $\mathbb{E}[\psi_{v,i}] = 0$,

$$\mathbb{E}[\bar{r}_{-i} \psi_{v,i}] = \mathbb{E}[\bar{r}_{-i}] \, \mathbb{E}[\psi_{v,i}] = 0, \qquad \mathbb{E}[g_M^{\mathrm{LOPO}}] = \mathbb{E}[g_M] = \nabla_\theta \mathcal{J}_\lambda. \tag{46}$$

This cross-fitted construction is an analytical control, not the implemented normalized group advantage. Its role is to pinpoint the source of the factors in Eq. (44): the ordinary pooled mean contains $r_{v,i}$ and its partner reward, creating one same-response and two same-pair reward–score contractions.

The actual method additionally divides by the same random $s_{\mathcal{G}} + \delta$. That scale is correlated with rewards and scores and cannot generally leave the expectation; token averaging, off-policy ratios, and clipping add further nonlinearities. Thus Eq. (44) characterizes the *centered numerator*, not the standardized clipped update.

Equations (15), (44), (39), and (41) delimit raw estimation, centered estimation, sample algebra, and the practical surrogate, respectively. No unbiasedness or convergence guarantee crosses the last boundary; equal expected correctness also does not imply equal response distributions, identical traces, or zero per-instance flips.

H AI USE STATEMENT

We disclose the use of GPT-6-Astra to assist with language refinement and manuscript preparation, including literature and code analysis, mathematical formulation and proof drafting, interpretation of supplied results, and visualization preparation. The measured training traces and evaluation results were supplied by the authors, and AI-assisted analysis did not replace model training or evaluation. The authors retain responsibility for our ConsistRoll method and implementation, verification of mathematical arguments and empirical claims, and all research decisions, as well as the accuracy, originality, and integrity of the final manuscript and accompanying artifacts.