Reward Fusion is a Pairwise Contrast Problem: Why Ceiling Contact Can Harm Group Relative RL

Zhenyu Zhang
Amazon
zhenyuzh@amazon.com

Dingyang Chen
Amazon
dingerwa@amazon.com

Yuan Ling
Amazon
yualing@amazon.com

Xinyang Shen
Amazon
xinyans@amazon.com

Li Dong
Amazon
ldonga@amazon.com

Abstract

Reward fusion combines an exact outcome signal with a dense but noisy process signal before group relative policy optimization (GRPO). For a fixed rollout group, we show that the part of the local GRPO update due to reward is an exact sum of pairwise policy score differences weighted by normalized reward gaps. This identity separates two roles for process reward: it can order incorrect trajectories, but it can also erase or reverse the contrast between correct and incorrect trajectories. We distinguish *ceiling collision*, where an incorrect reward ties the correct reward, from *strict reversal*, where it exceeds it. Under stochastic monotonicity, both probabilities are nondecreasing in the quality of an incorrect solution; positive risk also requires process channel mass near the relevant threshold. A positive outcome margin gives a distribution free guarantee that every contrast across outcome classes has the correct sign. Empirically, process score tails track observed contact rates to about 1.5 percentage points. On hard math at 3B, fusion with a margin reaches $+2.8$ percentage points over outcome reward alone; at 0.5B, CLIP degrades by -8.9 percentage points while methods with a margin remain stable. A separate intervention shows that sufficiently frequent strict reversals can cause comparable degradation.

1 Introduction

Reasoning models increasingly combine a sparse, exact outcome reward with a dense process score during post training [Shao et al., 2024, DeepSeek-AI, 2025]. The usual view treats fusion as an estimation problem: combine the two signals so that the resulting scalar reward better reflects solution quality. That view is incomplete for GRPO because rewards are centered and standardized within each rollout group. The optimization step depends not only on which reward is larger, but also on the gaps among rewards in the same group.

This paper develops a pairwise view of that dependence. For any fixed group, the local update due to reward can be written exactly as a sum over trajectory pairs. Each term is a policy score difference multiplied by the corresponding reward gap. The sign of the gap determines which trajectory that pair favors, while its magnitude determines the strength of the comparison. General monotone transformations preserve reward order but need not preserve this update.

The decomposition reveals two distinct jobs for process reward. Among incorrect trajectories, process score differences can provide supervision that outcome reward alone cannot. Across outcome classes, the same process score can weaken a comparison, erase it through a ceiling collision, or reverse

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Transitioning from Pre-Training to Post-Training.

it. This leads to a hierarchical design principle: *outcome sets the tier, while process reward orders trajectories within the tier*. We formalize the principle through the smallest outcome contrast

$$\gamma_\varphi := r_+ - \sup_s h(s),$$

where r_+ is the reward for a correct trajectory and $h(s)$ is the reward for an incorrect trajectory with process score s. A positive γ_φ is exactly the condition for a uniform positive outcome margin.

The experiments are organized around this decomposition. First, pairwise diagnostics measure how often a learned process reward reaches the correct reward and how much normalized contrast remains. Second, end to end training compares fusion with and without a margin at 3B and 0.5B. Third, a controlled intervention injects strict reversals to test the more severe wrong sign regime. The results support the pairwise account, but they do not establish that a margin alone guarantees final accuracy.

Scope. The formal results describe the local update at the rollout policy, before repeated PPO epochs and excluding the KL term. They characterize the coefficients attached to pairwise policy score differences rather than the final sum of interacting gradients. The main analysis uses a constant reward r_+ for correct trajectories and a nondecreasing branch $h(s)$ for incorrect trajectories. If the correct branch also depends on process score, robust separation instead requires

$$\inf_s h_+(s) > \sup_s h(s).$$

This extension changes the margin definition but not the pairwise logic.

Our contributions are:

1. an exact pairwise decomposition of the local reward driven GRPO update;
2. a distributional characterization of ceiling contact and an exact characterization of robust outcome separation;
3. empirical evidence connecting pairwise diagnostics to distinct training regimes, with explicit limits on the causal interpretation.

2 Pairwise Theory

2.1 Local GRPO Update

Given prompt x, GRPO samples $G \geq 2$ trajectories $Y_{1:G} := \{Y_1, \ldots, Y_G\}$. Trajectory Y_i receives reward R_i. Define

$$\bar{R} := \frac{1}{G}\sum_{j=1}^{G} R_j, \qquad s_R^2 := \frac{1}{G}\sum_{j=1}^{G}(R_j - \bar{R})^2, \qquad A_i := \frac{R_i - \bar{R}}{s_R + \epsilon}, \tag{1}$$

where $\epsilon > 0$ is a numerical stabilizer. For rollout $Y_i = (y_{i,1}, \ldots, y_{i,T_i})$, let

$$\ell_i(\theta) := \frac{1}{T_i}\sum_{t=1}^{T_i} \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}), \qquad g_i := \nabla_\theta \ell_i(\theta).$$

At the rollout policy, the PPO importance ratio equals one. The local update attributable to reward is

$$U_R := \frac{1}{G}\sum_{i=1}^{G} A_i g_i. \tag{2}$$

The KL term is omitted because it does not depend on the fusion rule.

Theorem 1 (Pairwise decomposition)**.** *For any fixed rollout group,*

$$U_R = \frac{1}{G^2(s_R + \epsilon)} \sum_{1 \leq i < j \leq G} (R_i - R_j)(g_i - g_j). \tag{3}$$

The proof is in Appendix A.1. For pair (i, j), the sign of $R_i - R_j$ determines the direction of its contribution and $|R_i - R_j|/(s_R + \epsilon)$ determines its normalized strength. Thus reward ranking matters, but ranking alone does not determine the update.

Corollary 2 (Reward transformations)**.** *For a fixed group:*

1. *any strictly increasing transformation preserves the ordering of rewards and advantages;*
2. *a positive affine transformation $R_i' = aR_i + b$, with $a > 0$ and $b \in \mathbb{R}$, changes U_R only by the positive group factor*
$$\frac{a(s_R + \epsilon)}{as_R + \epsilon};$$
3. *a nonlinear increasing transformation can change relative pairwise gaps and therefore can change the update direction.*

When $\epsilon = 0$, positive affine transformations leave the local update unchanged. This is the precise sense in which group standardization removes calibration scale. It does not make GRPO invariant to general monotone recalibration, because such a transformation can change the relative weights assigned to different pairs. The derivation is in Appendix A.1.

2.2 Two Roles for Process Reward

Each trajectory has an exact outcome $O_i \in \{0, 1\}$ and a process score $S_i \in \mathcal{S} \subseteq \mathbb{R}$. We study fusion rules of the form

$$R_i = \varphi(O_i, S_i), \qquad \varphi(1, s) = r_+, \qquad \varphi(0, s) = h(s), \tag{4}$$

where h is nondecreasing. Define

$$\Delta_\varphi(s) := r_+ - h(s), \qquad \gamma_\varphi := \inf_{s \in \mathcal{S}} \Delta_\varphi(s) = r_+ - \sup_{s \in \mathcal{S}} h(s). \tag{5}$$

Let $\mathcal{C} := \{i : O_i = 1\}$ and $\mathcal{W} := \{i : O_i = 0\}$. Substituting (4) into Theorem 1 gives

$$\begin{aligned} U_\varphi = \kappa_\varphi \Bigg[& \sum_{c \in \mathcal{C}} \sum_{w \in \mathcal{W}} \Delta_\varphi(S_w)(g_c - g_w) \\ & + \sum_{\{j,k\} \subset \mathcal{W}} \big(h(S_j) - h(S_k)\big)(g_j - g_k) \Bigg], \end{aligned} \tag{6}$$

where $\kappa_\varphi := [G^2(s_R + \epsilon)]^{-1} > 0$. The first line contains comparisons across outcome classes. Its coefficient is positive, zero, or negative when the comparison agrees with the outcome, collides, or reverses. The second line contains comparisons among incorrect trajectories. Its usefulness depends on whether the process score meaningfully orders those trajectories. These are separate channel properties. Outcome reward alone makes the second line vanish, so it provides no preference inside a group in which every trajectory is incorrect. A process reward can activate those comparisons. In a mixed group, however, the first line is the direct outcome supervision that distinguishes correct from incorrect trajectories. Fusion is useful when it adds information through the second line without destroying the signs in the first. This is why within class ranking quality and across class separation must be evaluated separately.

2.3 Ceiling Contact and Robust Separation

An incorrect score creates a *ceiling collision* when $h(s) = r_+$ and a *strict reversal* when $h(s) > r_+$. Their union, $h(s) \geq r_+$, is *weak contact*. To study how contact varies with solution quality, let $Q \in [0, 1]$ denote latent quality, with $Q = 1$ representing a fully correct solution. The exact outcome satisfies $O = 1$ if and only if $Q = 1$. For incorrect trajectories, write

$$S \mid Q = q \sim \nu_q, \qquad \bar{F}_q(t) := \Pr(S \geq t \mid Q = q).$$

We assume first order stochastic monotonicity: for every threshold t and qualities $q \leq q' < 1$,

$$\bar{F}_q(t) \leq \bar{F}_{q'}(t) \tag{A1}$$

A1 permits equality and therefore does not require the process channel to be informative.

Table 1: Pointwise fusion rules and their smallest outcome contrast. For CLIP, equality assumes the score domain reaches the clipping ceiling. For NASH, $\sup_s u(s) = 1$ and $0 < \lambda\widehat{\sigma} < 1$.

Fusion	r_+	$h(s)$	γ_φ
ORM	1	0	1
CLIP	1	$\mathrm{clip}(s, 0, 1)$	0
Advantage Margin	$1 + \delta$	$\mathrm{clip}(s, 0, 1)$	$\geq \delta$
NASH	1	$(1 - \lambda\widehat{\sigma})u(s)$	$\lambda\widehat{\sigma}$

Proposition 3 (Quality and contact)**.** *Let*

$$\mathcal{E}_{\geq} := \{s : h(s) \geq r_+\}, \qquad \mathcal{E}_{>} := \{s : h(s) > r_+\},$$

and define

$$\pi_{\geq}(q) := \Pr(S \in \mathcal{E}_{\geq} \mid Q = q), \qquad \pi_{>}(q) := \Pr(S \in \mathcal{E}_{>} \mid Q = q).$$

Under A1, both probabilities are nondecreasing in q.

When $\mathcal{E}_{\geq} = [\tau_{\geq}, \infty)$, Proposition 3 reduces to $\pi_{\geq}(q) = \bar{F}_q(\tau_{\geq})$. Thus A1 orders contact risk but does not imply positive risk. Positive risk also requires process score mass beyond the decision threshold. For CLIP, defined by $r_+ = 1$ and $h(s) = \mathrm{clip}(s, 0, 1)$, weak contact is the event $S \geq 1$ and strict reversal is impossible.

Corollary 4 (Gaussian channel example)**.** *For CLIP, suppose $S = Q + Z$ with $Z \sim \mathcal{N}(0, \sigma^2)$. If Φ is the standard normal cumulative distribution function, then*

$$\pi_{\geq}(q) = \Phi\left(\frac{q - 1}{\sigma}\right) \longrightarrow \frac{1}{2} \quad \text{as } q \uparrow 1, \qquad \pi_{>}(q) = 0.$$

For fixed $q < 1$, $\pi_{\geq}(q) \to 0$ as $\sigma \downarrow 0$.

The two limits separate quality from noise. For a fixed noisy channel, incorrect solutions close to the correctness boundary can contact the CLIP ceiling often. For a fixed incorrect quality, better estimation can instead drive contact probability toward zero. The channel controls whether contact occurs; a margin supplies a stronger guarantee that does not depend on that channel.

Theorem 5 (Hierarchy and margin)**.** *For the fusion rule in* (4)*:*

1. *strict reversal is impossible for every admissible score if and only if $\gamma_\varphi \geq 0$;*
2. *a uniform positive outcome margin exists if and only if $\gamma_\varphi > 0$;*
3. *if h attains its supremum, weak contact is impossible if and only if $\gamma_\varphi > 0$.*

Without attainment, weak contact is impossible exactly when $h(s) < r_+$ for every admissible s; this pointwise condition can hold even when $\gamma_\varphi = 0$. Proofs and the boundary case are in Appendix A.2. The outcome reward model baseline, denoted ORM, uses $\varphi(o, s) = o$. Advantage Margin uses $r_+ = 1 + \delta$ and $h(s) = \mathrm{clip}(s, 0, 1)$, giving $\gamma_\varphi \geq \delta > 0$. NASH and RANK implement the same hierarchy in different ways; their definitions are in Appendix B.

A normalized hierarchy. Let $u : \mathcal{S} \to [0, 1]$ be any nondecreasing readout. The family

$$\varphi_\delta(o, s) := o + (1 - o)(1 - \delta)u(s), \qquad \delta \in (0, 1),$$

assigns reward one to correct trajectories and confines every incorrect reward to $[0, 1 - \delta]$. It therefore has $\gamma_\varphi \geq \delta$ while preserving weak process score order among incorrect trajectories. The parameter δ reserves part of the reward range for outcome separation. A larger value strengthens the worst case cross class margin but compresses the range available to distinguish incorrect trajectories. The theorem identifies why $\delta > 0$ is structurally robust; it does not determine the value that maximizes final accuracy.

A positive margin guarantees the sign of every coefficient across outcome classes in (6). It does not guarantee final accuracy because process noise still affects comparisons among incorrect trajectories, normalized gap magnitudes, and optimization dynamics. Conversely, an operator that touches the boundary can still help when its within class signal outweighs the lost outcome contrast.

Proposition 6 (Group contact)**.** *Suppose a mixed group contains $k := |\mathcal{C}|$ correct and $m := |\mathcal{W}|$ incorrect trajectories. For $j \in \mathcal{W}$, let*

$$I_j := \mathbf{1}\{h(S_j) \geq r_+\}, \qquad \pi_j(Q_{\mathcal{W}}) := \Pr(I_j = 1 \mid Q_{\mathcal{W}}),$$

where $Q_{\mathcal{W}} := (Q_j)_{j \in \mathcal{W}}$. If incorrect process scores are conditionally independent given their qualities, then

$$\Pr(\textit{at least one weak contact} \mid Q_{\mathcal{W}}) = 1 - \prod_{j \in \mathcal{W}} \big(1 - \pi_j(Q_{\mathcal{W}})\big).$$

Without independence, the expected number of weakly contacting pairs across outcome classes is

$$\mathbb{E}\left[k \sum_{j \in \mathcal{W}} I_j \,\middle|\, Q_{\mathcal{W}} \right] = k \sum_{j \in \mathcal{W}} \pi_j(Q_{\mathcal{W}}).$$

One contacting incorrect trajectory affects its comparison with every correct trajectory. The number of incorrect trajectories controls whether a group contains any contact, while the number of correct trajectories controls how many pairwise coefficients are affected. The proof is in Appendix A.3.

2.4 Diagnostics Implied by the Theory

The decomposition suggests three complementary diagnostics.

1. **Contrast across outcome classes.** Pair and group contact rates together with $A_c - A_w$ show whether exact outcome supervision is preserved.
2. **Ordering within the incorrect class.** A rank metric measures whether the second line of (6) contains useful information.
3. **Exposure through group composition.** Mixed groups and groups in which every trajectory is incorrect determine how often either role is active.

Together they distinguish an uninformative process channel, limited group exposure, and a fusion map that erases useful outcome contrast.

3 Experiments

We use Qwen2.5-Math-PRM-7B as the process reward model and exact answer match as the outcome [Zhang et al., 2025]. The experiments ask three questions suggested by (6). How often does process reward weaken contrast across outcome classes? Can process reward help when that contrast is protected? What happens when the contrast is deliberately reversed? Full protocols and secondary results are in Appendix C.

3.1 Pairwise Contrast Diagnostics

We score rollout groups from Qwen2.5-{0.5B, 7B}-Instruct on GSM8K and MATH-500 [Qwen et al., 2024, Cobbe et al., 2021, Hendrycks et al., 2021]. Diagnostics use mixed groups containing at least one correct and one incorrect trajectory. The pair rate is the fraction of pairs across outcome classes with $R_w \geq R_c$. The group rate is the fraction of mixed groups containing at least one such pair. The margin is the mean normalized gap $A_c - A_w$ across outcome classes.

CLIP contacts the ceiling in 12.6% to 14.7% of pairs and in 31.9% to 57.6% of mixed groups. Its normalized outcome margin is also smaller. Empirical process score tails at the CLIP threshold agree with observed pair contact rates to about 1.5 percentage points on average. This is a descriptive check of Proposition 3, not an independent prediction, because both quantities use the same score distribution.

The monotonicity statistic is a finite sample proxy for A1 rather than proof of latent stochastic ordering. A paired permutation test with 10,000 resamples over problem level contact indicators gives $p \leq 10^{-4}$ at the reported resolution for the difference between CLIP and methods with a margin in all three settings. Zero contact for the latter is already guaranteed by construction.

Table 2: Contrast diagnostics under a learned PRM. N_{mix} is the number of mixed groups. Methods with a margin have zero weak contact by construction.

Setting	Fusion	Pair rate ↓	Group rate ↓	Margin ↑	N_{mix}
GSM8K / 0.5B	ORM	0.000	0.000	2.45	313
	CLIP	0.147	0.572	1.66	313
	NASH	0.000	0.000	2.17	313
	RANK	0.000	0.000	1.75	313
GSM8K / 7B	CLIP	0.147	0.319	2.03	94
	NASH	0.000	0.000	2.48	94
MATH-500 / 0.5B	CLIP	0.126	0.576	1.75	257
	NASH	0.000	0.000	2.18	257

Table 3: Empirical upper tail diagnostic for Proposition 3. "Mono." is the fraction of evaluated threshold comparisons satisfying the empirical stochastic ordering check. "Observed" and "Margin" are contact rates for CLIP and the method with a margin.

Setting	Mono.	Tail	Observed	Margin
GSM8K / 0.5B	95%	0.137	0.147	0.000
MATH-500 / 0.5B	86%	0.141	0.126	0.000
GSM8K / 7B	100%	0.125	0.147	0.000

The difference between pair and group rates follows from Proposition 6. For illustration, four independent incorrect trajectories with contact probability 0.147 would yield

$$1 - (1 - 0.147)^4 \approx 0.47.$$

The observed group rate is 0.572 on GSM8K at 0.5B. The larger value can reflect heterogeneous trajectory qualities, dependence within prompts, or variation in group composition.

3.2 Training With and Without a Margin

We next compare ORM, CLIP, NASH, and Advantage Margin in end to end GRPO. The base policy is Qwen2.5-3B-Instruct, and process scores come from live Qwen2.5-Math-PRM-7B inference. GRPO uses group size $G = 8$, learning rate 1×10^{-6}, KL coefficient $\beta = 0.04$, sampling temperature $\tau = 1.0$, and the AdamW optimizer. The 3B experiment trains for 300 steps on 800 problems from MATH Levels 4 and 5. The 0.5B stress test trains for 100 steps with the same main hyperparameters. Both use three seeds and an independent verifier.

At 3B, every arm using process reward exceeds ORM in mean accuracy, suggesting that the process channel contains useful signal. Advantage Margin has the largest gain, $+2.8$ percentage points. This pattern is consistent with useful comparisons among incorrect trajectories together with protected outcome contrast. The table compares each method with ORM, not Advantage Margin directly with CLIP, so it does not isolate the margin as the unique cause of the additional gain.

At 0.5B, CLIP loses 8.9 percentage points relative to the untrained base, while ORM and both methods with a margin remain near or above the base. This is the strongest downstream evidence associated with boundary contact, but it is not a causal estimate of contact alone. CLIP also changes normalized gap magnitudes, and three seeds leave substantial optimization uncertainty. Seed ranges and the GSM8K controls are reported in Appendix C.3.

3.3 Intervention With Strict Reversals

CLIP creates collisions rather than strict reversals. To test the wrong sign regime directly, we start from ORM and, with probability ρ, swap the advantages of the best correct and best incorrect trajectories in a mixed group. The model, data, and hyperparameters remain fixed.

Performance is approximately flat through $\rho = 0.50$, then degrades by 5.1 and 8.5 percentage points at $\rho = 0.75$ and $\rho = 1.00$. Thus sufficiently frequent strict reversals can causally damage training under this intervention. The result should not be equated numerically with the CLIP result because

Table 4: End to end GRPO on MATH Levels 4 and 5 with three seeds. At 3B, Δ and p compare with ORM. At 0.5B, Δ compares with the untrained base and no p value is reported.

Scale	Fusion	Accuracy	Δ	p
3B	ORM	0.595 ± 0.005	N/A	N/A
	CLIP	0.607 ± 0.008	$+1.2$ pp	0.038
	NASH	0.612 ± 0.015	$+1.7$ pp	0.021
	Advantage Margin	$\mathbf{0.623 \pm 0.007}$	$\mathbf{+2.8}$ **pp**	$\mathbf{< 0.01}$
0.5B	Base	0.266	N/A	N/A
	CLIP	0.177	-8.9 pp	N/A
	ORM	0.271	$+0.5$ pp	N/A
	Advantage Margin	0.271	$+0.5$ pp	N/A
	NASH	0.276	$+1.0$ pp	N/A

Table 5: Strict reversal intervention on GSM8K at 0.5B with three seeds and 100 training steps.

Injection rate ρ	Accuracy	Δ vs. ρ=0
0.00	0.544	N/A
0.25	0.538	-0.6 pp
0.50	0.545	$+0.1$ pp
0.75	0.493	-5.1 pp
1.00	0.459	-8.5 pp

the datasets differ and swapping advantages is more severe than setting an incorrect reward equal to the correct ceiling.

3.4 What the Results Support

The two terms in (6) imply different conditions for risk and gain. Risk across outcome classes requires mixed groups and process score mass near the correct ceiling. Gain beyond ORM requires a process channel that usefully orders incorrect trajectories. The four settings separate these requirements. On MATH at 0.5B, contact is frequent and CLIP degrades, while the methods with a margin remain stable. On GSM8K at 0.5B, the area under the receiver operating characteristic curve among incorrect trajectories is about 0.53, close to chance. Methods with a margin recover the gain achieved by ORM but do not exceed it, consistent with protected outcome contrast and weak within class information. On MATH at 3B, all process based methods improve over ORM and Advantage Margin has the largest mean. On saturated GSM8K at 3B, mixed groups and groups in which every trajectory is incorrect are less common, and all methods lie within 1.4 percentage points. These observations support the two gate interpretation, but four settings do not establish a universal scaling law.

4 Related Work

Policy optimization and reward shaping. Reinforcement learning from human feedback [Ouyang et al., 2022] commonly optimizes a learned reward through PPO [Schulman et al., 2017] or GRPO [Shao et al., 2024, DeepSeek-AI, 2025], while direct preference optimization avoids an online scalar reward [Rafailov et al., 2023]. Classical reward shaping studies transformations that preserve optimal policies [Ng et al., 1999]. Our transformation result is narrower and local: positive affine changes preserve the direction of the fixed group update, while general monotone changes preserve order but can alter pairwise gap geometry.

Group relative reward shaping. RLRR addresses groups with zero reward variance [Niu et al., 2026], while concurrent work studies biased GRPO advantages across difficulty levels and gradient starvation under binary rewards [Yang et al., 2026, Nie et al., 2026]. Our decomposition connects but distinguishes these cases. Process reward can activate comparisons when every trajectory is incorrect, whereas a margin controls comparisons across outcome classes in mixed groups. The first issue concerns whether a gradient exists; the second concerns which pairwise directions it reinforces.

Reward model quality and post training failures. Low reward variance can flatten the RLHF landscape [Razin et al., 2025], and reward overoptimization and reward hacking arise when policy optimization exploits an imperfect proxy [Gao et al., 2023, Skalse et al., 2022]. Ceiling contact instead occurs in the reward computation map before the policy update. A collision attenuates an outcome comparison, while a strict reversal gives that comparison the wrong sign. Neither event requires the policy to have already exploited the reward model. Aggregate reward accuracy, calibration, or mean squared error need not reveal the cross class margin that controls these coefficients.

Process reward models. PRMs provide supervision at the level of reasoning steps [Uesato et al., 2022, Lightman et al., 2024, Wang et al., 2024]. We use Qwen2.5-Math-PRM [Zhang et al., 2025]; methods such as PRIME also use process rewards for online shaping [Cui et al., 2025]. This literature motivates the dense channel, while our analysis isolates a separate design question: how outcome and process signals should share the scalar reward consumed by group relative optimization. Process score quality governs comparisons within an outcome class, while the fusion map governs whether comparisons across classes are preserved, erased, or reversed.

5 Limitations

The theory concerns the local reward driven update and does not order final accuracies after repeated PPO epochs, KL regularization, and interacting gradients. It assumes an exact verifier, scalar latent quality, and a constant correct branch; A1 is only approximated through empirical strata. The experiments use one mathematical PRM, three seeds, and limited training horizons, while problem level tests are not substitutes for independent training runs. The reversal intervention is stronger than CLIP collision, and we do not report a full margin sweep or a direct test between Advantage Margin and CLIP. Generalization beyond mathematical reasoning remains open.

6 Conclusion

Reward fusion in GRPO is naturally a pairwise contrast problem. Process reward can add supervision among incorrect trajectories, but it can also weaken the outcome contrast that tells the optimizer which trajectory is correct. Ceiling collision erases that contrast, strict reversal changes its sign, and a positive outcome margin prevents both uniformly. The experiments show regimes in which process signal helps, boundary contact accompanies instability, and frequent injected reversals cause harm. The conclusion is deliberately structural rather than universal: *outcome sets the tier, while process reward orders within the tier*.

References

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

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999.

Wenhua Nie, Jianan Wu, Junlin Liu, Ziwei Li, Zheng Lin, Zijian Zhang, Yilong Fan, Haoran Zheng, and Jyh-Shing Roger Jang. Gradient starvation in binary-reward GRPO: Why group-mean centering fails and why the simplest fix works. *arXiv preprint arXiv:2605.07689*, 2026.

Wenzhe Niu, Wei He, Zongxia Xie, Jinpeng Ou, Huichuan Fan, Yuchen Ge, Yanru Sun, Ziyin Wang, Yizhao Sun, Chengshun warehouse Shi, Jiuchong Gao, Jinghua Hao, and Renqing He. From absolute to relative: Rethinking reward shaping in group-based reinforcement learning. *arXiv preprint arXiv:2601.23058*, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. In *Advances in Neural Information Processing Systems*, 2025.

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

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

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, pages 23253–23267, 2022.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin ... Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

A Proofs and Technical Details

A.1 Pairwise Decomposition

Proof of Theorem 1. Because the rewards are centered at their group mean, $\sum_i A_i = 0$. For arbitrary vectors $g_{1:G}$,

$$\sum_{i=1}^{G} A_i g_i = \frac{1}{G} \sum_{i<j} (A_i - A_j)(g_i - g_j).$$

Moreover,

$$A_i - A_j = \frac{R_i - R_j}{s_R + \epsilon}.$$

Multiplying the first identity by $1/G$ and substituting the second gives (3). □

Proof of Corollary 2. A strictly increasing transformation preserves every pairwise order. For $R'_i = aR_i + b$, $\bar{R}' = a\bar{R} + b$ and $s'_R = as_R$, so

$$A'_i = \frac{a(R_i - \bar{R})}{as_R + \epsilon} = \frac{a(s_R + \epsilon)}{as_R + \epsilon} A_i.$$

The same positive factor therefore multiplies U_R. A nonlinear transformation need not preserve ratios among pairwise reward gaps. Equation (3) then shows that it can change the direction of the weighted sum when the score difference vectors are not collinear. □

A.2 Contact and Hierarchy

Proof of Proposition 3. Because h is nondecreasing, both $\mathcal{E}_{\geq}$ and $\mathcal{E}_{>}$ are upper sets. First order stochastic dominance orders the probability of every upper set, so $\pi_{\geq}(q)$ and $\pi_{>}(q)$ are nondecreasing. The same argument shows that, whenever the expectation exists,

$$M_{>}(q) := \mathbb{E}[(h(S) - r_+)_+ \mid Q = q]$$

is nondecreasing, because $(h(s) - r_+)_+$ is a nondecreasing function of s and $(z)_+ := \max\{z, 0\}$. □

If the crossing set is empty or the channel has no mass in it, contact probability is zero. For CLIP, $\mathcal{E}_{\geq} = \mathcal{S} \cap [1, \infty)$ and $\mathcal{E}_{>} = \emptyset$.

Proof of Corollary 4. For CLIP, weak contact occurs when $q + Z \geq 1$. Therefore

$$\pi_{\geq}(q) = 1 - \Phi\left(\frac{1-q}{\sigma}\right) = \Phi\left(\frac{q-1}{\sigma}\right).$$

The stated limits follow directly. CLIP cannot exceed one, so $\pi_{>}(q) = 0$. □

Proof of Theorem 5. Strict reversal is impossible exactly when $h(s) \leq r_+$ for every s, which is equivalent to $\sup_s h(s) \leq r_+$ and therefore to $\gamma_\varphi \geq 0$. A uniform positive margin exists exactly when $\inf_s \Delta_\varphi(s) = \gamma_\varphi > 0$. Weak contact is impossible exactly when $h(s) < r_+$ for every s. If h attains its supremum, this strict pointwise inequality holds exactly when $\sup_s h(s) < r_+$, or equivalently $\gamma_\varphi > 0$. □

The attainment condition distinguishes pointwise separation from uniform separation. For example, $h(s) = r_+ - e^{-s}$ never reaches the ceiling at a finite score, but $\gamma_\varphi = 0$ and the outcome contrast can become arbitrarily small. A continuous branch on a compact score domain attains its maximum, so the two conditions coincide in that common setting.

A.3 Group Contact

Proof of Proposition 6. The first identity is the complement of no incorrect trajectory contacting the ceiling. One contacting incorrect trajectory forms a weakly contacting pair with each of the k correct trajectories. Linearity of expectation gives the second identity. □

If trajectory j depends on the group only through Q_j, then $\pi_j(Q_{\mathcal{W}}) = \pi_{\geq}(Q_j)$. The number of incorrect trajectories controls whether any contact occurs, while the number of correct trajectories controls how many pairwise coefficients are affected.

B Fusion Operators

Advantage Margin uses

$$r_+ = 1 + \delta, \qquad h(s) = \mathrm{clip}(s, 0, 1).$$

Its margin is at least δ, with equality when the clipped branch can attain one. After division by $1 + \delta$, it belongs to the normalized family with margin parameter $\delta/(1 + \delta)$.

NASH uses $r_+ = 1$ and

$$h_{\mathrm{NASH}}(s) := (1 - \lambda\widehat{\sigma})u(s),$$

where $\sup_s u(s) = 1$, $\widehat{\sigma} > 0$ is an estimated process noise scale, and $\lambda > 0$ is a chosen margin multiplier. When $0 < \lambda\widehat{\sigma} < 1$, its margin is $\lambda\widehat{\sigma}$.

For RANK, let $\mathrm{rank}_{\mathrm{lex}}(i) \in \{1, \ldots, G\}$ be the position of trajectory i when trajectories are sorted first by outcome and then by process score, with exact ties resolved deterministically. RANK assigns

$$R_i^{\mathrm{rank}} := \frac{\mathrm{rank}_{\mathrm{lex}}(i) - 1}{G - 1}.$$

RANK is group dependent and lies outside the pointwise class in (4), but it directly implements the same hierarchy.

C Experimental Details and Secondary Results

C.1 Pairwise Diagnostic Protocol

The diagnostic policies are Qwen2.5-{0.5B, 7B}-Instruct. We use 500 GSM8K problems with group size $G = 8$ and MATH-500. The outcome is exact answer match. The process score is the minimum Qwen2.5-Math-PRM-7B score across reasoning steps. The reported weak contact rate is computed over pairs formed by correct and incorrect trajectories. The group rate is computed over mixed groups.

C.2 Training Protocol

The base policy is Qwen2.5-3B-Instruct, and process scores come from live Qwen2.5-Math-PRM-7B inference. GRPO uses group size $G = 8$, learning rate 1×10^{-6}, KL coefficient $\beta = 0.04$, sampling temperature $\tau = 1.0$, and the AdamW optimizer. The arms are ORM, CLIP, NASH with $\lambda = 0.3$, and Advantage Margin with $\delta = 0.1$. The 3B experiment trains for 300 steps and evaluates greedy decoding on 800 MATH Level 4 and 5 problems with an independent verifier. The 0.5B experiment uses the same main hyperparameters for 100 steps.

At 3B, statistical significance is computed by paired permutation tests across problems, pooling seeds. These tests describe conditional benchmark differences and should not be interpreted as having many independent training runs.

C.3 Additional Accuracy Results

On GSM8K at 0.5B, CLIP shows no gain over the untrained base, whereas methods with a margin recover the $+7.2$ percentage point gain achieved by ORM. They do not exceed ORM. The area under the receiver operating characteristic curve among incorrect trajectories is about 0.53, so the second term in (6) is nearly uninformative in this setting.

Table 6: MATH Levels 4 and 5 at 0.5B with three seeds.

Fusion	Mean accuracy	Seed range	Δ vs. base
Base	0.266	N/A	N/A
CLIP	0.177	[0.120, 0.221]	-8.9 pp
ORM	0.271	[0.262, 0.288]	$+0.5$ pp
Advantage Margin	0.271	[0.268, 0.274]	$+0.5$ pp
NASH	0.276	[0.269, 0.288]	$+1.0$ pp

On GSM8K at 3B, base accuracy is about 87%. Mixed groups and groups in which every trajectory is incorrect are less frequent. The four training arms span only 1.4 percentage points with no consistent ordering.

C.4 Reversal Intervention Details

Table 7: Strict reversal intervention with seed ranges.

Injection rate ρ	Mean accuracy	Seed range	Δ vs. ρ=0
0.00	0.544	[0.522, 0.560]	N/A
0.25	0.538	[0.530, 0.550]	-0.6 pp
0.50	0.545	[0.518, 0.570]	$+0.1$ pp
0.75	0.493	[0.484, 0.508]	-5.1 pp
1.00	0.459	[0.426, 0.498]	-8.5 pp