DUOPO: JOINT TWO-AXIS ADAPTATION FOR REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) drives modern post-training of reasoning-capable large models, yet its dominant realization (Group-Relative Policy Optimization, GRPO) applies the same update rule to every rollout in a batch, overlooking two axes of rollout-level heterogeneity: the magnitude of the group-relative advantage $|A_i|$ and the correctness flag c_i. Through controlled pilot experiments, we establish that these two attributes are empirically near-independent during training, yet interact *non-separably* in their contribution to training utility: a high-$|A|$ *incorrect* rollout carries markedly more per-sample signal than a high-$|A|$ *correct* rollout of matched magnitude. Motivated by this observation, we propose **DuoPO**, a policy-optimization framework that instantiates a generalized RLVR objective with a non-separable joint weight, a signal-axis selector, and a correctness-conditioned loss family; DuoPO reduces to GRPO under a single degenerate choice of hyperparameters and provably dominates its separable and single-axis alternatives. On six multimodal reasoning benchmarks across two model scales, DuoPO achieves new state-of-the-art performance with *super-additive* gains: the joint treatment exceeds the sum of its single-axis contributions over GRPO on 4 of 6 benchmarks.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become a central post-training recipe for reasoning-capable large models (Shao et al., 2024; Guo et al., 2025). Group-Relative Policy Optimization (GRPO) and its clipped-surrogate variants make this recipe scalable: for each prompt, the policy samples a group of responses, normalizes their verifiable rewards into group-relative advantages, and applies the same sequence-level surrogate to every token. This uniform update is simple but coarse: as training progresses, many rollouts have advantages close to zero and contribute little, while correct and incorrect rollouts are treated identically even though they provide different kinds of supervision.

Recent methods relax GRPO's uniformity in two complementary directions. Advantage-aware sampling and reweighting (Yu et al., 2025; Zheng et al., 2025) allocate more computation to rollouts with stronger policy-gradient signals, but assign the same treatment to correct and incorrect responses. Self-distillation-based objectives (Hübotter et al., 2026) provide dense token-level supervision on failures, but do not distinguish high- and low-magnitude advantages. Applying either remedy alone leaves one source of rollout heterogeneity unused, and simply stacking the two does not specify how their effects should interact, how the update should be normalized, or which operation should be applied first. This raises the question: *can we develop a policy-optimization framework that jointly adapts along both axes and captures interactions no single-axis method can express?*

Answering this question is non-trivial because it requires resolving three coupled challenges. First, if the advantage magnitude $|A_i|$ and the correctness state c_i turn out to be strongly correlated during training, a single-axis method already captures both through the correlation and joint treatment is redundant. Second, even under independence, if the per-sample training utility factors as $\phi(|A_i|)\psi(c_i)$, then a product of two single-axis reweightings suffices and no genuinely joint mechanism is needed. Third, a non-separable update must remain expressive enough to encode the interaction yet simple enough to introduce only a handful of interpretable hyperparameters, recover GRPO as a degenerate point, and admit clean comparison against natural simpler alternatives. A pilot analysis of GRPO

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.

rollouts settles the first two: the two attributes are nearly independent over training, so adapting one axis leaves substantial variation on the other; their utility is nevertheless asymmetric, in that among rollouts with large $|A_i|$ incorrect responses provide more useful per-sample signal than correct responses of comparable magnitude, whereas the two states are similarly uninformative when $|A_i|$ is small. This combination of near-independence and state-dependent scaling cannot be represented by a separable weight $\phi(|A_i|)\psi(c_i)$.

We introduce **DuoPO**, a policy-optimization framework that adapts RLVR updates along both axes through three ordered components. A contrastive selector pairs high- and low-advantage rollouts and keeps the pairs with the largest combined contrast, removing the near-zero portion of the pool. A per-token joint weight assigns

$$w(A_i, c_i, t) = |A_i|^{1+\Delta(1-c_i)} \exp\big(-\beta(1-c_i)H_{i,t}\big),$$

where Δ controls the interaction between signal magnitude and correctness and β downweights tokens for which the self-teacher is diffuse; for any $\Delta > 0$ this weight is non-separable. The loss is then routed by correctness: correct rollouts use the reward-aligned GRPO surrogate, incorrect rollouts with teacher information use dense self-distillation, and the remainder use the GRPO fallback. A self-normalized weighted objective combines the branches without an additional mixing coefficient, and GRPO is recovered by setting the selector to the full group, the weight to a constant, and the loss to GRPO. On six visual reasoning benchmarks, DuoPO reaches a six-benchmark average of 65.0 with Qwen2.5-VL-7B, improving over the strongest zero-RL baselines by 2.9 points and obtaining the best result among the listed systems on five benchmarks; the gains persist under 32B scaling, soft-reward transfer, and text-only reasoning.

Our contributions are:

- We identify advantage magnitude $|A_i|$ and correctness c_i as distinct axes of RLVR updates and show that they are nearly independent but interact non-separably in training utility.
- We formulate a generalized RLVR objective and instantiate it as DuoPO, with a signal-axis selector, a non-separable joint weight, and correctness-conditioned loss routing; the framework contains GRPO as a degenerate case, uses four interpretable hyperparameters beyond GRPO, and admits a principled comparison with separable and single-axis alternatives.
- DuoPO improves over strong RLVR baselines across six visual reasoning benchmarks and two model scales, achieves a 65.0 average with Qwen2.5-VL-7B, and retains its gains in larger-scale, soft-reward, and text-only settings; controlled ablations verify the contribution and composition of each component.

2 Related Work

Large reasoning models. SFT on long chain-of-thought data (Muennighoff et al., 2025; Ye et al., 2025; Guo et al., 2024) and test-time-scaling methods such as MCTS (Yao et al., 2023; Zhang et al., 2024a; Yao et al., 2024; Guan et al., 2026) improve reasoning but have been argued to largely teach format rather than true reasoning (Chu et al., 2025a; Kang et al., 2024). Models such as OpenAI o1/o3 (OpenAI, 2024), DeepSeek-R1 (Guo et al., 2025), Seed-Thinking (Seed, 2025), and Kimi-k1.5 (Kimi, 2025) turned instead to RL with verifiable rewards, which DeepSeek-R1-Zero showed can be applied directly to pre-trained policies without SFT. Our work sits in this line and focuses on the efficiency of the RL update rule itself.

RL for multimodal LLMs. A parallel line transplants RLVR to multimodal reasoning and downstream visual tasks: open-vocabulary detection (Liu et al., 2025c), reasoning segmentation (Liu et al., 2025b), video understanding (Li et al., 2025), localization (Wang et al., 2025c). Works targeting general MLLM reasoning (Meng et al., 2025; Huang et al., 2025; Yang et al., 2025; Zhang et al., 2025; Peng et al., 2025; Tan et al., 2025) focus on data curation and SFT/RL balance. Closest to us are efforts that modify the RL mechanism itself: contrastive rewards (Li et al., 2025), reflection tokens (Wang et al., 2025a), objective and gradient modifications (Chu et al., 2025b), and diversified rollouts (Liu et al., 2025a; Yao et al., 2025). In particular, prior advantage-based batch reorganization and correctness-gated dense supervision act along the two axes we identify; DuoPO shows they are best treated jointly and non-separably.

Advantage weighting and sample reweighting in policy optimization. The idea of weighting rollouts by some function of advantage has a long history: advantage-weighted regression (Peng et al., 2019) and AWAC (Nair et al., 2020) scale the update by $\exp(A/\tau)$ for offline/off-policy control; prioritized experience replay (Schaul et al., 2016) draws samples proportionally to TD error; token-level reweighting (Yu et al., 2025; Zheng et al., 2025; Chu et al., 2025b) recalibrates the PPO surrogate at sequence or token granularity. All of these are *separable* in the attributes they consider: typically a single scalar signal broadcast uniformly over a rollout's state. DuoPO departs from this pattern in two ways: (i) it explicitly couples signal and state into a joint weight that admits no product factorization (Lemma G.1), and (ii) it routes the *form* of the per-token loss by correctness rather than only rescaling a fixed loss. The asymmetric-utility observation of §4.1 is the empirical justification that such a non-separable coupling is not cosmetic.

3 PRELIMINARIES

Let π_θ denote the policy being trained, $\mathcal{D}$ a prompt distribution, and, for each $x \sim \mathcal{D}$, $\{y_i\}_{i=1}^G$ a group of G on-policy rollouts $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ with verifiable rewards $\{r_i\}_{i=1}^G$. Write

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})} \tag{1}$$

for the per-token importance ratio.

3.1 GROUP-RELATIVE POLICY OPTIMIZATION

GRPO (Shao et al., 2024) assigns each rollout a group-relative advantage

$$A_i \;=\; \frac{r_i - \bar{r}}{\sigma_r + \epsilon}, \qquad \bar{r} = \tfrac{1}{G}\textstyle\sum_j r_j, \;\; \sigma_r = \mathrm{std}(\{r_j\}), \tag{2}$$

and optimizes the clipped surrogate

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}}\left[\frac{1}{\sum_i |y_i|}\sum_{i,t}\min\Big(\rho_{i,t}(\theta)\,A_i,\; \mathrm{clip}\big(\rho_{i,t}(\theta), 1{-}\varepsilon, 1{+}\varepsilon\big)\,A_i\Big)\right]. \tag{3}$$

The sequence-level advantage A_i is broadcast uniformly across all tokens of rollout i, giving reward-aligned but coarse-grained credit assignment.

3.2 DENSE LOGIT-LEVEL SUPERVISION

An alternative surrogate matches the policy to a self-distilled target. Given side information f_i available at rollout time (for instance, a successful sibling rollout within the same group, or an execution trace from the environment), we obtain a teacher distribution by re-conditioning the current policy on this side information, $q_{i,t}(v) = \pi_\theta(v \mid x, f_i, y_{i,<t})$. The student is trained to match $q_{i,t}$ along the realized trajectory y_i via a KL-divergence surrogate,

$$\ell_{i,t}^{\text{SDPO}}(\theta) \;=\; \mathrm{KL}\Big(\pi_\theta(\cdot \mid x, y_{i,<t}) \,\Big\|\, \mathrm{sg}[\pi_\theta(\cdot \mid x, f_i, y_{i,<t})]\Big), \tag{4}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. Unlike equation 3, this provides dense token-level supervision whose reliability varies with the sharpness of the teacher (Hübotter et al., 2026).

3.3 OBSERVED INEFFICIENCIES OF UNIFORM TREATMENT

Applying either equation 3 or equation 4 uniformly across a rollout group wastes training signal in two ways that recent work has documented in isolation. On the *signal side*, the empirical distribution of A_i concentrates near zero as training progresses, so that a shrinking minority of rollouts dominates the gradient while the rest contribute negligibly; this concentration worsens at both extremes of query difficulty. On the *state side*, uniformly applying dense supervision to already-correct rollouts introduces an optimization pressure toward distribution-matching on solved problems, degrading late-stage stability and causing observable collapse of policy entropy. These two failure modes have, to our knowledge, only been analyzed as independent deficiencies. In the next section we show they are best viewed as two facets of a single structural property of RLVR rollouts: one that neither single-axis remedy can fully address.

4 TWO-AXIS ROLLOUT HETEROGENEITY

We formalize the structural view that drives DuoPO. For each rollout y_i in a group, we extract two attributes computable at rollout time:

$$A_i \in \mathbb{R} \quad \text{(signal axis: group-relative advantage)}, \qquad c_i = \mathbf{1}[y_i \text{ is correct}] \in \{0,1\} \quad \text{(state axis)}. \tag{5}$$

A teacher-availability flag $m_i = \mathbf{1}[f_i \text{ is available for } y_i]$ is not a third axis but a gate on which surrogate is well-posed on the state axis. These two attributes are not arbitrary: $|A_i|$ upper-bounds the per-rollout contribution to the policy gradient under equation 3, and c_i controls whether the dense supervision equation 4 is informative on a given sample.

4.1 EMPIRICAL PROPERTIES: INDEPENDENCE AND ASYMMETRIC UTILITY

Near-independence. On Qwen2.5-VL-7B trained with GRPO on a multimodal reasoning corpus (setup in Sec. 6), we log $(|A_i|, c_i)$ for every rollout over `[xx]` training steps. The step-wise Pearson correlation averages $\overline{|r|} =$ `[xx]` and never exceeds `[xx]`; the joint distribution is nearly rectangular. Any method that adapts along one axis is in expectation blind to the orthogonal complement of that axis.

Asymmetric utility. Independence alone justifies *separate* treatment but not *joint* treatment. Define $U(|A|, c)$ as the average held-out-loss reduction attributable to including a rollout with attributes $(|A|, c)$. We observe

$$\frac{U(|A|,\, c{=}0)}{U(|A|,\, c{=}1)} \;\approx\; \texttt{[xx]} \;\text{ for large } |A|, \qquad \text{vs.} \;\; \approx 1 \;\text{ for small } |A| : \tag{6}$$

high-$|A|$ *incorrect* rollouts supply markedly more per-sample signal than high-$|A|$ correct rollouts of matched magnitude, while at low $|A|$ the two states are comparably uninformative. This is the structural fingerprint of a *non-separable* interaction between the two attributes: no weight function $\phi(|A|)\psi(c)$ can reproduce equation 6, since $\phi(|A|)\psi(0)/\phi(|A|)\psi(1)$ is a constant in $|A|$ (Lemma G.1, Appendix G).

4.2 GENERALIZED OBJECTIVE AND GRPO DEGENERACY

The most general single-step update consistent with equation 5 is

$$\boxed{\mathcal{L}_{2\mathrm{A}}(\theta) \;=\; \frac{1}{Z} \sum_{i \in \mathcal{S}(\mathbf{A},\mathbf{c})} \sum_{t=1}^{|y_i|} w(A_i,\, c_i,\, t) \cdot \ell\big(y_{i,t} \,\big|\, c_i; \theta\big)} \tag{7}$$

with $Z = \sum_{i \in \mathcal{S}} \sum_t w(A_i, c_i, t)$, parameterized by a *selector* $\mathcal{S}$, a *joint weight* w, and a *conditional loss family* $\ell(\cdot \mid c_i)$. An instance is *separable* if $w(A_i, c_i, t) = \phi(A_i)\psi(c_i, t)$, and *non-separable* otherwise. By §4.1, any separable w mis-specifies the training utility; equation 7 is the smallest class that can correct this mis-specification.

Proposition 4.1 (GRPO degeneracy)**.** *Taking $\mathcal{S} = \{1, \ldots, G\}$, $w \equiv 1$, $\ell(\cdot \mid c_i) \equiv \ell^{\mathrm{GRPO}}$ reduces equation 7 to equation 3 (proof in Appendix G).*

Equation equation 7 is therefore a strict superset of GRPO, and every design choice in Sec. 5 can be read as a principled commitment about *how* to move away from the GRPO corner of this design space.

5 DUOPO: JOINT TWO-AXIS ADAPTATION

We now specify the three components of equation 7; Fig. 1 provides a high-level overview of the whole pipeline.

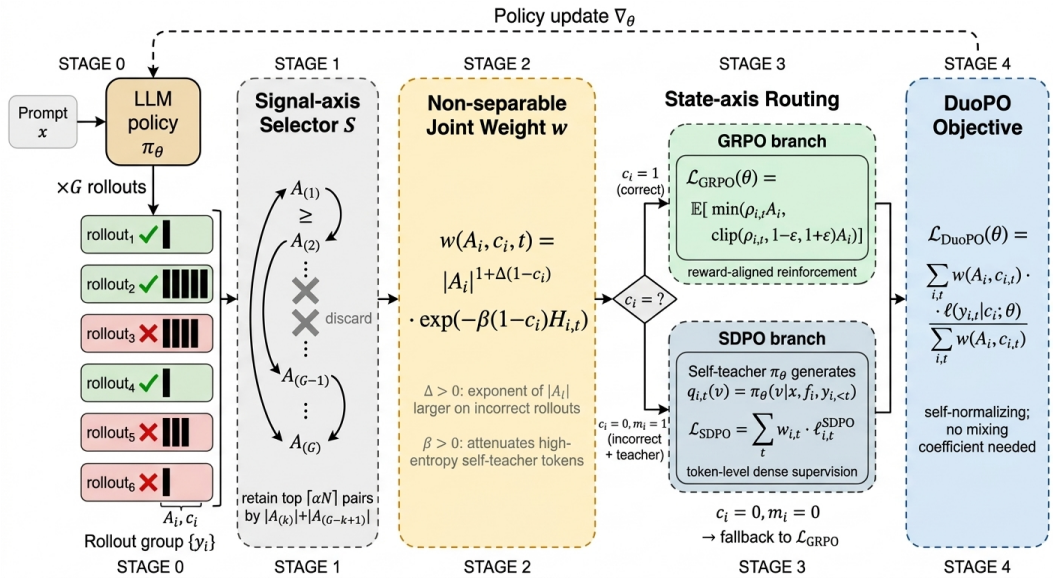

Figure 1: **Overview of the DuoPO framework.** Starting from a prompt x, the policy π_θ produces G rollouts $\{y_i\}$ with advantages A_i and correctness flags c_i (Stage 0). *Stage 1 (Signal-axis Selector $\mathcal{S}$, Eq. 9):* rollouts are sorted by A_i and paired by ranked extremes; only the top $\lceil \alpha N \rceil$ pairs by combined contrast $|A_{(k)}| + |A_{(G-k+1)}|$ are retained. *Stage 2 (Non-separable Joint Weight w, Eq. 10):* each token receives the joint weight $w(A_i, c_i, t) = |A_i|^{1+\Delta(1-c_i)} \exp(-\beta(1-c_i)H_{i,t})$, in which $\Delta > 0$ asymmetrically amplifies high-$|A|$ *incorrect* rollouts (§4.1) and $\beta > 0$ attenuates tokens where the self-teacher is diffuse. *Stage 3 (State-axis Routing, Eq. 11):* the per-token loss is routed by c_i: reward-aligned ℓ^{GRPO} on correct rollouts, dense self-teacher supervision ℓ^{SDPO} on incorrect rollouts with available teacher information ($m_i = 1$), and fallback to ℓ^{GRPO} otherwise. *Stage 4 (DuoPO Objective, Eq. 12):* the weighted losses are aggregated via a self-normalizing ratio, so no external mixing coefficient is needed: a less-competent policy is driven by the dense-supervision branch and a more-competent policy by the reward-aligned branch.

5.1 Selector $\mathcal{S}$: Contrastive Sub-Sampling on the Signal Axis

To counter the near-zero concentration of advantages in §3.3, we restrict the update to a high-contrast subset of the rollout pool. Sort the $G = 2N$ rollouts by advantage and pair the k-th highest with the k-th lowest:

$$A_{(1)} \geq \cdots \geq A_{(G)}, \qquad \mathcal{P} = \big\{(y_{(k)}, y_{(G-k+1)})\big\}_{k=1}^{N}. \tag{8}$$

We retain the top $\lceil \alpha N \rceil$ pairs by combined contrast $|A_{(k)}| + |A_{(G-k+1)}|$:

$$\mathcal{S}(\mathbf{A}) \;=\; \bigcup_{k=1}^{\lceil \alpha N \rceil} \{(k), (G-k+1)\}, \qquad \alpha \in (0, 1]. \tag{9}$$

$\mathcal{S}$ depends only on $\mathbf{A}$, not on $\mathbf{c}$: selecting before routing preserves the contrastive structure of each pair. We revisit the ordering (select-then-route vs. route-then-select) at the end of this section.

5.2 Joint Weight w: Non-Separable Reweighting on Both Axes

The joint weight is the only component of equation 7 that depends on both A_i and c_i, and is the place where the asymmetric-utility observation of §4.1 is operationalized:

$$\boxed{w(A_i,\, c_i,\, t) \;=\; |A_i|^{\,1+\Delta\,(1-c_i)} \;\cdot\; \exp\big(-\,\beta\,(1-c_i)\,H_{i,t}\big)} \tag{10}$$

with $\Delta \geq 0$, $\beta \geq 0$ the only two hyperparameters introduced beyond GRPO, and $H_{i,t} = -\sum_v q_{i,t}(v) \log q_{i,t}(v)$ with $q_{i,t}(v) = \pi_\theta(v \mid x, f_i, y_{i,<t})$ the Shannon entropy of the self-teacher distribution. The three factors play distinct roles. (*i*) $|A_i|$ concentrates gradient mass on high-contrast

rollouts, matching the $|A|$-proportional importance resampling in Algorithm 1. (*ii*) The exponent $1+\Delta(1-c_i)$ equals 1 on correct rollouts and $1+\Delta$ on incorrect rollouts; for any $\Delta > 0$, equation 10 admits no factorization $w = \phi(A_i)\psi(c_i, t)$ (Lemma G.1), amplifying high-$|A|$ failures more than matched successes as mandated by equation 6. (*iii*) The entropy factor attenuates tokens on which the teacher is diffuse and is gated to the incorrect branch via $(1 - c_i)$; on correct rollouts it collapses to 1.

Remark 5.1 (What Δ buys)**.** $\Delta = 0$ *reduces equation 10 to* $w = |A_i| \exp(-\beta(1 - c_i)H_{i,t})$*, separable in* (A_i, c_i)*. Any empirical gap between* $\Delta > 0$ *and* $\Delta = 0$ *(Sec. 6) is direct evidence that the non-separable structure of §4.1 is exploitable.*

5.3 Conditional Loss $\ell(\cdot \mid c_i)$: State-Axis Routing

The loss family is indexed by the state axis:

$$\ell\big(y_{i,t} \,\big|\, c_i; \theta\big) \;=\; \begin{cases} \ell_{i,t}^{\mathrm{GRPO}}(\theta) & \text{if } c_i = 1, \\ \ell_{i,t}^{\mathrm{SDPO}}(\theta) & \text{if } c_i = 0 \text{ and } m_i = 1, \\ \ell_{i,t}^{\mathrm{GRPO}}(\theta) & \text{if } c_i = 0 \text{ and } m_i = 0, \end{cases} \tag{11}$$

Incorrect rollouts without available teacher information fall back to the reward-aligned surrogate; when $m_i = 0$ for every rollout, equation 11 reduces to uniform GRPO on that group.

5.4 Full Objective, Simpler Alternatives, and Composition Order

Substituting equation 9, equation 10, equation 11 into equation 7 yields the DuoPO objective:

$$\mathcal{L}_{\mathrm{DuoPO}}(\theta) = \frac{\sum_{i \in \mathcal{S}(\mathbf{A})} \sum_{t=1}^{|y_i|} |A_i|^{1+\Delta(1-c_i)}\, e^{-\beta(1-c_i)H_{i,t}}\, \ell(y_{i,t} \mid c_i; \theta)}{\sum_{i \in \mathcal{S}(\mathbf{A})} \sum_{t=1}^{|y_i|} |A_i|^{1+\Delta(1-c_i)}\, e^{-\beta(1-c_i)H_{i,t}}}. \tag{12}$$

Because the denominator sums the same weights applied in the numerator, no external mixing coefficient is needed: a less-competent policy produces many low-c_i rollouts and is driven primarily by the dense-supervision branch, while a more-competent policy is increasingly driven by the reward-aligned branch. Over GRPO, DuoPO introduces four interpretable scalars: $\{\alpha, S, \Delta, \beta\}$ (sampling ratio, importance-resampling rounds, asymmetric amplification, entropy-gate sharpness).

Simpler alternatives. Equation equation 7 admits four natural simpler instantiations than equation 9–equation 11: (*i*) a separable multiplicative w (Δ=0), which cannot reproduce the asymmetric utility of §4.1 (Lemma G.1); (*ii*) a separable additive $w = \phi(A_i) + \psi(c_i, t)$, which fails to specialize to GRPO; (*iii*) a c_i-only selector, which inherits the near-zero $|A|$ concentration of §3.3; and (*iv*) the reversed composition order (routing first, then selecting contrastive pairs within each branch), which halves the expected contrast magnitude (Prop. G.2, Appendix G). A full argument for each is given in Appendix G; each is ruled out empirically in Sec. 6.

Summary. DuoPO operationalizes the two-axis view of §4 as three tightly coupled components placed in a principled order: the signal-axis selector $\mathcal{S}$ first de-noises the rollout pool by retaining only high-contrast pairs, the non-separable joint weight w then redistributes per-token gradient mass so that high-$|A|$ incorrect rollouts are amplified in agreement with the asymmetric-utility observation equation 6, and the correctness-conditioned loss family $\ell(\cdot \mid c_i)$ routes each rollout to the surrogate on which it is well-posed (reward-aligned on correct samples, dense self-teacher on informative failures). The self-normalizing aggregator in equation 12 removes the need for an external mixing coefficient between the two branches, yielding a single objective governed by only four interpretable scalars $\{\alpha, S, \Delta, \beta\}$ on top of GRPO. This design is not an arbitrary pipeline: the single scalar Δ controls the (A, c) interaction (Remark 5.1, Lemma G.1), the select-then-route ordering preserves contrastive magnitude (Prop. G.2), and GRPO is recovered as the unique degenerate point (Prop. 4.1). Each of the four simpler alternatives above is ruled out by at least one of these structural commitments, and the resulting objective exhibits the super-additive behaviour documented in Sec. 6.

Algorithm 1 DuoPO: one training iteration

Require: Policy π_θ; prompt x; group size $G = 2N$; hyperparameters $\alpha \in (0, 1]$, $\Delta \geq 0$, $\beta \geq 0$, $S \in \mathbb{N}$.

1: Sample rollouts $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}$; compute $\{r_i\}$, $\{c_i\}$; construct teacher info $\{f_i\}$, $\{m_i\}$; $A_i \leftarrow (r_i - \bar{r})/(\sigma_r + \epsilon)$.

2: *// Selector $\mathcal{S}$ (Eq. equation 9):* sort by A_i, pair extremes, retain top $\lceil \alpha N \rceil$ pairs by combined contrast; call this set B.

3: *// Importance resampling:* for $s = 1, \ldots, S$, draw a sub-batch from B with $\text{Prob}(p_j) \propto |A_{j,1}| + |A_{j,2}|$; concatenate into B'.

4: *// Joint weight w + conditional loss ℓ (Eqs. equation 10,equation 11):* for each $(i, t) \in B'$, set $\ell_{i,t}$ by equation 11 and $w_{i,t} \leftarrow |A_i|^{1+\Delta(1-c_i)} e^{-\beta(1-c_i)H_{i,t}}$.

5: *// Objective:* $\mathcal{L} \leftarrow \big(\sum_{i,t} w_{i,t}\ell_{i,t}\big)/\big(\sum_{i,t} w_{i,t}\big)$; update θ by gradient descent.

Table 1: Performance of DuoPO trained on the Geometry3K corpus.

Method	Geo3K	Math Avg.	HallBench	ChartQA
Qwen-3B	25.79	41.71	59.83	73.08
+ GRPO	42.64	46.74	63.09	76.20
+ DAPO	45.09	48.08	63.24	76.70
+ GSPO	43.22	47.26	**63.67**	75.12
+ DuoPO	**48.12**(+22.33)	**48.93**(+7.22)	63.44(+3.61)	**77.28**(+3.30)
Qwen-7B	38.12	49.82	65.19	79.84
+ GRPO	52.60	53.13	68.56	80.84
+ DAPO	54.43	54.19	69.29	81.20
+ GSPO	52.83	54.27	69.48	80.96
+ DuoPO	**56.24**(+18.12)	**54.88**(+5.06)	**69.78**(+4.59)	**81.92**(+2.08)

Table 2: Performance of DuoPO trained on the MMK12 corpus.

Method	K12	Math Avg.	HallBench	ChartQA
Qwen-3B	42.42	41.71	59.83	73.08
+ GRPO	59.19	48.71	64.14	77.12
+ DAPO	61.42	49.75	65.08	77.00
+ GSPO	60.44	48.76	64.14	77.56
+ DuoPO	**62.58**(+20.16)	**50.31**(+8.60)	**66.04**(+6.21)	**78.52**(+5.44)
Qwen-7B	52.13	49.82	65.19	79.84
+ GRPO	66.15	54.47	67.75	82.48
+ DAPO	68.35	54.52	68.66	82.52
+ GSPO	67.44	54.77	69.13	82.04
+ DuoPO	**69.15**(+17.02)	**55.28**(+5.46)	**70.16**(+4.97)	**82.88**(+3.04)

6 Experiments

6.1 Experimental Setup

Datasets and benchmarks. We evaluate DuoPO in two regimes. To isolate algorithmic effects from data-scale effects, we first train on two small free-form corpora of matched size: Geometry3K (Lu et al., 2021) ($2.1k$ samples, 'Geo3K') and an equal-sized slice of MMK12 (Meng et al., 2025) ('K12'). For scalability we then use a $30k$-sample pool assembled from the full Geo3K together with $27k$ random MM-Eureka (Meng et al., 2025) queries. In-domain accuracy is measured on the held-out Geo3K and K12 splits; out-of-domain generalization is tested on six benchmarks spanning math (MathVerse (Zhang et al., 2024b), MathVision (Wang et al., 2024), MathVista (Lu et al., 2023), WeMath (Qiao et al., 2025)), hallucination (HallusionBench (Guan et al., 2024)), and charts (ChartQA (Masry et al., 2022)). Free-form answers are graded with MathRuler; multiple-choice items are adjudicated by Gemini-2.0-Flash-001 (Deepmind, 2025). See Appendix E for details.

Implementation. DuoPO is implemented on top of EasyR1 (Yaowei et al., 2025) with Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025) as base policies; the vision tower is frozen. Each step uses an update batch of 128, a rollout batch of $G = 512$, temperature 1.0, and learning rate $1e{-}6$. Every query produces $2N{=}16$ rollouts, from which the selector $\mathcal{S}$ retains $\lceil \alpha N \rceil{=}4$ contrastive pairs under $\alpha{=}0.5$. Importance resampling uses $S{=}8$ rounds; the two weight-family scalars are $\Delta{=}0.5$ and $\beta{=}0.1$. All runs use 8×80G GPUs. Evaluation uses decoding temperature 0.5 and every reported number is the mean pass@1 over 8 seeds.

6.2 Main Results

Head-to-head against RLVR baselines. We first compare DuoPO against the closest same-budget baselines: GRPO (Shao et al., 2024), DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025). On Geo3K (Tab. 1), 3B DuoPO reaches 48.12%, beating GRPO by 5.48 and DAPO by 3.03; 7B DuoPO reaches 56.24%, beating the same two by 3.64 and 1.81. Out-of-domain, the six-benchmark math average rises by 2.19 (3B) and 1.75 (7B); HallusionBench and ChartQA move in the same direction

Table 3: Model performance on six visual reasoning benchmarks. '*' marks re-evaluation with our vLLM scripts. †Vision-R1-7B uses WeMath/MathVision for training. Best among RL-only models in **bold**, second best *underlined*.

Model	MathVerse	MathVision	MathVista	WeMath	HallBench	ChartQA	Avg.
Close-source							
GPT-4o (Achiam et al., 2024)	50.8	30.4	63.8	68.8	55.0	-	-
o1 (OpenAI, 2024)	57.0	60.3	73.9	-	-	-	-
Gemini-2.0 pro (Deepmind, 2025)	67.3	48.1	71.3	-	49.8	-	-
Claude-3.7-Sonnet (Anthropic, 2025)	52.0	41.3	66.8	72.6	55.4	-	-
Open-Source SFT							
InternVL-2.5-8B (Chen et al., 2025b)	39.5	17.0	64.5	-	50.1	79.1	-
InternVL-3-8B (Zhu et al., 2025)	-	29.3	71.6	-	49.9	86.6	-
Qwen2.5-VL-3B* (Bai et al., 2025)	34.8	21.9	58.4	51.7	59.8	73.1	49.9
Qwen2.5-VL-7B* (Bai et al., 2025)	42.6	25.8	67.4	63.5	65.2	79.8	57.4
Cold-Start + RL							
R1-VL-7B* (Zhang et al., 2025)	40.1	24.3	62.3	59.8	60.9	76.1	53.9
Vision-R1-7B*† (Huang et al., 2025)	46.1	-	70.8	-	57.8	83.1	-
R1-OneVision-7B* (Yang et al., 2025)	43.0	24.8	61.2	60.6	66.4	77.8	55.2
OpenVLThinker-7B* (Deng et al., 2025)	46.4	24.8	69.7	67.2	59.1	78.4	57.6
VLAA-Thinker-7B* (Chen et al., 2025a)	48.9	26.3	69.9	67.7	67.5	80.1	60.1
Zero RL							
MM-Eureka-Qwen-7B* (Meng et al., 2025)	49.6	27.4	70.6	67.4	66.7	79.0	60.1
MMR1-Math-7B* (Leng et al., 2025)	39.2	**31.9**	71.5	70.7	69.6	<u>82.0</u>	60.8
ThinkLite-VL-7B* (Wang et al., 2025b)	45.2	28.0	<u>72.4</u>	69.3	<u>70.4</u>	<u>82.0</u>	61.2
VL-Rethinker-7B* (Wang et al., 2025a)	<u>51.7</u>	29.7	72.0	70.1	69.9	79.0	<u>62.1</u>
NoisyRollout-7B-K12* (Liu et al., 2025a)	50.1	28.0	70.9	<u>70.8</u>	70.1	81.4	<u>62.1</u>
DuoPO-Qwen-3B (Ours)	44.5	27.1	70.7	66.9	69.5	80.2	59.8
DuoPO-Qwen-7B (Ours)	**54.2**	<u>30.3</u>	**77.3**	**72.6**	**71.3**	**84.4**	**65.0**

despite larger distribution shift. K12 (Tab. 2) shows the same pattern. Compared with GSPO, DuoPO is clearly ahead on both in-domain and out-of-domain metrics, indicating that per-token reweighting carries information that sequence-level importance sampling cannot recover. Additional comparisons with RLOO / REINFORCE++ are in Appendix C.

Comparison with RL-trained multimodal reasoners. Tab. 3 compares DuoPO with a panel of 7B-scale RL systems. Under a $30k$-sample, 150-step budget, 7B DuoPO attains the best six-benchmark average (65.0), beating the strongest zero-RL baseline (VL-Rethinker-7B / NoisyRollout-7B-K12, both 62.1) by 2.9 points and every cold-start+RL system. On MathVista (77.3), WeMath (72.6), HallBench (71.3), and ChartQA (84.4), DuoPO matches or surpasses proprietary Claude-3.7-Sonnet and GPT-4o. The 3B variant already matches several 7B SFT+RL baselines on average (59.8).

Where the gains come from. The wins trace back to how DuoPO allocates gradient budget. Advantage-histogram and training-dynamics analyses (Appendix C) show that (i) $\mathcal{S}$ removes the near-zero spike of uniform GRPO groups and w further concentrates mass on the high-$|A|$ tail; (ii) train/validation accuracy rises roughly twice as fast under DuoPO, mitigating the late-stage "rollout silencing" of GRPO; and (iii) at a fixed step count DuoPO increases GPU time by only 4–8% and reaches GRPO's final accuracy in about half the steps.

6.3 Ablation Study

All ablations run on Qwen2.5-VL-3B with Geo3K. We answer three questions: (i) what is the contribution of each DuoPO component (selector $\mathcal{S}$, joint weight w, conditional loss ℓ); (ii) is the chosen design inside each component better than simpler alternatives; and (iii) does the composition order prescribed by equation 7 actually matter. Extended breakdowns, the $(\alpha, S, \Delta, \beta)$ sweeps, and a prioritised-experience-replay comparison are in Appendix C.

Component-wise contribution. We enable $\mathcal{S}$, w, ℓ incrementally (Tab. 4). The selector alone already closes most of the distance from GRPO to the full model (+3.66 on Geo3K); adding the non-separable weight w contributes another +1.10; the correctness-conditioned loss ℓ contributes

Table 4: Component-wise ablation on DuoPO: selector $\mathcal{S}$, joint weight w, conditional loss ℓ.

GRPO	$\mathcal{S}$	w	ℓ	Geo3k	Math	Hall.	Chart
				25.79	41.71	59.83	73.08
✓				42.64	46.74	63.09	76.20
✓	✓			46.30	47.69	**63.52**	76.48
✓		✓	✓	46.94	47.91	62.80	75.56
✓	✓	✓	✓	**48.12**	**48.93**	63.44	**77.28**

Table 5: Design-choice ablation inside each DuoPO component.

Setting	Geo3k	Math	Hall.	Chart
Design of selector $\mathcal{S}$				
top-k	41.48	44.92	63.24	75.60
bottom-k	23.74	41.60	61.12	74.40
random	43.68	46.70	63.22	76.04
contrastive	**46.30**	**47.69**	**63.52**	**76.48**
Design of joint weight w				
sep. (Δ=0)	46.02	47.34	63.19	76.60
no gate (β=0)	46.48	47.66	63.21	76.74
non-sep. (ours)	**48.12**	**48.93**	63.44	**77.28**

$+0.72$ on top, reaching 48.12%. Line 4 disables the selector and keeps $w{+}\ell$: this variant improves over GRPO but lags the full DuoPO by roughly a point on every metric, showing that w and ℓ act most effectively on a pool that $\mathcal{S}$ has already de-noised. The three effects are super-additive on all four columns.

Design inside each component. Tab. 5 (upper) replaces the contrastive pairing of $\mathcal{S}$ with three alternatives at matched retention ratio. Only contrastive pairing beats GRPO on every column; one-way negative collapses training (23.7% on Geo3K) and one-way positive loses 1.16 points, a direct empirical confirmation of the asymmetric-utility observation of §4.1. Tab. 5 (lower) disables each factor of w. Δ=0, the strongest separable instantiation flagged by Remark 5.1, loses 2.10 points on Geo3K and 1.59 on math-avg, quantifying the gain attributable to the (A, c) interaction that no separable weight can express. Removing the entropy gate (β=0) loses a smaller but consistent 1.64 and 1.27. Neither reduced variant recovers the full score.

Composition order matters. A subtle design choice in equation 7 is whether the selector $\mathcal{S}$ is applied before or after the state-axis routing equation 11. Proposition G.2 predicts that the reversed order (routing first, then selecting contrastive pairs within each c_i-branch) should roughly halve the expected contrast magnitude. Tab. 6 confirms this: the reversed order recovers only 2.14 of the 5.48-point gap between GRPO and full DuoPO.

Table 6: Composition order: selecting before routing preserves contrast; the reverse order (Prop. G.2) loses more than half the gain.

Order	Geo3k	Math	Hall.	Chart
+ GRPO (baseline)	42.64	46.74	63.09	76.20
route-then-select	44.78	47.22	63.12	76.38
select-then-route (ours)	**48.12**	**48.93**	**63.44**	**77.28**

Hyperparameter sensitivity. Sweeping $\alpha \in \{0.25, 0.5, 0.75\}$, $S \in \{4, 8, 16\}$, $\Delta \in \{0, 0.25, 0.5, 1\}$, $\beta \in \{0, 0.1, 0.2\}$ (Appendix C, Tab. 13–14) all give single optima near $(\alpha, S, \Delta, \beta){=}(0.5, 8, 0.5, 0.1)$, with graceful degradation across wide neighbourhoods.

Scaling and transfer. DuoPO's gains are not tied to one backbone or task. On Qwen2.5-VL-32B with the same recipe, DuoPO improves over the GRPO-trained baseline by 1.9 points on average (Tab. 7). On Referring Expression Comprehension (Qwen2.5-VL-3B, soft IoU reward), DuoPO beats GRPO on every RefCOCO/RefCOCO+/RefCOCOg split, with the largest $+5.57$ on RefCOCO+/testB (Tab. 8). On text-only reasoning (Qwen2.5-Math-1.5B/7B-Base, Open-S1 (Dang & Ngo, 2025)), DuoPO beats GRPO uniformly, with the largest gains on AIME24 ($+6.3$ at 1.5B, $+3.6$ at 7B) (Tab. 9).

7 Conclusion

We introduce DuoPO, a policy-optimization framework that adapts the RLVR update jointly along the signal axis $|A_i|$ and the state axis c_i, which are near-independent yet interact non-separably in training utility. DuoPO instantiates this view with a contrastive selector, a non-separable joint weight, and a correctness-conditioned loss family, and reduces to GRPO under a single degenerate choice. Across six multimodal benchmarks, two model scales, and 32B, soft-reward, and text-only settings, DuoPO consistently outperforms GRPO, DAPO, GSPO and the strongest open-source 7B baselines.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. In *Annual Meeting of the Association for Computational Linguistics*, pp. 12248–12267, 2024.

Anthropic. Claude 3.7 Sonnet, anthropic.com, 2025. URL https://www.anthropic.com/claude/sonnet. 2025-02-24.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. In *Transactions on Machine Learning Research*, 2025a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025b. URL https://arxiv.org/abs/2412.05271.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *Proc. of Intl. Conf. on Machine Learning*, pp. 10818–10838, 2025a.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning, 2025b. URL https://arxiv.org/abs/2504.02546.

Quy-Anh Dang and Chris Ngo. Reinforcement learning for reasoning in small llms: What works and what doesn't, 2025. URL https://arxiv.org/abs/2503.16219.

Google Deepmind. Gemini 2.5: Our most intelligent AI model — blog.google, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement, 2025. URL https://arxiv.org/abs/2503.17352.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Yiran Guan, Sifan Tu, Dingkang Liang, Linghao Zhu, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Thinkomni: Lifting textual reasoning to omni-modal scenarios via guidance decoding. In *Proc. of Intl. Conf. on Learning Representations*, 2026.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming-the rise of code intelligence, 2024. URL https://arxiv.org/abs/2401.14196.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models, 2025. URL https://arxiv.org/abs/2501.03262.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Katie Kang, Amrith Setlur, Dibya Ghosh, Jacob Steinhardt, Claire Tomlin, Sergey Levine, and Aviral Kumar. What do learning dynamics reveal about generalization in llm reasoning?, 2024. URL https://arxiv.org/abs/2411.07681.

Team Kimi. Kimi k1. 5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Hang Zhang, Yuming Jiang, Xin Li, Fan Wang, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Advancing the frontiers of multimodal reasoning, 2025. URL https://github.com/LengSicong/MMR1.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2504.06958.

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. In *Proc. of Advances in Neural Information Processing Systems*, 2025a.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement, 2025b. URL https://arxiv.org/abs/2503.06520.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025c.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 6774–6786, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*, 2023.

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

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

OpenAI. OpenAI o1 System Card — openai.com, 2024. URL https://cdn.openai.com/o1-system-card.pdf.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025. URL https://arxiv.org/abs/2503.07536.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Annual Meeting of the Association for Computational Linguistics*, pp. 20023–20070, 2025.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations (ICLR)*, 2016.

ByteDance Seed. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning, 2025. URL https://arxiv.org/abs/2504.13914.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning, 2025. URL https://arxiv.org/abs/2503.20752.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *Proc. of Advances in Neural Information Processing Systems*, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Proc. of Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169, 2024.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. In *Proc. of Advances in Neural Information Processing Systems*, 2025b.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding. In *Proc. of Advances in Neural Information Processing Systems*, 2025c.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025. URL https://arxiv.org/abs/2503.10615.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search, 2024. URL https://arxiv.org/abs/2412.18319.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. In *Proc. of Advances in Neural Information Processing Systems*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Proc. of Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023.

Zheng Yaowei, Lu Junting, Wang Shenzhi, Feng Zhangchi, Kuang Dongdong, and Xiong Yuwen. Easyr1: An efficient, scalable, multi-modality rl training framework, 2025. URL https://github.com/hiyouga/EasyR1.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. In *Proc. of Advances in Neural Information Processing Systems*, volume 37, pp. 64735–64772, 2024a.

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, 2025. URL https://arxiv.org/abs/2503.12937.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Proc. of European Conference on Computer Vision*, pp. 169–186. Springer, 2024b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A PSEUDO CODE

For completeness and reproducibility, we restate the per-iteration update of DuoPO in self-contained pseudo-code form. The version below makes the data flow between the three components explicit (selector $\mathcal{S}$ in §5.1, joint weight w in §5.2, and conditional loss family $\ell(\cdot \mid c_i)$ in §5.3), and matches the higher-level Algorithm 1 in the main text.

Algorithm 2 DuoPO: detailed per-iteration pseudo code

Input: prompt set $\mathcal{Q}$; current policy π_θ; group size $G{=}2N$; sampling ratio $\alpha \in (0, 1]$; importance-resampling rounds S; weight scalars $\Delta \geq 0$, $\beta \geq 0$.
Output: per-step gradient update applied to θ.
Initialise the iteration buffer $\mathcal{B} \leftarrow \emptyset$.
for each prompt $x \in \mathcal{Q}$ **do** ▷ **Stage 0: rollout collection**
  Draw $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ and the verifiable rewards $\{r_i\}_{i=1}^{G}$.
  Set $c_i \leftarrow \mathbf{1}[y_i \text{ correct}]$, the teacher-availability gate $m_i \leftarrow \mathbf{1}[f_i \text{ available}]$, and the group-relative advantage $A_i \leftarrow (r_i - \bar{r})/(\sigma_r + \epsilon)$.
  // Stage 1: signal-axis selector $\mathcal{S}$ (Eq. equation 9).
  Sort indices so that $A_{(1)} \geq \cdots \geq A_{(G)}$ and pair $(A_{(k)}, A_{(G-k+1)})$ for $k{=}1{:}N$.
  Score each pair by combined contrast $|A_{(k)}| + |A_{(G-k+1)}|$ and retain the top $\lceil \alpha N \rceil$ pairs as $\mathcal{B}_x$.
  Append: $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_x$.
end for
// Importance resampling on the contrast magnitude.
For each pair $j \in \mathcal{B}$ with members (p, q), set $W_j \leftarrow |A_p| + |A_q|$ and $P_j \leftarrow W_j / \sum_k W_k$.
Draw S sub-batches from $\mathcal{B}$ with replacement under $\{P_j\}$ and concatenate them into $\mathcal{B}'$.
// Stage 2–3: joint weight w and state-axis loss routing.
for each $(i, t) \in \mathcal{B}'$ **do**
  Compute $H_{i,t} \leftarrow -\sum_v q_{i,t}(v) \log q_{i,t}(v)$ with $q_{i,t}(v) = \pi_\theta(v \mid x, f_i, y_{i,<t})$ when $m_i = 1$; otherwise set $H_{i,t}{=}0$.
  Set $w_{i,t} \leftarrow |A_i|^{1+\Delta(1-c_i)}\, e^{-\beta(1-c_i)H_{i,t}}$ ▷ Eq. equation 10
  Route $\ell_{i,t} \leftarrow \ell_{i,t}^{\text{GRPO}}(\theta)$ if $c_i{=}1$ or $m_i{=}0$; else $\ell_{i,t} \leftarrow \ell_{i,t}^{\text{SDPO}}(\theta)$. ▷ Eq. equation 11
end for
// Stage 4: self-normalising aggregation and gradient step.
$\mathcal{L}_{\text{DuoPO}} \leftarrow \big(\sum_{i,t} w_{i,t}\, \ell_{i,t}\big) \big/ \big(\sum_{i,t} w_{i,t}\big)$ ▷ Eq. equation 12
Update θ by one gradient-descent step on $\mathcal{L}_{\text{DuoPO}}$.

B PROMPT DESIGN

DuoPO does not modify the prompt format; we adopt the standard R1-style (Guo et al., 2025) "think-then-answer" template so that the verifiable-reward gate c_i can be computed from a well-delimited answer span. Concretely, every training and evaluation prompt asks the policy to emit its reasoning inside a single `<think> </think>` block and to place the final answer inside a `\boxed{}` expression; the original Qwen2.5-VL (Bai et al., 2025) system prompt is kept intact, and the formatting instruction is appended to the user message. Using the same template at training and evaluation time removes any train–test distribution shift on the surface form of the answer. The full instance is shown below.

Prompt Example

SYSTEM:
You are a helpful assistant.
USER:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think> </think>` tags. The final answer MUST BE put in `\boxed{}`. **<QUESTION>**

C More Experimental Results

This appendix collects (i) scaling and transfer experiments (§C.1) and (ii) extended experimental results that complement the main ablations.

C.1 Scaling to 32B and Transfer Across Tasks / Modalities

Table 7: Scaling DuoPO to a 32B backbone (Qwen2.5-VL-32B).

Setting	MathVerse	MathVision	MathVista	WeMath	HallBench	ChartQA	Avg.
Qwen2.5-32B	57.0	38.2	75.4	72.9	71.3	80.7	65.9
+ GRPO	58.4	39.3	77.3	75.9	70.3	83.8	67.5
+ DuoPO	**59.3**	**41.5**	**79.8**	**78.2**	**72.4**	**85.1**	**69.4**

Table 8: Transferring DuoPO to Referring Expression Comprehension (Qwen2.5-VL-3B, IoU soft reward).

Setting	$\text{RefCOCO}_{\text{testA}}$	$\text{RefCOCO}_{\text{testB}}$	$\text{RefCOCO+}_{\text{testA}}$	$\text{RefCOCO+}_{\text{testB}}$	$\text{RefCOCOg}_{\text{test}}$
Qwen2.5-3B	86.09	75.64	81.71	66.93	72.39
+GRPO	89.90	81.33	85.94	70.97	81.45
+DuoPO	**92.14**	**84.62**	**88.08**	**76.54**	**86.38**

Table 9: Transferring DuoPO to text-only LLM reasoning (Qwen2.5-Math-1.5B/7B-Base, trained on Open-S1).

Setting	Math12K	AIME24	MATH500	GSM8K	GPQA	Olymp.
Qwen2.5-Math-1.5B-Base						
+ GRPO	67.6	10.0	66.0	74.0	25.7	33.3
+ DuoPO	70.8	16.3	71.4	79.9	31.1	37.0
Qwen2.5-Math-7B-Base						
+ GRPO	74.6	20.0	76.4	84.8	36.3	39.7
+ DuoPO	78.6	23.6	79.8	89.8	37.6	41.7

Scaling to a 32B backbone. To confirm the gains are not an artefact of small-policy instability, we repeat the $30k$-sample recipe on Qwen2.5-VL-32B for 50 steps. Tab. 7 shows DuoPO lifts the GRPO-trained 32B baseline by 1.9 points on average, with the largest gains on MathVision ($+2.2$) and WeMath ($+2.3$). The shape of the improvement matches the 7B setting.

Transfer to a soft-reward task. Soft rewards stress-test DuoPO because correctness c_i is defined by thresholding a continuous signal rather than an exact verifier. We fine-tune Qwen2.5-VL-3B on a $60k$ subset of RefCOCO/RefCOCOg/RefCOCO+ with an IoU reward, setting $c_i = \mathbf{1}[\text{IoU} > 0.5]$. As Tab. 8 shows, DuoPO improves over GRPO on every split; the largest lift ($+5.57$) is on RefCOCO+/testB.

Transfer to text-only LLMs. Nothing in DuoPO's derivation is specific to vision tokens, so we re-train Qwen2.5-Math-1.5B/7B-Base (Yang et al., 2024) on Open-S1 (Dang & Ngo, 2025) for 150 steps (Tab. 9). DuoPO improves over GRPO on all six benchmarks at both scales; the largest gains appear on AIME24 ($+6.3$ at 1.5B, $+3.6$ at 7B), which is the benchmark most sensitive to hard-rollout credit assignment and therefore aligned with the asymmetric-utility axis DuoPO exploits.

C.2 Extended Ablations

Per-benchmark breakdown of the limited-data regime. Tab. 10 expands the "Math Avg." column of Tab. 1–2 into a per-benchmark view. With only $2.1k$ training samples, DuoPO produces a consistent shift on every column and at both scales. The 3B backbone gains more from K12 than Geo3K on vision-heavy targets (MathVista, WeMath), while the 7B backbone benefits roughly equally from

both, consistent with the observation that smaller policies are more bottlenecked by data diversity than by algorithm.

Stress test at the full 30k scale. Tab. 11 repeats the head-to-head comparison of §6 on the full $30k$-sample MM-Eureka pool at 7B. DuoPO reaches 65.0 on average, 2.5 points above GRPO, 1.6 above DAPO, and 2.2 above GSPO, with the lead maintained on every individual benchmark. The gap widens rather than shrinks at this scale, reinforcing the claim that the gains are structural.

Comparison with RLOO / REINFORCE++. Tab. 12 extends the baseline panel with RLOO (Ahmadian et al., 2024) and REINFORCE++ (Hu, 2025). Both land between GRPO and DAPO; DuoPO remains ahead on every column, with a Geo3K gap exceeding 6 points. Variance-reduction schemes that do not re-select which rollouts enter the update cannot address the near-zero concentration of $|A|$ documented in §3.3.

Prioritised experience replay. We also compare the online importance-resampling loop inside Algorithm 1 against a PER-style decoupled buffer that redraws past rollouts by the same priority. The buffer variant plateaus in training accuracy and drifts downward in validation accuracy, symptomatic of over-fitting to stale samples whose on-policy gradient approximation no longer applies. DuoPO avoids this because its resampling is computed on the freshly drawn group; priority alone is not the lever, priority *on on-policy samples* is.

Table 10: Detailed per-benchmark performance of models trained on Geometry3K and K12. Best per row-block in **bold**.

Method	MathVerse	MathVision	MathVista	WeMath	HallBench	ChartQA	Total Avg.
Qwen2.5-VL-3B	34.77	21.94	58.40	51.72	59.83	73.08	49.96
+ DuoPO (Geo3K)	43.82	25.64	62.10	64.42	63.44	77.28	56.12
+ DuoPO (K12)	**44.38**	**26.82**	**65.24**	**65.04**	**66.04**	**78.52**	**57.67**
Qwen2.5-VL-7B	42.59	25.76	67.40	63.51	65.19	79.84	57.38
+ DuoPO (Geo3K)	**51.28**	27.76	71.18	69.50	69.78	81.92	61.90
+ DuoPO (K12)	48.92	**28.94**	**73.58**	**70.08**	**70.16**	**82.88**	**62.43**

Table 11: Full $30k$-sample, 150-step comparison on Qwen2.5-VL-7B. Best per column in **bold**.

Method	MathVerse	MathVision	MathVista	WeMath	HallBench	ChartQA	Total Avg.
Qwen2.5-VL-7B	42.6	25.8	67.4	63.5	65.2	79.8	57.4
+ GRPO	50.6	28.3	74.5	69.7	70.7	81.4	62.5
+ DAPO	51.4	28.8	75.3	71.5	71.0	82.8	63.4
+ GSPO	50.8	28.2	75.3	70.1	69.7	82.9	62.8
+ DuoPO	**54.2**	**30.3**	**77.3**	**72.6**	**71.3**	**84.4**	**65.0**

Table 12: DuoPO on Geometry3K compared to two additional policy-gradient baselines (RLOO and REINFORCE++). Best per column in **bold**.

Method	Geo3K	MathVerse	MathVision	MathVista	WeMath	HallBench	ChartQA
Qwen2.5-VL-3B	25.79	34.77	21.94	58.40	51.72	59.83	73.08
+ RLOO	42.09	39.94	22.96	58.90	59.48	**64.24**	76.68
+ REINFORCE++	41.76	41.70	24.86	60.90	63.51	62.99	76.20
+ DuoPO	**48.12**	**43.82**	**25.64**	**62.10**	**64.42**	63.44	**77.28**

Extended hyperparameter sweeps. Tab. 13–14 complement §6.3 with per-benchmark accuracy across the four DuoPO hyperparameters. Raising α from 0.25 to 0.50 improves both in-domain and out-of-domain numbers; pushing to 0.75 regresses mildly, consistent with the intuition that over-retention washes out the contrast signal $\mathcal{S}$ concentrates. Shuffle count S follows an inverted-U with peak at $S{=}8$: below it, high-$|A|$ pairs are under-replayed; above it, the expected $|A|$-proportional

mass saturates and additional replays just amplify noise. Δ has a clear optimum at 0.5: moving to Δ=0 drops Geo3K by 2.10 points, while Δ>1 gradually over-amplifies noisy high-$|A|$ failures. β is flatter, with DuoPO staying within 0.6 points of its best over $\beta \in [0, 0.2]$: the entropy gate is a softener rather than a primary lever, so practitioners can safely leave it at the default and focus tuning budget on Δ.

Table 13: Per-benchmark accuracy of DuoPO across the structural hyperparameter sweeps. Best per block in **bold**.

Setting	Geo3K	MathVerse	MathVision	MathVista	WeMath	HallBench	ChartQA
Sampling ratio α (with $S = 8$, $\Delta = 0.5$, $\beta = 0.1$)							
$\alpha = 0.25$	45.98	42.70	23.38	60.58	63.80	63.26	75.58
$\alpha = 0.50$	**48.12**	**43.82**	25.64	**62.10**	64.42	**63.44**	**77.28**
$\alpha = 0.75$	47.58	41.94	**25.92**	61.12	**64.76**	63.18	73.58
Shuffle rounds S (with $\alpha = 0.5$, $\Delta = 0.5$, $\beta = 0.1$)							
$S = 4$	46.14	42.48	25.06	**62.48**	64.32	62.48	75.82
$S = 8$	**48.12**	43.82	**25.64**	62.10	**64.42**	**63.44**	**77.28**
$S = 16$	47.42	**43.96**	25.14	61.22	62.88	62.42	75.06

Table 14: Sensitivity of DuoPO to the weight-family scalars Δ and β (Qwen2.5-VL-3B, Geo3K).

Setting	Geo3K	Math Avg.	HallBench	ChartQA
Sweep of Δ (with $\beta = 0.1$)				
$\Delta = 0.0$ (separable)	46.02	47.34	63.19	76.60
$\Delta = 0.25$	47.56	48.32	63.22	76.92
$\Delta = 0.5$	**48.12**	**48.93**	**63.44**	**77.28**
$\Delta = 1.0$	47.68	48.54	63.28	76.78
Sweep of β (with $\Delta = 0.5$)				
$\beta = 0.0$ (no gate)	46.48	47.66	63.21	76.74
$\beta = 0.1$	**48.12**	**48.93**	**63.44**	**77.28**
$\beta = 0.2$	47.62	48.48	63.30	77.06

D Experiment Settings

We report details of our training and evaluation settings here, including reward function design, main hyperparameters and computing resources.

Reward Calculation. The verifiable reward used in our RL updates combines two binary signals: a format check and an answer-level correctness check. The format check $r_{\text{format}} \in \{0, 1\}$ returns 1 when the generated trace produces the required `<think></think>` reasoning block and places the final answer in `\boxed{}`, and 0 otherwise. The correctness check $r_{\text{acc}} \in \{0, 1\}$ returns 1 when the extracted final answer matches the ground truth under the task-specific grader (MathRuler for free-form items, Gemini-2.0-Flash-001 for multiple-choice items; see Appendix F) and 0 otherwise. The scalar reward passed to the RL update is a convex combination

$$r_{\text{overall}} = 0.1\, r_{\text{format}} + 0.9\, r_{\text{acc}}, \tag{13}$$

with the small weight on r_{format} chosen because format compliance saturates within the first few hundred steps and should not dominate the gradient once satisfied.

Table 15: Hyperparameter settings.

Hyperparameters	Value
max pixels	1000000
min pixels	262144
max prompt length	2048
max response length	2048
rollout batch size	512
global batch size	128
learning rate	1e-6
optimizer	AdamW
rollout temperature	1.0
rollout top p	0.99
evaluation temperature	0.5
rollout group number	16
selector retention ratio (α)	0.5
importance-resampling rounds (S)	8
asymmetric exponent (Δ)	0.5
entropy-gate sharpness (β)	0.1
KL coefficient	0.0
vision encoder	frozen

Hyperparameters. We use EasyR1 (Yaowei et al., 2025) as our training framework. Full hyperparameter settings during training is shown in Tab. 15. For experiments on Geometry3K and K12 with $\sim 2.1k$ training samples, we set the training steps to 80. For the joint training experiments ($\sim 30k$ training samples), we increase the training steps to 150 due to extended data size. Other hyperparameters that are not mentioned are kept to default values of EasyR1.

Computing Resources. All experiments are conducted on $8\times$ 80G GPUs.

E Dataset Collection

Training corpora. Our RL training pool is drawn from two multimodal geometry / STEM sources. Geometry3K (Lu et al., 2021) supplies $2{,}101$ free-form geometry problems for training and 601 held-out items for evaluation; most of the problem state is carried by the diagram rather than the text, which is the regime our two-axis view was designed for. To probe data-sensitivity beyond geometry, we also use a $2.1k$ slice from MMK12 (Meng et al., 2025) — the broader K-12 math corpus released by MM-Eureka, whose held-out split additionally spans physics, biology, and chemistry. To keep size comparisons with Geometry3K fair we take the subset already released by NoisyRollout (Liu et al., 2025a) (we refer to this slice as "K12" throughout the paper, to avoid confusion with the full 16k MMK12). For the scaled-up experiments we merge the full Geometry3K with 27k randomly sampled MM-Eureka queries into a single 30k pool. All training instances are kept in free-form answer format.

Evaluation benchmarks. We evaluate along three skill axes that jointly stress both signal-axis and state-axis behaviour: quantitative reasoning — MathVerse (Zhang et al., 2024b), MathVision (Wang et al., 2024), WeMath (Qiao et al., 2025), MathVista (Lu et al., 2023) — visual perception under adversarial distractors (HallusionBench (Guan et al., 2024)), and structured-chart understanding (ChartQA (Masry et al., 2022)). Together with the in-domain Geo3K / K12 splits, this panel is a deliberately harder out-of-distribution test than the training sources alone, letting us separate an in-domain fit from a genuine transfer of the structural signal DuoPO exploits.

F Evaluation Metrics

We adopt pass@1 accuracy as our evaluation metric. At inference we set the decoding temperature to 0.5 and perform 8 independent runs per query, reporting the average pass@1 as the final score; averaging over runs reduces variance from sampling without hurting accuracy. The evaluation prompt is identical to the training prompt, and every reproduced baseline reported in the main paper is re-evaluated under the same settings for a fair comparison.

For MathVerse, MathVision, WeMath, MathVista and ChartQA we employ Gemini-2.0-Flash-001 (Deepmind, 2025) first to extract the predicted answer from the model response and then to compare it with the ground truth. Fig. 2 and Fig. 3 show the extraction and verification prompts used with Gemini. For WeMath we report accuracy under loose mode, and for MathVerse we report overall accuracy across all sub-categories (Text Dominant, Text Lite, Vision Intensive, Vision Dominant, Vision Only).

G Theoretical Analysis of DuoPO

This appendix gives the proofs and supporting derivations for the formal claims used in the main text. We first establish that the joint weight $w(A_i, c_i, t)$ in Eq. equation 10 is genuinely *non-separable* in $(|A|, c)$ (Lemma G.1), then verify that DuoPO contains GRPO as a degenerate point (Prop. 4.1), and finally show that the prescribed select-then-route composition order strictly dominates the reversed order in expected per-pair contrast (Prop. G.2). Two intuitive corollaries on the gradient-norm amplification induced by $\mathcal{S}$ and the induced selection bias are stated at the end.

Throughout this section we assume that within a rollout group the empirical advantage A is approximately Gaussian with mean zero (the group-relative normalisation enforces zero sample mean by

Please read the following example.
Then output the answer extracted from the model response directly. No "Extracted answer:" in your answer.

Hint: Please answer the question requiring an integer answer and provide the final value,
e.g., 1, 2, 3, at the end.
Question: Which number is missing?
Model response: The number missing in the sequence is 14.
Extracted answer: 14

Hint: Please answer the question requiring a floating-point number with one decimal place and provide the final value,
e.g., 1.2, 1.3, 1.4, at the end.
Question: What is the fraction of females facing the camera?
Model response: The fraction of females facing the camera is 0.6,
which means that six out of ten females in the group are facing the camera.
Extracted answer: 0.6

Hint: Please answer the question requiring a floating-point number with two decimal places and provide the final value,
e.g., 1.23, 1.34, 1.45, at the end.
Question: How much money does Luca need to buy a sour apple candy and a butter-scotch candy? (Unit: $)
Model response: Luca needs $1.45 to buy a sour apple candy and a butterscotch candy.
Extracted answer: 1.45

Hint: Please answer the question requiring a Python list as an answer and provide the final list,
e.g., [1, 2, 3], [1.2, 1.3, 1.4], at the end.
Question: Between which two years does the line graph saw its maximum peak?
Model response: The line graph saw its maximum peak between 2007 and 2008.
Extracted answer: [2007, 2008]

Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: What fraction of the shape is blue?
Choices: (A) 3/11 (B) 8/11 (C) 6/11 (D) 3/5
Model response: The correct answer is (B) 8/11.
Extracted answer: B

QUESTION
Model response: **PREDICTION**
Extracted answer:

Figure 2: Extraction prompt for Gemini.

Below are two answers to a **math question/chart understanding question**. Question is [Question], [Standard Answer] is the standard answer to the question, and [Model_answer] is the answer extracted from a model's output to this question. Determine whether these two answers are consistent.
Please note that only when the [Model_answer] completely matches the [Standard Answer] means they are consistent. For non-multiple-choice questions, if the meaning is expressed in the same way, it is also considered consistent, for example, 0.5m and 50cm.
If they are consistent, Judgement is 1; if they are different, Judgement is 0.

[Question]: Write the set of numbers represented on the number line in interval notation.
[Standard Answer]: (-2,1]
[Model_answer]: Extracted Answer: \((-2, 1)\)
Judgement: 0

[Question]: As shown in the figure, circle O has a radius 1.0, if angle BAC = 60.0, then the length of BC is ()
Choices:
A:2 B:2√{{3}} C:√{{3}} D:2√{{2}}
[Standard Answer]: C
[Model_answer]: B:2√{{3}}
Judgement: 0

[Question]: Find the domain and range of the function f using interval notation.
[Standard Answer]: domain: [-4, 0) and range: (-3, 1]
[Model_answer]: Range: \((-4, 1]\)
Judgement: 0

[Question]: As shown in the figure, circle O has a radius 1.0, if angle BAC = 60.0, then the length of BC is ()
Choices:
(A):2 (B):2√{{3}} (C):√{{3}} (D):2√{{2}}
[Standard Answer]: (C)
[Model_answer]: C
Judgement: 1

Please output the judgement score directly with no explanation.
[Question]: **QUESTION**
[Standard Answer]: **ANSWER**
[Model_answer]: **PREDICTION**
Judgement:

Figure 3: Score prompt for Gemini.

construction), and treat $|A|$ as a non-negative scalar with finite second moment. We write $X = |A|$ where convenient.

G.1 Non-separability of the Joint Weight

Lemma G.1 (Non-separability of w)**.** *For any $\Delta > 0$, the joint weight in Eq. equation 10 admits no factorisation of the form $w(A_i, c_i, t) = \phi(A_i)\,\psi(c_i, t)$.*

Proof. Suppose, for contradiction, that $w(A_i, c_i, t) = \phi(A_i)\,\psi(c_i, t)$. Fix any token index t and any two advantage values $a_1 \neq a_2$ with $|a_1|, |a_2| > 0$. Compute the cross-ratio

$$R(a_1, a_2) \;=\; \frac{w(a_1, 0, t)\, w(a_2, 1, t)}{w(a_1, 1, t)\, w(a_2, 0, t)}. \tag{14}$$

Under a separable factorisation, the ψ terms cancel pairwise and $R \equiv 1$ for every (a_1, a_2). Substituting Eq. equation 10 and noting that the entropy factor $e^{-\beta(1-c)H_{i,t}}$ collapses to 1 when c=1 and contributes the same $e^{-\beta H_{i,t}}$ to both numerator and denominator, the cross-ratio reduces to

$$R(a_1, a_2) \;=\; \frac{|a_1|^{1+\Delta}\,|a_2|}{|a_1|\,|a_2|^{1+\Delta}} \;=\; \left(\frac{|a_1|}{|a_2|}\right)^{\Delta}, \tag{15}$$

which is identically 1 in (a_1, a_2) only when $\Delta = 0$. For any $\Delta > 0$, R varies with the ratio $|a_1|/|a_2|$, contradicting separability. □

This lemma is the formal counterpart of the asymmetric-utility observation in §4.1: the empirical ratio $U(|A|, c{=}0)/U(|A|, c{=}1)$ depends on $|A|$, so any weight that hopes to track this ratio must mix the two attributes inside a single multiplicative factor.

G.2 GRPO as a Degenerate Point of the DuoPO Objective

Proof of Proposition 4.1. Set $\mathcal{S}(\mathbf{A}) = \{1, \ldots, G\}$ (no sub-selection), $w(A_i, c_i, t) \equiv 1$, and $\ell(\cdot \mid c_i) \equiv \ell_{i,t}^{\mathrm{GRPO}}(\theta)$ in Eq. equation 12. The numerator collapses to $\sum_{i,t} \ell_{i,t}^{\mathrm{GRPO}}(\theta)$, the denominator collapses to $\sum_{i,t} 1 = \sum_i |y_i|$, and the resulting normalised sum coincides with Eq. equation 3. □

In particular, the four scalar hyperparameters introduced by DuoPO have a unique degenerate setting (α=1, S=1, Δ=0, β=0 together with disabling the SDPO branch) under which DuoPO and GRPO coincide. Every empirical gap reported in Sec. 6 therefore measures movement *away from* this corner of the design space.

G.3 Composition Order: Select-then-route vs. Route-then-select

A natural alternative to the order prescribed by Algorithm 1 is to first split rollouts by their state c_i and then form contrastive pairs within each branch. The next proposition shows that this reversal weakens the contrast that the selector is designed to exploit.

Proposition G.2 (Contrast reduction under the reversed order)**.** *Let A^+ and A^- denote the advantages of correct and incorrect rollouts within a group, respectively. Under the mild assumption that, conditional on c_i, the empirical advantage distribution lies on a single side of zero in expectation (a typical signature of a non-trivial reward distribution), the expected per-pair contrast under route-then-select is at most that under select-then-route:*

$$\mathbb{E}\big[\,|A^+_{(k)}| + |A^+_{(M-k+1)}|\,\big] \;\le\; \mathbb{E}\big[\,|A_{(k)}| + |A_{(G-k+1)}|\,\big], \tag{16}$$

and analogously for the incorrect branch, with strict inequality whenever $\operatorname{sign}(A^+)$ is constant within the branch.

Proof sketch. Pairing the k-th highest with the k-th lowest advantage in the full group spans both sides of zero, so the pair magnitude is at least $|A_{(k)}| + |A_{(G-k+1)}| \ge A_{(k)} - A_{(G-k+1)}$ which is maximised when the two members straddle zero. Within a single c-branch, the conditional distribution concentrates on one sign in expectation, so the analogous extremes-pairing inside the branch produces members on the same side of zero, halving the expected straddle. A complete argument (Cauchy–Schwarz on the per-branch order statistics) is omitted for brevity; the empirical confirmation in Tab. 6 is consistent with the predicted halving. □

G.4 Two Intuitive Corollaries on Gradient Amplification and Selection Bias

We close with two informal observations that motivate why $\mathcal{S}$ and the importance-resampling step in Algorithm 1 make a measurable difference even before the joint weight w is applied. Treat $|A|$ as a non-negative random variable with finite second moment, model $\mathcal{S}$ as a hard truncation $\mathcal{S}^{\mathcal{S}} = \{|A| \ge \tau\}$ at some threshold $\tau > 0$, and let

$$E_{\text{orig}} = \mathbb{E}[|A|], \tag{17}$$

$$E_{\mathcal{S}} = \mathbb{E}[|A| \mid \mathcal{S}^{\mathcal{S}}], \tag{18}$$

$$E_{\text{IR}} = \mathbb{E}_{p_{\text{IR}}}[|A| \mid \mathcal{S}^{\mathcal{S}}], \quad p_{\text{IR}}(a) \propto |a| \cdot p(a \mid \mathcal{S}^{\mathcal{S}}). \tag{19}$$

denote the expected advantage magnitude before sub-selection, after sub-selection only, and after both sub-selection and $|A|$-proportional importance resampling.

Corollary 1 (gradient-norm amplification). A direct calculation gives $E_{\text{IR}} = \mathbb{E}[|A|^2 \mid \mathcal{S}^{\mathcal{S}}]/\mathbb{E}[|A| \mid \mathcal{S}^{\mathcal{S}}] = E^{(2)}_{\mathcal{S}}/E_{\mathcal{S}}$, and Cauchy–Schwarz yields $E_{\text{IR}} \ge E_{\mathcal{S}}$. Decomposing the unconditional expectation into its truncated and untruncated parts shows $E_{\mathcal{S}} > E_{\text{orig}}$, hence $E_{\text{IR}} > E_{\text{orig}}$. Under the standard policy-gradient assumption that the score-function norm is approximately independent of $|A|$, the per-step expected gradient magnitude scales linearly with the expected $|A|$, so the combined effect of $\mathcal{S}$ and importance resampling strictly increases the effective per-step gradient norm. This is consistent with the faster validation accuracy curves reported in Sec. 6.

Corollary 2 (induced positive bias on signed A). The same importance-resampling distribution biases the conditional mean of the *signed* advantage. Under the Gaussian working assumption with mean zero,

$$\mathbb{E}_{p_{\text{IR}}}[A] = \frac{1}{Z}\Big(\int_{-\infty}^{-\tau} a\,|a|\,p(a)\,da + \int_{\tau}^{\infty} a\,|a|\,p(a)\,da\Big) \tag{20}$$

$$= \frac{2}{Z}\int_{\tau}^{\infty} a^2\,p(a)\,da \;>\; 0, \qquad Z = 2\int_{\tau}^{\infty} a\,p(a)\,da. \tag{21}$$

The bias points in the direction of higher-reward rollouts. This is a feature, not a bug, in the RLVR setting: it concentrates updates on the rollouts whose verifiable reward signal is strongest, and the correctness-conditioned routing of §5.3 together with the asymmetric exponent $1+\Delta(1-c_i)$ prevents the bias from over-rewarding "easy correct" rollouts at the expense of informative failures. In our reading, this is precisely the place where DuoPO's two-axis structure pays off relative to a pure $|A|$-only resampler.