SELECTIVE PARTITION REPAIR FOR GROUP-RELATIVE POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Executable verifiers provide inexpensive rewards for visual-symbolic reasoning, but false negatives can eliminate useful training signals. In group-relative policy optimization (GRPO), an all-rejected group receives no ranking signal even when it contains correct responses. Correcting individual verdicts can still give a missed correct response a negative advantage, because each advantage depends on the rewards of the entire group. We formulate this problem as hidden-partition recovery and introduce *Structured Group-Relative Correction* (SGRC). SGRC learns a joint distribution over binary correctness masks from judge-annotated rollout groups. Its energy combines evidence from each response with interactions between the predicted correct and incorrect subsets. On subsequent groups, a confidence gate evaluates each complete predicted mask before applying it, leaving uncertain groups unchanged. The repaired rewards enter the original GRPO normalizer without changing the functional form of the policy objective. On the `geometry_sym` subset of MathV360K, the complete SGRC system, using Qwen2.5-VL-3B-Instruct and 6,912 annotation attempts per run, achieves mean strict answer accuracy 2.96 percentage points above fixed-verifier GRPO across five paired training seeds. On held-out mixed groups, its joint predictor achieves exact-mask recovery 16.7 points above a parameter-matched factorized predictor. These results motivate evaluating group-relative reward repair by the complete partition it recovers, rather than only by individual verdict accuracy.

1 INTRODUCTION

Reinforcement learning with executable verifiers offers a practical way to improve reasoning models: a policy generates candidate solutions, a program checks their answers, and the resulting rewards provide a training signal without annotating every response (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). For visual-symbolic reasoning, this approach connects free-form solutions to mechanically checkable answers in geometry and diagram problems. Its effectiveness depends on whether the verifier recognizes the answers the policy produces. Equivalent expressions, formatting variants, and extraction failures can turn valid responses into false negatives (Xu et al., 2026b).

The effect of such errors depends on how the optimizer uses rewards. GRPO normalizes the rewards of several rollouts for the same question to determine which responses to reinforce. If the verifier rejects every rollout, all normalized advantages are zero. This removes no useful ranking when all responses are wrong, but it also discards the signal when some responses are correct and unrecognized. The relevant target is therefore the partition between correct and incorrect responses within the group.

Correcting individual verdicts does not by itself ensure that this partition is recovered. Consider a group in which two responses are correct but both receive zero reward. Repairing only one makes the reward vector nonconstant, yet group normalization gives the other correct response a negative advantage. Partial repair can recover reward variance while distorting the ordering it was intended to restore. Figure 1 contrasts leaving the group unchanged, repairing responses separately, and predicting the complete mask before deciding whether to act, alongside a geometry example.

Existing approaches address parts of this problem. Learned verifiers improve per-response decisions (Zhang et al., 2025b; Xu et al., 2026b; Liu et al., 2025a); dynamic sampling replaces groups with constant rewards (Yu et al., 2025); and alternative advantage rules retain learning signals from verifier-negative groups (Nan et al., 2025; Nie et al., 2026). These choices do not explicitly align

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

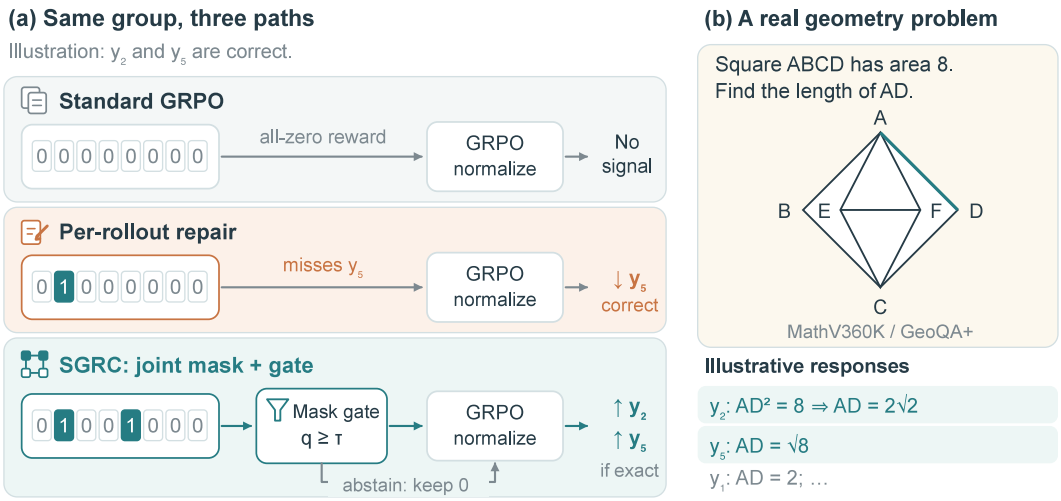

Figure 1: **Recover the group partition before normalization.** (a) Standard GRPO, per-rollout repair, and SGRC operate on the same all-rejected group. Partial repair can suppress a missed correct response; an accepted exact mask restores the reference ordering. (b) A geometry question from MathV360K (GeoQA+ image 1448; Shi et al., 2024) in which y_2 and y_5 give equivalent correct answers. The diagram is redrawn from the source; responses and repair paths are illustrative. SGRC retains the original rewards when its gate abstains.

repair confidence with recovery of the entire group partition. A factorized predictor can output a complete mask and gate its probability, but its joint probability remains a product of marginals. This is restrictive when responses share answer content and correctness labels depend on one another.

We introduce **Structured Group-Relative Correction (SGRC)** to recover this hidden partition. A group-interaction energy scores complete correctness masks using both rollout-level evidence and statistics of the proposed partition. SGRC learns from judge-annotated groups and calibrates the probability of complete-mask recovery on separate prompts. For subsequent all-rejected groups, it supplies a confident mixed mask to the GRPO normalizer and otherwise retains the verifier reward. Prediction, confidence, and intervention are all defined on the complete group partition. Removing the interaction term recovers a factorized predictor. Our parameter-matched baseline retains the same inputs and annotation protocol while reallocating capacity to the item network.

We evaluate SGRC on the `geometry_sym` subset of MathV360K (Shi et al., 2024), which contains symbolic-answer geometry problems from GeoQA+, UniGeo, and Geometry3K. Across five paired training seeds with Qwen2.5-VL-3B-Instruct, the complete system achieves mean strict answer accuracy 2.96 percentage points above fixed-verifier GRPO. It also exceeds the factorized predictor with the same gate in every seed, and its joint predictor's exact-mask recovery on held-out mixed groups is 16.7 points higher. Component comparisons examine joint prediction, selective application, and rollout features. A random-partition control retains little of the accuracy gain, consistent with the benefit depending on which responses receive the repaired rewards.

Our contributions are:

- We formulate verifier repair in all-rejected GRPO groups as hidden-partition recovery. Per-response correction can restore reward variance while suppressing missed correct responses, leaving the group-level training signal incorrectly ordered.
- We propose SGRC, which jointly predicts the correctness partition and uses complete-mask confidence to select reward repairs before the original GRPO normalization.
- We evaluate SGRC on MathV360K's `geometry_sym` subset through policy training and held-out mask recovery. The complete system achieves mean strict accuracy 2.96 points above fixed-verifier GRPO, and its joint predictor achieves mixed-mask recovery 16.7 points above the factorized baseline.

2 Related Work

RLVR and group-relative optimization. GRPO was introduced for mathematical reasoning and later scaled in DeepSeek-R1 and DAPO (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). Analyses expose its optimization and base-model biases (Liu et al., 2025b; Shao et al., 2026), and Visual-RFT, R1-VL, and DEVA adapt it to vision-language tasks (Liu et al., 2025c; Zhang et al., 2025a; Das et al., 2026). DAPO's dynamic sampling drops zero-variance groups and refills the batch; for a hidden mixed group, this matches the standard skip, since its missed correct rollouts are never reinforced, while extra generation changes compute and the training distribution. By contrast, SGRC keeps the sampler and recovers the partition of existing rollout groups. NGRPO and fixed-reference sign advantages treat verifier-negative rollouts as wrong (Nan et al., 2025; Nie et al., 2026), so false negatives can suppress hidden correct responses; SGRC instead estimates whether an all-rejected group contains a hidden correct subset.

Imperfect verifiers and learned verification. Learned verifiers score or generate verification decisions (Zhang et al., 2025b), and CompassVerifier trains a unified verifier for formulas and free-form answers (Liu et al., 2025a). TinyV rechecks rule-based negatives with a small model (Xu et al., 2026b). These methods improve the per-response verdict. Group-wise verification also appears in software agents: R4P jointly assesses candidate patches and identifies their correct subset (Xu et al., 2026a). SGRC focuses on all-rejected GRPO groups, predicting a complete mask and gating on confidence in its correctness before supplying the accepted mask to the original normalizer. Its interface accepts binary rewards from either a rule-based or a learned verifier. Audits expose verifier bugs and reward hacking (Ray, 2026; Wang et al., 2026), and controlled studies probe how the rate and structure of label noise shape RLVR dynamics (Plesner et al., 2026; Zhu & Kang, 2026; Egashira et al., 2026); SGRC targets group-conditional false negatives.

Correcting rewards and advantages. BiCC conditions responses on the observed positive–negative partition, while RCC adjusts the advantage baseline using reward–confidence covariance (Li et al., 2026b). GraphAE propagates rewards over a graph derived from reward-model representations (Li et al., 2026a). Noise-corrected GRPO estimates reward-flip rates to debias policy updates (El Mansouri et al., 2026). PGBC and PGFC likewise correct learning under an estimated verifier-noise channel, with online appeals for false-negative-rate estimation (Cai et al., 2025). SGRC learns a different object: the hidden binary partition of an individual all-rejected group. It selectively replaces the reward vector while retaining the original normalizer and objective.

Reward adaptation, structured prediction, and selective prediction. Reward models can be adapted to policy shift (Huang et al., 2026). SGRC likewise refreshes its correction model as the policy changes, here for binary correctness masks in all-rejected groups. In multilabel prediction, the joint mode is the appropriate target for exact-mask loss under label dependence (Dembczyński et al., 2012); set models give equivariant scorers (Zaheer et al., 2017; Lee et al., 2019); and selective prediction supplies abstention and finite-family risk control (Geifman & El-Yaniv, 2019; Angelopoulos et al., 2025). SGRC combines these tools so that the prediction target, confidence event, and reward intervention all refer to the complete correctness partition.

3 Method

SGRC predicts the partition of correct and incorrect responses before GRPO assigns relative advantages. A joint predictor proposes the mask, and a confidence gate decides whether to apply it. Figure 2 shows their training and application.

3.1 Problem Definition

Given an image and question x, the policy π_θ samples G responses $\mathcal{Y} = (y_i)_{i=1}^{G}$, each with a committed answer. These form a group $\mathcal{G} = (x, \mathcal{Y})$. An executable verifier assigns binary rewards $v_i = V(x, y_i)$, collected in $\mathbf{v} = (v_1, \ldots, v_G)$. GRPO converts a reward vector $\mathbf{s}$ into relative advantages through

$$\mathcal{N}_\epsilon(\mathbf{s})_i = \frac{s_i - \bar{s}}{\mathrm{std}_{\mathrm{pop}}(\mathbf{s}) + \epsilon}, \tag{1}$$

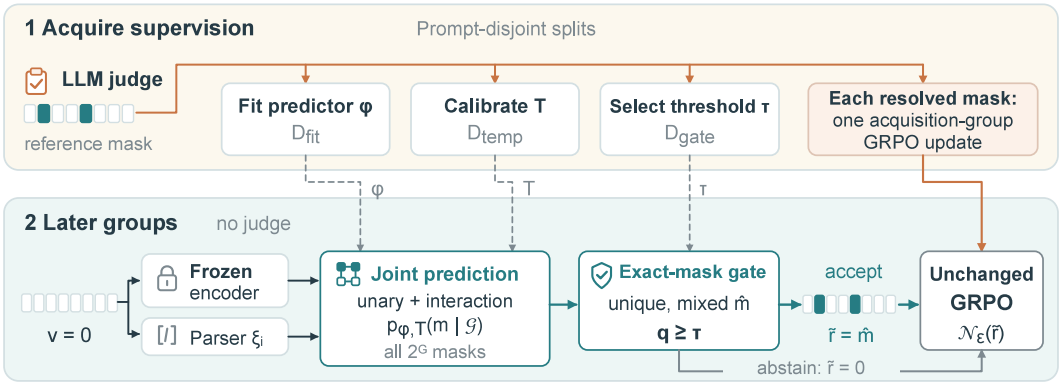

Figure 2: **Learning a partition and deciding when to apply it.** At each acquisition checkpoint, prompts are assigned before annotation to disjoint predictor-fit, temperature-calibration, and gate-selection splits. Each resolved mask also supplies one acquisition-group update. Later groups use a frozen encoder and parser traces without querying the judge. The gate either accepts a predicted mixed mask or preserves zero reward; both paths use the original GRPO normalizer.

where $\bar{s}$ is the group mean, $\mathrm{std}_{\mathrm{pop}}$ is the population standard deviation, and $\epsilon > 0$ stabilizes the denominator. An *all-rejected* group has $\mathbf{v} = \mathbf{0}$, so every reward-derived advantage is zero. SGRC repairs these groups while keeping the verifier and normalizer unchanged.

A reference annotator supplies a complete binary correctness mask $\mathbf{m}^\star$, or abstains if any response cannot be judged. A *mixed* mask contains both correct and incorrect responses; empty and full masks yield zero advantages. For a mixed mask with k positives, $0 < k < G$, the positive and negative advantages tend to $\sqrt{(G-k)/k}$ and $-\sqrt{k/(G-k)}$, respectively, as $\epsilon \to 0$. Their magnitudes depend on the partition of the entire group. If a predicted mixed mask misses a correct response, that response receives a negative advantage in place of its positive reference advantage. Recovering the complete mask reproduces all reference advantages under Eq. 1; the resulting policy update also depends on policy gradients and clipping.

The task is to recover this hidden partition on later *unqueried* groups from annotated groups, abstaining when an uncertain prediction could introduce misleading advantages. Reference correctness refers to the judge's committed-answer labels; final policy accuracy is measured separately by a strict grader.

3.2 Joint Prediction of the Complete Mask

SGRC scores partitions using evidence shared among responses to the same question. A frozen checkpoint encoder provides a stop-gradient representation $\mathbf{z}_i$ for each rollout, and the reward-time parser provides a trace $\boldsymbol{\xi}_i$. Shared maps combine them into item features $\mathbf{a}_i$, their mean group context $\mathbf{g}$, and context-aware embeddings $\mathbf{h}_i$. Only rollout features and parser traces enter the predictor; the dataset's published answer and judge outputs are excluded. The pooled context lets each response be represented relative to the other candidates.

Each candidate mask $\mathbf{m} \in \mathcal{M} = \{0,1\}^G$ separates the responses into predicted correct and incorrect sets. Let $u_i = \mathbf{w}^\top \mathbf{h}_i$ denote the evidence for response i. For each mask, we form $\mathbf{d}_\mathbf{m}$ from the group context, the mean embeddings of the two sets, their contrast, and the mask cardinality. The mask score combines individual evidence with an interaction term that evaluates this partition:

$$e_\phi(\mathbf{m}; \mathcal{G}) = \sum_i m_i u_i + \rho_\phi(\mathbf{d}_\mathbf{m}), \qquad p_\phi(\mathbf{m} \mid \mathcal{G}) = \frac{\exp e_\phi(\mathbf{m}; \mathcal{G})}{\sum_{\mathbf{m}' \in \mathcal{M}} \exp e_\phi(\mathbf{m}'; \mathcal{G})}. \tag{2}$$

We train the predictor by minimizing the complete-mask negative log-likelihood $-\log p_\phi(\mathbf{m}^\star \mid \mathcal{G})$. Shared maps and symmetric pooling preserve mask probabilities when responses and mask bits are permuted together. Empty and full masks remain in the support, allowing the model to recognize groups whose reference rewards also provide no ranking signal.

Changing one mask bit alters both its individual term and the pooled evidence for the partition. SGRC can thus distinguish masks with similar sums of individual scores: group context informs each response's score, while ρ_ϕ evaluates the proposed assignment.

Relationship to per-rollout prediction. When $\rho \equiv 0$, the normalizing sum is $\prod_i (1 + e^{u_i})$, giving

$$p_\phi(\mathbf{m} \mid \mathcal{G})\big|_{\rho \equiv 0} = \prod_i \sigma(u_i)^{m_i} (1 - \sigma(u_i))^{1 - m_i}. \tag{3}$$

The factorized predictor uses group context but predicts labels independently given its features. We widen its item network to match SGRC's parameter count, retaining the same features, data, and gate. All predictors enumerate the full mask space: with $G = 8$, 256 candidates allow exact normalization and maximum-probability prediction without merging duplicate answers.

3.3 Selective Application through an Exact-Mask Gate

One incorrect label changes every advantage in the group. The gate therefore evaluates confidence in the complete prediction. We fit a scalar temperature $T > 0$ on a separate temperature-calibration split (Guo et al., 2017), producing $p_{\phi,T}(\mathbf{m} \mid \mathcal{G}) = \mathrm{softmax}_{\mathbf{m}}(e_\phi(\mathbf{m}; \mathcal{G})/T)$. The proposed mask $\widehat{\mathbf{m}}$ is the most probable mask, with confidence $q = p_{\phi,T}(\widehat{\mathbf{m}} \mid \mathcal{G})$. SGRC abstains if two masks tie for the maximum or the proposed mask is empty or full. For a factorized predictor, q is the product of the calibrated confidences assigned to its predicted labels.

An independent gate-selection split determines the threshold τ. At each threshold on a grid specified before labeling, we count accepted groups and complete-mask errors. We choose the threshold with the largest coverage among those whose finite-sample upper risk bound is at most α and whose coverage is at least $c_{\min}$. If none qualifies, $\tau = \infty$ disables correction until the next refresh. This bound applies to resolved annotations under the checkpoint distribution; a separate runtime audit assesses error as the live policy changes.

For an all-rejected group, an accepted mixed mask replaces the zero reward vector: $\widetilde{\mathbf{r}} = \widehat{\mathbf{m}}$. In all other cases, $\widetilde{\mathbf{r}} = \mathbf{v}$. The trainer normalizes $\widetilde{\mathbf{r}}$ once and performs its usual GRPO update, retaining the loss, clipping, KL term, sampler, and optimizer. An exact prediction recovers the reference advantages; abstention preserves the trainer's original update or skip.

3.4 Training with Periodic Refresh

As the policy changes, its responses can differ from those used to fit the predictor. SGRC periodically refreshes the correction model. At each refresh, it freezes the current policy as the encoder and requests annotations for B naturally occurring all-rejected groups. Prompts are assigned to disjoint predictor-fit, temperature-calibration, and gate-selection splits before labeling. Each resolved mask enters its assigned split and is also used once in an ordinary GRPO update on its acquisition group. An unresolved annotation supplies neither a label nor an update but still counts toward B.

The encoder, predictor, temperature, and threshold remain unchanged within each segment as the live policy trains. At the next checkpoint, the current policy replaces the encoder, and the predictor, temperature, and gate are refitted using only the new annotations. Subsequent groups use the predictor and gate without querying the judge, transferring supervision beyond the annotated groups.

4 Experiments

4.1 Experimental Setup

Dataset and splits. We use `geometry_sym`, a symbolic-answer geometry subset derived from the public MathV360K dataset (Shi et al., 2024). The subset combines 2,827 image–question pairs from GeoQA+, UniGeo, and Geometry3K. A seeded shuffle reserves 500 items for held-out policy evaluation, D_{acc}. Removing items whose image paths overlap with this set leaves 1,600 training items and 219 items in an auxiliary diagnostic split. SGRC uses disjoint predictor-fit, temperature-calibration, and gate-selection splits.

Algorithm 1 SGRC within one training segment

Freeze checkpoint encoder; attempt B annotations on all-rejected groups.
Fit ϕ, then T, then τ on the predictor-fit, temperature-calibration, and gate-selection splits; freeze all three.
Apply each resolved reference mask once to its acquisition-group update.
for each later group $\mathcal{G}$ sampled by the live policy **do**
    $\mathbf{v} \leftarrow V(\mathcal{G})$; $\tilde{\mathbf{r}} \leftarrow \mathbf{v}$
    if $\mathbf{v} = \mathbf{0}$ **then**
        Predict the complete-mask distribution using frozen features.
        if mode $\hat{\mathbf{m}}$ is unique and mixed, and $q \geq \tau$ **then**
            $\tilde{\mathbf{r}} \leftarrow \hat{\mathbf{m}}$
    Apply the original GRPO update using $\mathcal{N}_\epsilon(\tilde{\mathbf{r}})$.

We train Qwen2.5-VL-3B-Instruct (Bai et al., 2025) with GRPO and $G = 8$ rollouts per question. Predictor evaluation uses D_{mask}, a separate collection of 1,000 all-rejected rollout groups: 900 empty, 90 mixed, and 10 full reference masks. Its prompts are disjoint from the predictor-fit, temperature-calibration, and gate-selection splits, and from D_{acc}. Each group contains eight generated responses.

Metrics. The primary policy metric is *strict committed-answer accuracy*: the fraction of held-out questions whose final answer is accepted by `math_verify`, using one greedy response per question. We report the mean and sample standard deviation over five paired training seeds. Predictor evaluation measures *exact-mask accuracy before gating*: every predicted label must match its reference label. We emphasize mixed groups, since an empty-mask prediction is already correct on most of D_{mask}. For selective correction, *error rate* (selective error) is the fraction of accepted mixed masks that differ from the judge's complete mask, and *coverage* is the accepted fraction of all evaluated groups. Coverage measures how often a method acts and is interpreted together with its error rate. The operating threshold is selected on the gate-selection split before either metric is evaluated on D_{mask}. Strict answer acceptance and judge-mask agreement measure different aspects of the system; neither replaces an independent assessment of semantic correctness.

Annotation and comparison protocol. A two-stage judge first solves the image question and then labels an anonymous, randomly ordered group of rollouts. Any unresolved verdict causes whole-group abstention; blinded human audits assess reference quality. All sparse-reference systems use the same judge and candidate-prompt order, with 6,912 annotation attempts per run, including unresolved cases. SGRC distributes these attempts across training, whereas static SGRC makes them all at initialization. Each resolved mask supplies one acquisition-group update. The main comparisons share the model, prompt pool, sampling, optimizer schedule, and final grader. Sampled groups diverge as the policies receive different updates. Equal annotation counts control access to reference supervision; computation varies across systems. Fixed-verifier GRPO uses no annotations, whereas the dense comparator queries every eligible runtime group. Implementation details and per-seed results are provided in the appendix.

4.2 End-to-End Performance

On MathV360K's `geometry_sym` subset, SGRC achieves the highest mean strict accuracy among the sparse-reference methods in Table 1. Its mean is 2.96 percentage points above fixed-verifier GRPO, with a 95% paired t interval of [1.83, 4.09]. The most direct structural comparison is the parameter-matched factorized predictor with the same gate; SGRC exceeds it in every paired seed, with a mean paired difference of +1.36 points (95% interval [0.67, 2.05]). Direct sparse supervision also exceeds the fixed-verifier baseline, while systems that learn corrections for subsequent groups show larger mean gains. These comparisons also reflect the resulting differences in policy trajectories and acquired supervision.

The DAPO sampling control replaces constant-reward groups, whereas SGRC attempts to recover the partition of groups already sampled. The dense-judge comparator directly annotates every eligible runtime group and attains higher accuracy, but this reference access places it outside the sparse-reference comparison. End-to-end computation is not equalized across these systems.

Table 1: **Strict held-out accuracy after policy training** (%, mean ± sample SD over five paired seeds). Sparse-reference rows match annotation attempts. Fixed-verifier GRPO uses no annotations; DAPO changes sampling cost and the dense comparator queries every eligible runtime group. In all tables, gray marks SGRC; bold marks the best scores within each comparison block.

System	Use of reference information	Accuracy
Fixed-verifier GRPO	No annotations	20.0 ± 0.8
Direct sparse supervision	Acquisition-group updates only	20.5 ± 0.8
Continuous correction	Learned residual reward	20.9 ± 0.8
Rollout-local + exact gate	Per-rollout prediction	21.6 ± 0.6
Factorized + exact gate	Group context; independent labels	21.6 ± 0.7
SGRC	Joint mask prediction + exact gate	$\mathbf{23.0 \pm 1.0}$
DAPO sampling control	No annotations; extra generation	20.7 ± 0.8
Dense-judge comparator	Direct judge access at runtime	$\mathbf{23.7 \pm 1.1}$

Table 2: **Mask recovery and selective correction on** D_{mask} (%). Exact-mask accuracy is measured before gating; error rate and coverage describe applied mixed masks. Mixed groups contain both correct and incorrect responses.

Predictor and gate	Exact, all ↑	Exact, mixed ↑	Error rate ↓	Coverage
Always empty	90.0	0.0	–	0
Rollout-local + exact gate	91.3	42.2	12.5	4.0
Factorized + exact gate	92.1	55.6	6.7	4.5
Joint, no gate	**94.7**	**72.2**	42.0	11.2
Joint + exact gate (SGRC)	**94.7**	**72.2**	**2.0**	5.0

4.3 Ablation Study

We organize the ablations around the decisions made by SGRC: how to represent a correction, when to apply it, which features to use, and when to refresh supervision. Offline comparisons use the same held-out groups; policy-training comparisons use paired seeds but can follow different trajectories.

Joint mask prediction. The factorized baseline retains group context and matches the joint model's total parameter count, while removing the interaction term ρ_ϕ. On the shared mixed groups, the joint predictor achieves 72.2% exact-mask recovery, compared with 55.6% for the factorized predictor (Table 2). The largest gap is on masks with two positives; neither predictor recovers masks with three or more positives in this pool. This comparison contrasts joint partition scoring with independent labels, although the two models allocate capacity differently.

Exact-mask gating. Holding the joint predictor fixed, the gate reduces the error rate among applied masks while accepting fewer groups (Table 2). In the training ablation (Table 3), the joint model without a gate does not attain the accuracy of gated SGRC. The two components address different decisions: the predictor proposes the partition, and the gate restricts unreliable interventions. The observed means are higher with gating within each predictor family and higher for the joint predictor within each gate condition; the predictor–gate interaction remains unresolved.

Stable rollout features. Table 4 changes only the feature source while retaining the acquisition schedule, predictor, and gate. Segment-frozen features have the highest observed mean policy accuracy and mixed-mask recovery, and the lowest observed error rate. Live-policy features drift after the predictor is fitted. An initial encoder avoids this within-segment drift but does not track later policy representations; its predictor and gate still refresh at each checkpoint. Trace-only inputs remove rollout representations altogether. The ordering of the observed means favors a stable, periodically refreshed representation. The feature comparison also changes the available information, so representation drift is one possible explanation rather than an isolated cause.

Table 3: **Predictor and gate ablation.** Strict accuracy (%, mean ± sample SD over five paired seeds). Both predictor families use the same feature inputs and annotation protocol. Bold marks the best score in each column.

Predictor	Without gate	With exact-mask gate
Group-context factorized	20.5 ± 0.6	21.6 ± 0.7
Joint	$\mathbf{21.0 \pm 0.7}$	$\mathbf{23.0 \pm 1.0}$

Table 4: **Feature-source ablation** (%). Policy accuracy is the mean over five seeds; the remaining metrics use D_{mask}. Initial-snapshot features are fixed throughout training, but the predictor, temperature, and gate still refresh.

Feature source	Policy accuracy ↑	Exact, mixed ↑	Error rate ↓	Coverage
Live policy	22.1	67.8	6.2	4.8
Initial snapshot	21.6	62.2	4.8	4.2
Parser trace only	21.2	56.7	7.7	3.9
Segment-frozen (SGRC)	**23.0**	**72.2**	**2.0**	5.0

Acquisition refresh and partition information. Static SGRC makes the same number of annotation attempts at initialization and reaches $21.9 \pm 0.6\%$ strict accuracy. Periodic refresh has higher mean accuracy, although this system comparison combines changes in acquisition timing and representation refresh. The random-partition control keeps the acceptance rule and predicted mask cardinality but places the positive labels uniformly at random. It reaches $20.4 \pm 0.7\%$ while activating a similar share of eligible groups. This control retains little of SGRC's gain, consistent with the benefit depending on which responses are reinforced, not only on restored reward variance. As with the other training comparisons, its policy trajectory diverges once the interventions differ.

4.4 Partition Recovery and Training Signal

Recovery across partition sizes. Figure 3a separates mixed groups by the number of reference-positive rollouts. The largest gap between the joint and factorized predictors occurs in groups with two positives, where independent decisions often fail to recover the complete mask. Both predictors perform better on groups with a single positive, and neither exactly recovers masks with three or more positives in this pool. Most successful repairs therefore concern small correct subsets. A single missed or extra positive makes the whole mask incorrect; larger subsets are recovered less often in this pool. Determining the sources of these prediction errors would require additional error annotations.

Selective error beyond one threshold. Figure 3b compares selection behavior using the largest threshold for each predictor that reaches a common minimum coverage target. At each reported target, the selected joint operating point has the lowest observed error; actual coverage can differ across predictors because of score ties. Error rises as the target increases. At the lowest target, joint prediction accepts no erroneous masks in this sample. The sweep describes selection behavior on held-out groups; training uses the threshold chosen on the independent gate-selection split.

Recovered advantages and abstention. Table 5 examines the signal passed to GRPO on mixed groups only. Advantage MAE is the absolute difference from the reference-normalized advantage, averaged first over rollouts and then over groups. Pair-order accuracy pools all reference-positive/negative rollout pairs and counts the fraction whose predicted advantages have the correct strict order; ties count as errors. An abstention therefore incurs error on these mixed groups even when it avoids a risky intervention. Evaluating these metrics on all of D_{mask}, where empty masks dominate, would obscure this cost.

SGRC has lower advantage MAE and higher pair-order accuracy than the rollout-local and factorized predictors that also use the exact-mask gate. Continuous correction preserves more pair orders overall, and the pointwise confidence rule has lower MAE. The latter thresholds individual rollout confidences rather than complete-mask confidence. SGRC nevertheless exceeds it in final policy accuracy in all five paired seeds, with a mean gain of +1.56 points. Neither diagnostic alone matches the ordering of

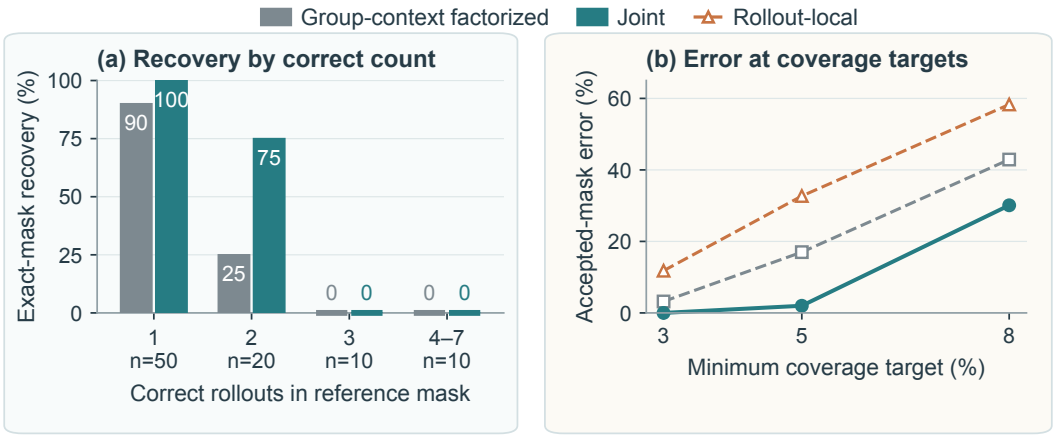

Figure 3: **Partition recovery and selective error.** (a) Ungated exact-mask accuracy by reference-positive count; n is the number of groups. (b) Error rate at common minimum coverage targets; score ties can change actual coverage. Lines connect recorded points; training gates are selected separately.

Table 5: **Optimizer-signal diagnostics on the 90 mixed groups.** Constant rewards yield zero advantages. Metrics evaluate all mixed groups, including abstentions, and are descriptive; each system applies its own acceptance rule.

System	Advantage MAE ↓	Pair-order accuracy (%) ↑
Always empty (verifier)	0.777	0.0
Continuous correction	0.436	**74.0**
Pointwise confidence rule	**0.283**	66.5
Rollout-local + exact gate	0.520	27.3
Factorized + exact gate	0.469	32.8
SGRC	0.417	38.3

final policy accuracy across these systems. Offline reconstruction quality does not by itself account for the observed benefit of selective repair during training.

4.5 Reference Quality and Runtime Behavior

On blinded mixed groups, judge and human masks agree exactly in 30 / 35 cases. A separate runtime audit compares applied masks with judge labels after the policy has moved beyond its acquisition checkpoint. Counting unresolved audited groups as errors gives an error rate of 5.7%.

SGRC adds a frozen-encoder pass to eligible runtime groups. Measured throughput is 0.90× relative to fixed-verifier GRPO, excluding acquisition and judge calls. SGRC acts selectively: most all-rejected groups remain unchanged (Table 12, Appendix E).

5 Conclusion

Group normalization can suppress a missed correct response even after another is repaired. SGRC predicts the complete correctness partition and selects repairs through complete-mask confidence. On MathV360K's `geometry_sym` subset, the joint predictor achieves higher mask recovery than the factorized predictor, and SGRC improves mean strict accuracy over fixed-verifier GRPO by 2.96 points and also exceeds its ungated counterpart. Experiments use Qwen2.5-VL-3B-Instruct with eight rollouts per group; transfer across models and domains remains open. The complete partition connects reward repair to GRPO's group-relative advantages.

AI USE STATEMENT

LLM-based writing assistance was used only for language polishing and grammatical revision.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe the method and training procedure; Section 4.1 defines the data splits, metrics, and comparison protocol. The appendix provides implementation and gate-selection details, annotation procedures, per-seed results, and audit measurements. We will release code, judge prompts, split manifests, acquisition logs, and reference masks with their labeled rollouts.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers. *arXiv preprint arXiv:2510.00915*, 2025.

Debasmit Das, Munawar Hayat, and Fatih Porikli. DEVA: Fine-tuning multimodal large language models for visual perception tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39141–39151, 2026.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1–2):5–45, 2012. doi: 10.1007/s10994-012-5285-8.

Kazuki Egashira, Mark Vero, Jasper Dekoninck, Florian E. Dorner, Robin Staab, and Martin Vechev. Delay, plateau, or collapse: Evaluating the impact of systematic verification error on RLVR. In *Proceedings of the Conference on Language Modeling (COLM)*, 2026.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected GRPO: From noisy rewards to unbiased gradients. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean

Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuefeng Xiao, Hongyan Xie, Huaqiu Li, Songshi Liang, Zhongxiang Dai, Fuzhen Zhuang, Jianxin Li, Yikun Ban, and Deqing Wang. Real-time aligned reward model beyond semantics. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Guozheng Li, Xiyan Fu, and Yiwen Guo. Representation-aware advantage estimation: Your reward model provides more than a scalar output. *arXiv preprint arXiv:2606.10528*, 2026a.

Yu Li, Tian Lan, and Zhengling Qi. When right meets wrong: Bilateral context conditioning with reward-confidence correction for GRPO. *arXiv preprint arXiv:2603.13134*, 2026b.

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F. Wong, Songyang Zhang, and Kai Chen. CompassVerifier: A unified and robust verifier for LLMs evaluation and outcome reward. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 33466–33494, 2025a. doi: 10.18653/v1/2025.emnlp-main.1698.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025b.

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 (ICCV)*, pp. 2034–2044, 2025c.

Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, and Xingzhong Xu. NGRPO: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*, 2025.

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.

Andreas Plesner, Francisco Guzmán, and Anish Athalye. An imperfect verifier is good enough: Learning with noisy rewards. *arXiv preprint arXiv:2604.07666*, 2026.

Jaideep Ray. Before the model learns the bug: Fuzzing RLVR verifiers. *arXiv preprint arXiv:2606.01066*, 2026.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

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.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4663–4680, 2024. doi: 10.18653/v1/2024.findings-emnlp.268.

Xuekang Wang, Zhuoyuan Hao, Shuo Hou, Hao Peng, Juanzi Li, and Xiaozhi Wang. Reproducing, analyzing, and detecting reward hacking in rubric-based reinforcement learning. *arXiv preprint arXiv:2606.04923*, 2026.

Junjielong Xu, Boyin Tan, Xiaoyuan Liu, Chao Peng, Pengfei Gao, and Pinjia He. Scalable supervision for software agents via patch reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026a. URL https://arxiv.org/abs/2510.22775.

Zhangchen Xu, Yuetai Li, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Basel Alomair, Bill Yuchen Lin, and Radha Poovendran. TinyV: Reducing false negatives in verification improves RL for LLM reasoning. *Transactions on Machine Learning Research*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244, 2025. doi: 10.52202/085713-3775.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, 2025a.

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

Yuxuan Zhu and Daniel Kang. Noisy data is destructive to reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2603.16140*, 2026.

A Implementation Details

Predictor inputs and maps. For rollout i, the frozen encoder supplies $\mathbf{z}_i$ from the final hidden layer, mean-pooled over the committed-answer span. The parser trace $\boldsymbol{\xi}_i$ contains answer length, parse status, numeric flag, symbolic flag, and stop reason; it excludes the verifier verdict, published answer, and all judge-derived fields. Shared maps compute

$$\mathbf{a}_i = \mathrm{LN}(W_z\mathbf{z}_i + W_\xi\boldsymbol{\xi}_i + \mathbf{b}), \qquad \mathbf{g} = \frac{1}{G}\sum_i \mathbf{a}_i, \qquad \mathbf{h}_i = \mathrm{MLP}([\mathbf{a}_i, \mathbf{g}, \mathbf{a}_i - \mathbf{g}]). \tag{4}$$

For a mask with k positives, the mean embeddings of the predicted correct and incorrect sets divide their sums by $\max(k, 1)$ and $\max(G-k, 1)$, respectively. Their concatenation with the group context, their contrast, k/G, and indicators of an empty or full mask gives $\mathbf{d_m}$. An MLP maps this vector to the interaction term $\rho_\phi(\mathbf{d_m})$ in Eq. 2.

Table 6: **Predictor configuration.** The base MLP architecture does not specify the widened item MLP of the parameter-matched factorized baseline.

Component	Setting
Dimensions $d_z/d_\xi/d_a/d_h$	2,048 / 16 / 256 / 256
Base MLP architecture	2 layers / width 256 / GELU / dropout 0.1
Parameters: joint / factorized / local	2.10M / 2.10M / 0.72M
Optimizer; learning rate; epochs; early stopping	AdamW; 0.0001; 10 epochs; early stop after 2

Predictor controls. The rollout-local baseline uses $u_i = f(\mathbf{a}_i)$ with no group context. The factorized baseline uses contextual features and a widened item MLP, with $\rho \equiv 0$. It matches the reported total parameter count of SGRC but allocates that capacity differently. Both baselines use the same fit, temperature-calibration, and gate-selection protocol. Under factorization the complete-mask likelihood is the product of Bernoulli likelihoods, so the loss reduces to per-rollout binary cross-entropy and the confidence is $q = \prod_i \sigma(|u_i|/T)$. Shared maps and symmetric pooling make the predictors equivariant to permutations of the input rollouts. All predictors retain the complete 2^G support.

Refresh schedule. Within each training run, SGRC acquires annotations at normalized training times $t_j \in \{0, \frac{1}{3}, \frac{2}{3}\}$, with $B = 2{,}304$ annotation attempts at each checkpoint and 6,912 attempts in total. These are successive stages of one run; the five seeds repeat this schedule independently. Static SGRC makes the same total number of attempts at initialization. The schedule is shared by the checkpoint-based sparse-reference systems.

Acquisition updates. Algorithm 1 uses each resolved mask once on its acquisition group. Acquisition rollouts come from the frozen checkpoint and are consumed over several minibatches while the live policy changes. Importance ratios use the segment-frozen checkpoint, with log-probabilities from stored acquisition rollouts; each minibatch is optimized for 1 epoch. An unresolved group produces no replacement mask or update. Empty and full masks retain the normalizer's zero-variance behavior. Annotations from earlier checkpoints are not pooled into later fits.

System controls. Direct sparse supervision uses acquisition updates without fitting a predictor. Continuous correction adds a learned real-valued residual to the verifier reward before the unchanged normalizer, using the same judge-annotation attempts and checkpoint schedule as the other sparse-reference systems. The pointwise confidence rule uses a per-rollout predictor and thresholds individual confidences instead of complete-mask confidence. It is an additional system control, outside the matched predictor–gate ablation. Static SGRC acquires all annotations at initialization and keeps one encoder, predictor, temperature, and threshold for the entire run. The random-partition control runs its own SGRC predictor and gate, then draws uniformly from all $\binom{G}{k}$ masks with the accepted cardinality k, including the accepted mask itself. This preserves the multiset of normalized advantages, but changes which rollouts receive them and can change the gradient magnitude. The dense comparator directly queries eligible runtime groups and applies resolved mixed masks. The DAPO sampling control implements dynamic sampling (Yu et al., 2025): it drops constant-reward groups and samples

further prompts to refill the batch, using an oversampling factor of 2.0 and no annotations. This control changes the sampler only.

B GATE SELECTION

Temperature $T > 0$ minimizes complete-mask negative log-likelihood on the temperature-calibration split. On the independent gate-selection split, a threshold accepts only unique, mixed MAP masks with $q \geq \tau$. Let n_τ be the accepted count, E_τ the number of complete-mask errors, and N the number of resolved gate-selection groups. The upper risk bound is the one-sided Clopper–Pearson binomial bound

$$U_\tau = \mathrm{Beta}^{-1}(1 - \delta'; E_\tau + 1, n_\tau - E_\tau), \qquad \delta' = \delta/|\Lambda|, \tag{5}$$

with $U_\tau = 1$ when $E_\tau = n_\tau$. Thresholds with no accepted groups are inadmissible. We choose the largest-coverage threshold satisfying $U_\tau \leq \alpha$ and $n_\tau/N \geq c_{\min}$; equal-coverage ties go to the largest threshold. If none qualifies, we set $\tau = \infty$.

Table 7: **Fixed parameters of the gate-selection rule.**

Parameter	Value
Risk target α	0.1
Minimum coverage $c_{\min}$	0.05
Bound level $1 - \delta$; per-threshold δ'	0.95; 0.0025
Threshold grid Λ	20 values in [0.5, 0.975], uniform step 0.025
Resolved groups by split	Predictor-fit / temperature-calibration / gate-selection: 400 / 150 / 1,600 resolved groups per checkpoint

The finite-grid adjustment accounts for threshold selection under independent checkpoint-distribution samples, conditional on annotation resolution. It does not bound true semantic errors of the judge, or guarantee the same risk on later groups from a changed policy. Temperature fitting alone provides no such guarantee. Neither offline evaluation nor final policy accuracy is used to select the operating threshold.

C DATA AND REFERENCE ANNOTATIONS

Data splits. The `geometry_sym` subset is selected from MathV360K (Shi et al., 2024) using the GeoQA+, UniGeo, and Geometry3K image-source prefixes and symbolic-answer patterns, including radicals, trigonometric expressions, π, and ratios. MathV360K includes both collected and synthesized question–answer pairs; our protocol selects and splits existing items from that release. The selected pool contains 2,827 items. After a seeded shuffle, the first 500 items form the held-out accuracy set. From the remaining items, we exclude those whose image paths also occur in the held-out set and then select 1,600 training items and 219 auxiliary diagnostic items. The training and auxiliary diagnostic splits are prompt-disjoint but can share images. The latter is used only for diagnostics in this study. The number of training images byte-identical to a held-out image is 0. Generic question strings can recur across figures: 24 strings (33 held-out items) also occur in training and 8 (13 items) in the auxiliary diagnostic split. The strict grader is `math_verify` on the committed answer, with greedy decoding and at most 256 new tokens.

Acquisition candidates come from the training prompts. The held-out accuracy split is never used for acquisition, predictor fitting, temperature calibration, gate selection, or offline mask evaluation. SGRC assigns acquisition prompts to disjoint predictor-fit, temperature-calibration, and gate-selection splits before labeling. Offline evaluation prompts are separate from these and from the held-out accuracy prompts. D_{mask} contains 900 empty, 90 mixed, and 10 full reference masks. An always-empty predictor attains high overall exact-mask accuracy on this set without recovering any mixed group; the main analysis therefore emphasizes mixed-mask recovery and selective errors.

Judge protocol. GPT-5.6 first solves the image question without seeing candidate rollouts. In the second stage it receives that solution and all rollouts in uniformly shuffled order under anonymous

indices, without verifier verdicts or the published answer. It returns `correct`, `incorrect`, or `unresolved` for every committed answer. An invalid first-stage solution, a malformed output, or any unresolved verdict causes whole-group abstention. Otherwise, the permutation is inverted to build the mask. Judge solutions, rationales, labels, and confidence never enter predictor inputs.

Table 8: **Judge configuration.** Both stage prompts will be released with the code.

Setting	Value
Model and recorded snapshot	GPT-5.6, 2026-09
Access dates	1–15 September 2026
Temperature / top-p / maximum output tokens	0 / 1 / 2,048
Retries and timeout	2 retries; 60 s timeout
Serialization and output	image bytes; shuffled rollout text; eight JSON verdicts

Annotation accounting. Attempts include unresolved outcomes, and realized groups depend on the policy and seed. Table 9 shows the accounting for seed 1. Each resolved mask supplies one acquisition-group update; only mixed masks carry a reward-derived ranking signal. Calls include the solving and adjudication stages actually executed.

Table 9: **SGRC annotation accounting for one run.** Attempts = resolved + abstained; resolved = empty + mixed + full. Each checkpoint uses the same number of annotation attempts.

Checkpoint	Attempts	Abstained	Empty	Mixed	Full	Resolved	Calls
0	2,304	154	1,900	230	20	2,150	4,454
1/3	2,304	154	1,900	230	20	2,150	4,454
2/3	2,304	154	1,900	230	20	2,150	4,454
Total	6,912	462	5,700	690	60	6,450	13,362

D Per-Seed Policy Results

Each seed fixes initialization and prompt order across compared systems; trajectories diverge after their updates first differ. The main tables report sample standard deviation across these five runs. Table 10 lists the individual outcomes underlying the paired comparisons in the main text.

Table 10: **Strict held-out accuracy by seed** (%). The upper block compares sparse-reference systems and fixed-verifier GRPO; the dense comparator uses runtime annotations.

System	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
Fixed-verifier GRPO	19.0	20.4	19.4	21.0	20.2
Direct sparse supervision	19.6	21.0	19.8	21.4	20.6
Continuous correction	20.2	21.2	20.0	22.0	21.0
Pointwise confidence rule	20.6	21.8	21.2	21.8	21.6
Rollout-local + exact gate	20.8	22.0	21.0	22.2	21.8
Factorized + exact gate	20.8	22.2	21.0	21.8	22.2
Factorized, no gate	19.8	20.8	20.0	21.2	20.8
Joint, no gate	20.2	21.6	20.4	21.6	21.4
Static SGRC	21.0	22.4	21.6	22.0	22.6
Random-partition control	19.6	20.8	19.8	21.4	20.6
SGRC	**21.6**	**23.6**	**22.8**	**22.6**	**24.2**
Dense-judge comparator	**22.0**	**24.4**	**23.2**	**24.0**	**24.8**

E Annotation and Runtime Audits

Reference quality. A repeated-permutation audit compares complete judge masks under a fresh rollout ordering. Human evaluation uses two blinded annotators, a stratified sample, and third-annotator

adjudication. Mixed-group agreement is reported separately because empty groups dominate the reference pool. Positive-label precision is computed over audited rollout labels; group false promotion is the fraction of audited groups containing at least one false-positive label. These are audits of the reference judge, not of the learned predictor.

Table 11: **Reference annotation audits.** Agreement and precision describe the audited sample.

Outcome	Count / total	Rate (%)
Valid stage-1 solution	232 / 240	96.7
Resolved adjudication	220 / 240	91.7
Repeated-permutation agreement	115 / 120	95.8
Human exact-mask agreement, all groups	110 / 120	91.7
Human exact-mask agreement, mixed groups	30 / 35	85.7
Positive-label precision	69 / 69	100.0
Group false promotion	0 / 120	0.0

Runtime behavior. The runtime audit samples 20 accepted groups uniformly at random per seed and segment. It compares applied masks against fresh judge labels; these labels never enter training. Across 300 audited groups, 293 resolve and contain 10 exact-mask errors. Treating every unresolved group as an error yields an error rate of 5.7%. Table 12 separates activation counts from the audited errors.

Table 12: **Runtime activation and audit by segment**, pooled across five seeds. Accepted share uses eligible all-rejected groups as the denominator; errors are counted among resolved audited groups.

Segment	Eligible	Accepted	Share (%)	Audited	Errors
$[0, 1/3)$	3,545	188	5.3	100	2
$[1/3, 2/3)$	3,280	180	5.5	100	4
$[2/3, 1]$	3,015	170	5.6	100	4

Runtime measurement. On a single NVIDIA B200 GPU, SGRC adds one frozen-encoder pass per eligible group and achieves $0.90\times$ eligible-group throughput relative to fixed-verifier GRPO. Acquisition and judge calls are excluded from this throughput measurement.

Scope of the comparisons. The component studies compare the proposed joint energy with a parameter-matched factorized predictor. Other joint parameterizations and alternative gate objectives remain untested. Section 2 compares complete noisy-reward training algorithms by mechanism; the experiments evaluate the systems listed in Section 4.