Hints, Critics, and Teachers: Prior Injection for Sparse-Reward RL in Vision–Language Math Reasoning

Qiqian Fu
qiqianf2@illinois.edu

August 2026

Abstract

Reinforcement learning for vision–language math reasoning starves under sparse reward: on a pool of 20,830 visual-math problems where Qwen2-VL-2B answers 3.6% of rollouts correctly, 85–97% of GRPO rollout groups are entirely wrong and contribute zero gradient. We train eleven methods under identical conditions in this regime, each injecting a different prior: text (reference-solution hints), distribution (on-policy distillation from a 7B teacher), and value (a value-pretrained critic with an MSE or HL-Gauss categorical loss). A prior helps exactly when it is *delivered*: the six arms whose prior effectively reaches the policy separate with no overlap from the remaining five—the no-prior baseline and four arms whose prior is teacher-capped, gated away, or lost to a mis-parameterized critic—both on the pooled in-domain metric and on cross-domain transfer (DynaMath). The central finding, however, concerns evaluation: one slice of the in-domain pool—long used as this project's general-distribution check—*anti-correlates* with genuine cross-domain transfer (Spearman $\rho = -0.74$, $n = 11$ arms, permutation $p = 0.011$), while the hardest in-domain slice predicts it closely ($\rho = +0.89$, $p < 0.001$). We attribute the inversion to a near-chance multiple-choice subset that rewards models for not having changed; read through it, the best cross-domain method looked mediocre and the worst looked like the champion. Among the methods, hint-guided exploration—not UFT's auxiliary loss—drives hint gains, and replacing the critic's MSE loss with HL-Gauss cross-entropy is worth +14.4 points in-domain. All accuracies are blind-judged, with paired exact tests.

1 Introduction

Group-relative policy optimization (GRPO) [1] has become the default recipe for reinforcement learning on math reasoning, but it carries a structural failure mode: its advantage is computed *within* a group of rollouts for the same prompt, so a group in which every rollout is wrong contributes exactly zero gradient. On problems a model rarely solves, most groups are all-wrong and training starves. This regime is not exotic. When we profile Qwen2-VL-2B [2] on the hardest problems of its own training distribution—a pool of 20,830 visual-math problems filtered from MathV360K [3]—the base model answers 3.6% of rollouts correctly, and 85–97% of GRPO groups open all-wrong.

The literature offers three families of remedies, each injecting a different *prior* into the starved policy. A **text** prior prefixes rollouts with part of a reference solution, so the model explores from a state where success is likely [4, 5]. A **distribution** prior replaces or augments the reward with a per-token signal from a stronger teacher, via on-policy distillation [6, 7]. A **value** prior trains a critic so that credit reaches individual tokens even when sequence outcomes are uniform [8], and recent work argues the critic's loss should be categorical rather than a scalar regression [9–11]. These families are studied in separate papers, on different models, tasks, and budgets; to our knowledge they have not been compared under one roof, and almost never on a vision–language model.

This paper runs that comparison. We construct a genuinely sparse training pool (verifying, before training anything, both that GRPO starves on it and that hints can unlock it), then train eleven arms—plain GRPO, three hint variants, four distillation variants, two critic variants, and one arm stacking the text and distribution priors—under identical conditions: the same pool, the same step and batch budget, the same seed, one 8×A800 node per arm. Every accuracy we report is blind-judged by an independent LLM judge (DeepSeek) [12, 13] that reads full outputs without knowing which arm produced them, and pairwise claims use McNemar exact tests.

The method-level results are clean. What separates the arms is not whether a prior is injected by design but whether it is *delivered*: on the only evaluation set outside the training domain (DynaMath [14]), the six arms whose prior demonstrably reaches the policy occupy the range 30.6–32.9%, and the remaining five—plain GRPO and four arms whose prior is capped at the teacher's ability, gated to failures, or lost in a mis-parameterized critic—occupy 27.0–29.1%, with no overlap. Within that headline, the comparison localizes *which* component earns the gain: hint-guided exploration works while UFT's auxiliary log-likelihood term contributes nothing; swapping the critic's clipped-MSE loss for HL-Gauss cross-entropy—a one-line change—moves in-domain accuracy by +14.4 pp and turns the weakest trained arm into the second-best cross-domain arm; annealed global distillation works where failure-gated persistent distillation does not; and stacking two priors buys nothing, because both solve the same early bottleneck.

The finding that reorganized this study, however, is about evaluation rather than methods. For four rounds of this project, a held-out MathV360K slice (old-319) served as the "general-distribution" check, and conclusions were drawn from it. With eleven arms evaluated on both domains—and the in-domain pool decomposed into its two slices—we could finally correlate the evaluation sets against each other, and old-319 ranks the arms *backwards*: its scores anti-correlate with genuine cross-domain transfer (Spearman $\rho = -0.74$, $n = 11$ arms), while the hardest in-domain slice (sparse-800) predicts transfer closely ($\rho = +0.89$). The mechanism, we argue, is compositional: 69% of old-319 is four-option multiple choice on which every arm sits near the 25% guessing floor, so scoring well there means still emitting short option letters like the base model. The set rewards *not having changed*—precisely the quantity that anti-correlates with capability gain. Three earlier conclusions of this project, including an apparent specialization–generalization trade-off, were artifacts of that lens; we retract them explicitly in Section 4. We report this in detail because the failure pattern is generic: any project that tracks generalization on a fixed held-out slice whose composition drifts toward chance-level subtasks can be steered backwards by it [15].

A third thread runs through the paper: verification as a deliverable. Earlier rounds of this project were rewritten by silent scoring and porting bugs, so round 4 treated every load-bearing mechanism as something to be gated before training and audited after. A scoring audit found that a heuristic answer extractor penalized different arms by different amounts—up to 30 pp—in a direction that inverted rankings; a hacking analysis tested and rejected guessing and format-reward explanations for the headline gains [16]. We describe this discipline in Section 6 because without it, most of the numbers in this paper would have been wrong in ways no significance test would catch.

Contributions.

- A controlled comparison of eleven training arms spanning text, distribution, and value priors for sparse-reward RL on a 2B vision–language model, under identical data, budget, and judging, with verified preconditions of sparsity (85–97% all-wrong groups) and hint efficacy (7.2% → 34.4% with a half-solution prefix).
- Method findings: arms with a delivered prior separate cleanly from the rest on cross-domain transfer; exploration, not the auxiliary objective, drives hint gains; the critic's loss parameteri-

zation alone is worth +14.4 pp (concurrent with a text-only version of this finding [11]); two priors do not compose.

- An evaluation-validity finding: a held-out "general" set that anti-predicts cross-domain transfer ($\rho = -0.74$) while the hardest in-domain tail predicts it ($\rho = +0.89$), a mechanism for the inversion, and three explicit retractions of conclusions the broken lens had produced.
- A reusable verification protocol for RL comparisons—blind LLM judging, scoring audits, pre-training hard gates, and reward-hacking analyses—together with the concrete failures it caught.

Organization. Section 2 situates the three prior families. Section 3 describes the pool construction (including a measurement trap that would have selected formatting failures instead of hard problems), the eleven arms, and the judging protocol. Section 4 presents the main comparison and the evaluation-set inversion. Section 5 details seven method-level findings. Section 6 documents the verification discipline, and Section 7 states limitations, chief among them the single training seed.

2 Related Work

Guided exploration with hints. UFT [4] unifies supervised and reinforcement fine-tuning by prefixing rollouts with part of a reference solution, annealing the hint proportion toward zero over training, and adding a supervised log-likelihood term on hint tokens to the RL objective. The hint acts as an exploration scaffold: it places the policy in the neighborhood of a correct solution, where reward becomes reachable. Our `hintonly`, `uft001`, and `uft002` arms instantiate this recipe on top of GRPO and deliberately factor it into its two components—hint-guided exploration and the auxiliary likelihood term—so that each component can be credited separately.

On-policy distillation. On-policy distillation trains the student on its own rollouts while a teacher scores every token, setting the per-token advantage to the negative reverse KL between student and teacher [6]. Liao et al. [7] extend this to multi-turn agents by replaying pre-collected teacher trajectories as rollout prefixes under a step-decaying length schedule (ReOPD); our `reopd` arm adapts the replay mechanism to our single-turn setting. Because reverse-KL imitation cannot exceed the behavior it imitates, distillation-only training inherits the teacher's ceiling—a property our comparison makes visible by running distillation arms alongside reward-driven arms that are free to surpass the teacher.

Value functions and critic parameterization. Classic actor–critic training pairs PPO [17] with generalized advantage estimation [18]. VAPO [8] adapts value-based RL to long-chain-of-thought reasoning through value pretraining and decoupled GAE, and adopts the clip-higher technique introduced by DAPO [19]; our `vapo` arm follows this recipe. A separate line of work replaces the critic's scalar regression with classification over a discretized return support: the histogram loss of Imani and White [9], which Farebrother et al. [10] showed to scale value learning across deep-RL domains, and which Zhou et al. [11] applied to LLM RL critics—concurrently with our work, and on text-only tasks. Our `vapohl` arm isolates exactly this substitution in a vision–language setting.

RL for vision–language math reasoning. GRPO [1] has become the default policy-gradient method for math reasoning and transfers directly to vision–language models such as Qwen2-VL [2]. Training corpora and benchmarks for this setting include MathV360K [3] and the program-generated

DynaMath [14]. Several works caution against taking RLVR gains at face value: accuracy-only rewards fail to improve the perception component of multimodal reasoning [20], even spurious rewards can raise benchmark scores [16], and GRPO's normalization terms carry biases of their own [21], which in all-failure groups can let auxiliary dense terms dominate the gradient [22]. Closest to our hint arms, Off-Context GRPO [5] also conditions rollouts on privileged information but corrects the resulting off-policyness with an importance-weighted objective, a correction our hint arms do not apply.

Evaluation validity. All accuracies we report are blind-judged by an LLM judge, a protocol whose agreement with human raters is well documented [13], as are its failure modes [23]. On the benchmark side, Alzahrani et al. [15] show that the surface composition of multiple-choice benchmarks—option order, answer-extraction format—can move leaderboard rankings by several positions: what a benchmark appears to measure and what actually drives its rankings can diverge. Our evaluation-validity finding is of this kind but sharper in direction—a held-out set used across every round of this project as its generalization check turns out to rank methods in the *reverse* order of their true cross-domain transfer.

3 Experimental Setup

3.1 Background: three rounds of honest negatives

This work is the fourth round of a longer investigation. Rounds 1–3 asked whether UFT-style hint-guided exploration [4] beats plain GRPO [1] for a 2B vision–language model on visual math, and concluded—after retracting two rounds contaminated by five port-fidelity bugs (whose fixes moved individual arm scores by up to +9.1 pp)—that *UFT ties GRPO*: in round 3, `grpo` reached 33.9% on the held-out set against 31.0/32.6% for the two UFT arms, with no significant differences in the hint arms' favor. The mechanism analysis of the earlier rounds contained the seed of this one: the reward was never sparse enough for hints to matter. GRPO's fraction of all-wrong rollout groups—groups that contribute zero gradient under group-normalized advantages—opened at 77% and fell to roughly 30% within training; the hint had nothing to rescue.

Round 4 tests the counterfactual: *make the reward genuinely sparse, then re-run the comparison*—and widens it to every prior-injection family proposed for that regime.

3.2 Constructing a genuinely sparse pool

We profiled Qwen2-VL-2B-Instruct [2] on all 52,964 problems of our MathV360K-derived training corpus [3]: 8 rollouts per problem at temperature 1.0, roughly 428k generations in total. Profiling immediately surfaced a measurement trap worth naming.

> **Measurement trap.** Scored with the training reward—which requires an `Answer: X` format—the base model gets **3.5%**. Rescored semantically, the same outputs score **43.4%**. The strict scorer measures format compliance, not ability; selecting "hard" problems with it would have built a pool of formatting failures. All difficulty selection therefore used the semantic score, and every reported result uses a blind LLM judge that reads full outputs and is immune to format drift.

The pool consists of the 21,630 problems the base model solves at most 1-in-8 times under semantic scoring. We hold out an 800-problem split (sparse-800) and train on the remaining 20,830. Within the pool the base model answers 3.6% of rollouts correctly.

Arm	Prior	Mechanism
`grpo`	none	plain GRPO, groups of n=4 (the baseline)
`hintonly`	text	reference-solution prefix, cosine-annealed 95%→5%, then off
`uft001`	text	`hintonly` + NLL on hint tokens, coef. 0.001 (full UFT)
`uft002`	text	`hintonly` + NLL on hint tokens, coef. 0.002
`opd`	distribution	pure on-policy distillation: advantage = −reverse KL, no reward
`reopd`	distribution	`opd` + teacher-trajectory prefix replay
`srpo`	distribution	GRPO + distillation gated to failed rollouts only
`saf`	distribution	GRPO + distillation clamped to ±2, annealed $\lambda(t)$: 1→0
`hintsaf`	text + dist.	`saf` + hint injection
`vapo`	value	PPO with value-pretrained critic, clipped-MSE value loss
`vapohl`	value	`vapo` with HL-Gauss categorical value loss

Table 1: The eleven arms. All train on the same 20,830-problem sparse pool for 500 steps with identical batch size, learning rate, and seed.

Two preconditions were verified *before* training a single arm. First, the pool is genuinely sparse for GRPO: the opening all-wrong group fraction is 85–97%, versus 77% in the earlier rounds. Second, hints can actually unlock it: prefixing half a reference solution lifts base accuracy from 7.2% to 34.4%—a property round 3's pool never had.

3.3 The eleven arms

Eleven methods were trained under identical conditions, each injecting a different prior into the starved regime (Table 1).

Text priors. `hintonly` prefixes each rollout with part of a reference solution; the fraction of hinted rollouts is cosine-annealed from 95% to 5% over the first 300 steps, then set to zero. `uft001` and `uft002` add the signature term of UFT [4]—a log-likelihood loss on the hint tokens—at coefficients 0.001 and 0.002, completing the full UFT objective.

Distribution priors. `opd` is pure on-policy distillation from a Qwen2-VL-7B teacher in the Thinking Machines recipe [6]: the per-token advantage is the negative reverse KL to the teacher (k1 estimator), with no task reward. `reopd` adds teacher-trajectory prefix replay with a geometrically sampled cut index ($\kappa = 0.6$), our single-turn adaptation of a method originally proposed for multi-turn agents [7]. `srpo` keeps the full GRPO objective and gates the distillation signal to failed rollouts only (sequence advantage ≤ 0), never decaying it. `saf` instead applies the distillation term globally, clamped to ±2 and cosine-annealed from 1 to 0 over the 500 steps. `hintsaf` stacks `saf` with hint injection to test whether the two priors compose.

Value priors. `vapo` follows the VAPO recipe [8]: PPO [17] with a same-backbone critic, 30 steps of value pretraining against Monte-Carlo targets with the actor frozen, decoupled GAE [18] with $\lambda_{\text{critic}} = 1$, and the clip-higher ratio bounds 0.2/0.28 introduced by DAPO [19] and adopted by VAPO. We set $\lambda_{\text{policy}} = 0.98$ (our choice; VAPO proposes a length-adaptive schedule). The value loss is clipped MSE. `vapohl` changes exactly one thing: the value loss becomes an HL-Gauss cross-entropy [9, 10] over 101 bins on $[-0.1, 1.1]$ with $\sigma = 0.75\times$ bin width, truncated-Gaussian targets, no value clipping, and a zero-initialized head.

Domain	Set	n	Nature	Base (%)
in-domain	MathV (pooled)	1119	sparse-800 + old-319, size-weighted 71.5/28.5	
	sparse-800	800	training-distribution hard tail, never trained on	≈ 0
	old-319	319	earlier MathV360K slice; 69% Geometry3K 4-option MC	—
cross-domain	DynaMath	477	programmatically rendered; zero overlap with MathV360K	25.8
	easy	373		30.3
	hard	104		11.5

Table 2: Evaluation datasets, grouped by domain. The combined DynaMath base figure is reported as measured; the size-weighted average of the two split figures is 26.2, a 0.4 pp bookkeeping difference we do not resolve here.

3.4 Training configuration

Every arm trains for 500 steps at batch size 256 prompts $\times$ 4 rollouts, learning rate 10^{-6}, KL coefficient 0.001, seed 42, on one 8×A800 node (distillation arms split the node 4/4 between student and teacher). The teacher's tokenizer is file-identical to the student's (MD5-verified; Section 6). All training runs on verl [24] with vLLM rollouts [25]; our modifications to the framework are environment-gated patches described in Section 6.

3.5 Evaluation sets and judging protocol

We evaluate along two axes (Table 2): one in-domain dataset drawn from MathV360K, and one cross-domain dataset with no overlap with it. The in-domain dataset (MathV, n=1119) pools two slices at their size weights (71.5/28.5): sparse-800, the held-out hard tail of the training distribution, and old-319, a different MathV360K slice (rows excluded from training by an earlier data-cleaning step), 69% of which is Geometry3K four-option multiple choice and which had served as this project's general-distribution check since round 1. The cross-domain dataset is DynaMath [14] (n=477, an easy split of 373 problems and a hard split of 104): its problems are programmatically rendered, with zero overlap with MathV360K. Because the two in-domain slices turn out to measure different quantities (Section 4.2), we report the pooled in-domain number together with its per-slice breakdown throughout.

Judging. All reported accuracies come from a model-blind LLM judge (DeepSeek; 12) that reads the full model output with the arm identity hidden and judges each answer against the ground truth. Pairwise comparisons use McNemar's exact test on paired per-problem outcomes. Judge run-to-run variance is small: three independent judging passes over identical `grpo` outputs scored 40.5, 41.1, and 41.0 ($\pm$0.5 pp).

4 Results

4.1 Main comparison

Table 3 reports blind-judged accuracy for all eleven arms on the in-domain metric (MathV pooled, with its two-slice breakdown) and the cross-domain metric (DynaMath), sorted by DynaMath. Every trained arm clears the near-zero base accuracy on the training distribution's hard tail by a wide margin, but the arms divide sharply in how far they get: six arms span 41.6–44.1 on the

Table 3: Blind-judged accuracy (%) of all eleven arms, sorted by the cross-domain metric (DynaMath, n=477, pooling the easy and hard splits). MathV (pooled, n=1119) is the in-domain metric, combining the sparse-800 and old-319 slices at their size weights (71.5/28.5); the two slice columns give its breakdown. The horizontal rule separates the six delivered-prior arms from the five undelivered-prior arms: the two groups do not overlap on either metric. Within-group orderings should not be read as rankings (single training seed; Section 7).

Arm	Prior	sparse-800	old-319	MathV (pooled)	DynaMath
`uft002`	text	52.4	19.1	42.9	**32.9**
`vapohl`	value (HL-Gauss)	**53.6**	19.7	44.0	31.9
`hintonly`	text	51.4	21.3	42.8	31.2
`hintsaf`	text + distribution	49.4	26.0	42.7	30.8
`uft001`	text	52.0	15.7	41.6	30.8
`saf`	distribution (annealed)	49.6	30.4	**44.1**	30.6
`srpo`	distribution (gated)	40.8	30.4	37.8	29.1
`opd`	distribution (pure)	40.5	22.9	35.5	29.1
`reopd`	distribution (replay)	38.8	25.1	34.9	28.1
`grpo`	none	41.0	27.6	37.2	27.5
`vapo`	value (MSE)	39.2	**32.6**	37.4	27.0
base (untrained)	—	∼0	—	—	25.8

in-domain metric (49.4–53.6% on the sparse-800 slice), while the other five span 34.9–37.8 (38.8–41.0 on sparse-800). The same six-versus-five split appears on both axes.

We refer to the upper six as the *delivered-prior* arms and the lower five as the *undelivered-prior* arms. The names are empirical rather than architectural—`opd`, `reopd`, `srpo`, and `vapo` all inject a prior by design—but every membership is explained by a design feature: the lower group comprises the no-prior baseline (`grpo`), pure distillation capped at its teacher's own ability (`opd`, `reopd`), distillation gated to failed rollouts and never decayed (`srpo`), and a value prior lost to a mis-parameterized critic (`vapo`; Section 5).

The cross-domain column exhibits a complete separation (Figure 1): the six delivered-prior arms occupy every position from 30.6 to 32.9, and the five undelivered-prior arms every position from 27.0 to 29.1, with no overlap between the groups. We emphasize what this claim is and is not: the *between-group* separation is the result; the ordering *within* either group spans at most 2.3 pp on a 477-item set under a single training seed, and we do not interpret it (Section 7).

The group boundary also clarifies why `vapo` and `vapohl`—which share every component except the critic's loss function—land on opposite sides. A value prior is only injected to the extent that the critic actually learns value; with a clipped-MSE loss the critic starts from a pathological cold state and never catches up, so `vapo` behaves like an arm without a prior, while the HL-Gauss parameterization delivers the prior and places `vapohl` with the delivered group (Section 5, F3).

4.2 The benchmark that ranked backwards

The finding that reorganized this report concerns the evaluation rather than the methods: the single in-domain number of Table 3 hides two components that measure different quantities. Since round 1 this project had used old-319—the slice that now forms 28.5% of the in-domain pool—as its "general-distribution" check. Read against the cross-domain column, that slice runs the wrong

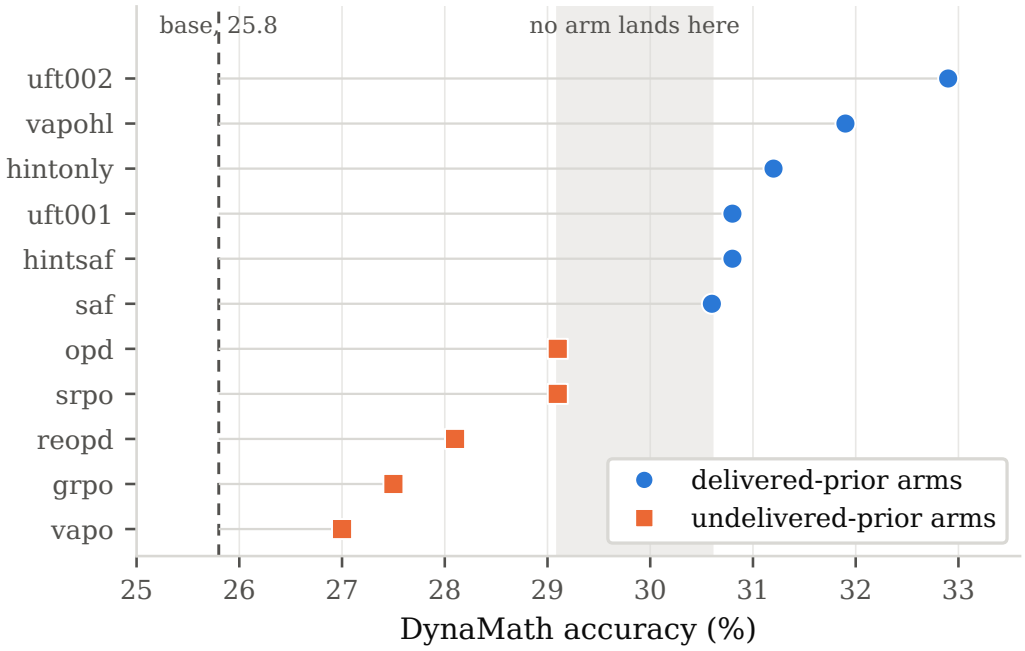

Figure 1: All eleven arms ranked by DynaMath accuracy. Every delivered-prior arm outperforms every undelivered-prior arm: the shaded band (29.1–30.6) separates the two groups and contains no arm. Differences within each group are inside judging noise and should not be ranked (Section 7). Dashed line: untrained base (25.8).

way: `vapo` tops old-319 at 32.6 while placing last among all trained arms on DynaMath (27.0, below `grpo`), and `vapohl` sits third-from-bottom on old-319 (19.7) while placing second overall on DynaMath (31.9).

Table 4 and Figures 2 and 3 quantify the pattern. The hardest in-domain set, sparse-800, predicts genuine cross-domain transfer closely ($\rho = +0.89$, n = 11 arms, permutation $p < 0.001$), while old-319 *anti-predicts* it ($\rho = -0.74$, permutation $p = 0.011$). The two in-domain sets anti-correlate with each other ($\rho = -0.66$): they are not noisy measurements of one quantity but measurements of two different quantities. Because the within-group orderings are inside noise (Section 7), these correlations are driven by the between-group separation and should be read as statements about which sets order the two groups correctly, not as evidence about fine-grained rankings of individual arms.

Why old-319 inverts. 69% of old-319 is Geometry3K four-option multiple choice, on which every arm—trained or untrained—sits near the 25% guessing floor. Scoring well there is therefore not about solving geometry; it is about still behaving like the base model, emitting a short bare option letter. Arms that genuinely learned something shifted their output distribution toward multi-step reasoning, which on a chance-level multiple-choice set converts lucky guesses into reasoned-but-wrong commitments. On this reading, the set rewards *not having changed*, and that is precisely the quantity that anti-correlates with capability gain. We state the epistemic status of this explanation plainly: it is an interpretation consistent with the set's composition, not a direct multiple-choice versus free-form decomposition of each arm's predictions, which we defer to a future revision.

Table 4: Spearman rank correlation between evaluation sets over the eleven trained arms (n = 11). The five correlations computable from Table 3 were independently recomputed from the published numbers and agree with the values below to within 0.004; the two easy-split rows use per-arm accuracies on DynaMath-easy alone.

X	Y	ρ
sparse-800	DynaMath	+0.89
old-319	DynaMath	−0.74
MathV pooled (71.5/28.5)	DynaMath	+0.76
MathV pooled (50/50)	DynaMath	+0.40
sparse-800	old-319	−0.66
sparse-800	DynaMath-easy	+0.87
old-319	DynaMath-easy	−0.81

Table 5: Earlier readings and their corrections after the evaluation-validity analysis.

Earlier reading	Corrected
All arms lie on one specialization-versus-generalization trade-off curve; none advances both ends.	The apparent curve is capacity reallocation *within* MathV360K. Across domains there is no trade-off: stronger in-domain predicts stronger out-of-domain.
`vapohl` is the specialization extreme, paying a generalization tax.	`vapohl` is second-best cross-domain (31.9). Its low old-319 score reflects only that it stopped guessing option letters.
`vapo` is the robustness champion—best on the "general-distribution" set.	`vapo` is the weakest trained arm cross-domain (27.0, below plain `grpo`). old-319 had the ranking inverted.

Consequences for pooled metrics. A corollary is that any single in-domain number inherits the flaw in proportion to how much old-319 it contains: at the 71.5/28.5 size weighting the pooled metric correlates with DynaMath at $\rho = +0.76$, but at equal weighting only +0.40, because equal weighting amplifies the anti-correlated half. The cleanest single in-domain number is sparse-800 alone.

Corrected readings. Three interim conclusions of this project were artifacts of reading old-319 as a generalization measure. We retain them here, with their corrections, as a record of how a miscalibrated held-out set distorts method comparison (Table 5).

5 Method Findings

The seven findings below are stated against the numbers in Table 3; all pairwise claims use exact McNemar tests on blind-judged outputs.

F1 — Sparse reward is the regime where hints earn their keep. All three hint arms beat `grpo` on sparse-800 by +10.4 to +11.4 pp, every comparison $p < 10^{-4}$ with confidence intervals

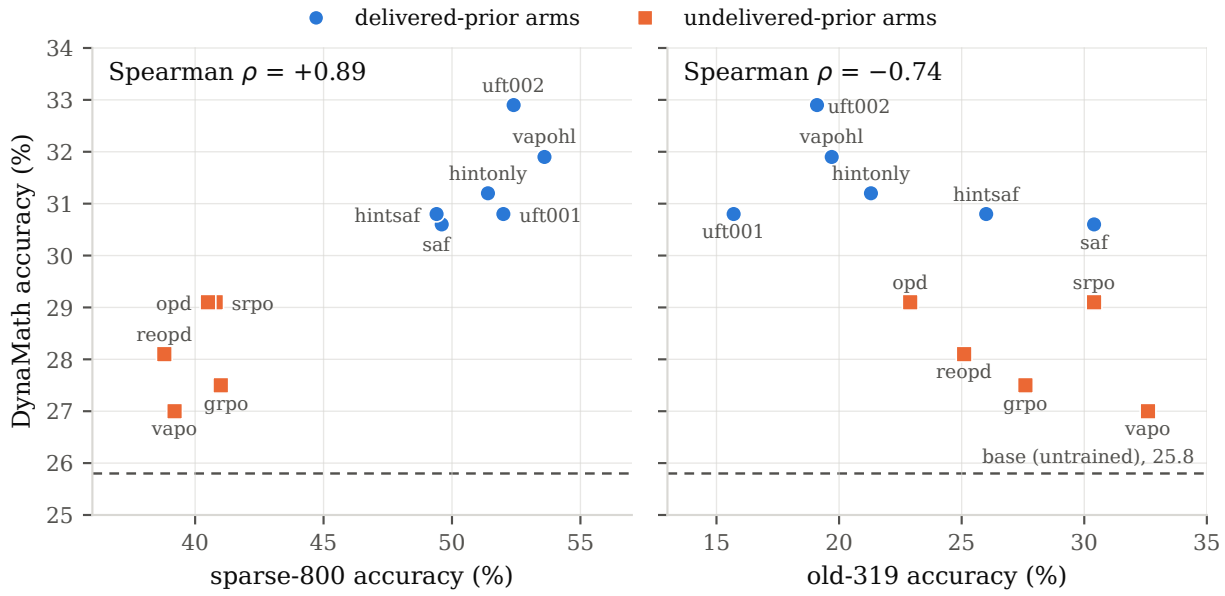

Figure 2: Which held-out set predicts cross-domain transfer? Each point is one trained arm; the y axis is accuracy on DynaMath (n=477), the only evaluation pool with no overlap with the training distribution. **Left:** the hardest in-domain held-out set (sparse-800) ranks arms nearly as DynaMath does (Spearman $\rho = +0.89$, $n = 11$). **Right:** the set used as a "general-distribution" check since round 1 (old-319) anti-predicts transfer ($\rho = -0.74$): `vapo` tops old-319 while placing last among trained arms on DynaMath, and `vapohl` nearly inverts that. Dashed line: untrained base model (25.8).

bounded away from zero. Round 3 of this project had concluded that UFT ties GRPO [4]; that conclusion is thereby relocated, not refuted. In the earlier rounds the reward was never sparse enough for the mechanism to bind—the fraction of all-wrong rollout groups started at roughly 77% and fell to about 30% over training, so the hint had little to rescue. On the present pool, which opens at 85–97% all-wrong groups, the same mechanism is worth ten points.

F2 — It is the exploration, not the objective. `hintonly`, `uft001`, and `uft002` are statistically indistinguishable in-distribution (51.4 / 52.0 / 52.4 on sparse-800, all pairwise n.s.), so UFT's auxiliary hint-token log-likelihood term contributes nothing where it was supposed to help. The best UFT configuration is the one with UFT's own signature term removed; what remains is hint-guided exploration.

F3 — The critic's loss function is worth more than the critic. `vapo` → `vapohl` changes exactly one component—clipped MSE becomes HL-Gauss cross-entropy [9, 10] over 101 bins—and moves sparse-800 from 39.2 to 53.6 (+14.4 pp, significant), DynaMath from 27.0 to 31.9, and explained variance from 0.59 to 0.66 (Figure 4). The categorical head also removes a cold-start pathology: zero-initialized, it starts at a uniform distribution ($V = 0.5$, explained variance ≈ 0), where the scalar head starts at an explained variance of -755. A good critic decides how deep the policy can dig; a badly parameterized one wastes the dig. A concurrent study reports the same categorical-critic effect for text-only LLM reinforcement learning [11]; the two results were obtained independently and corroborate each other across modalities.

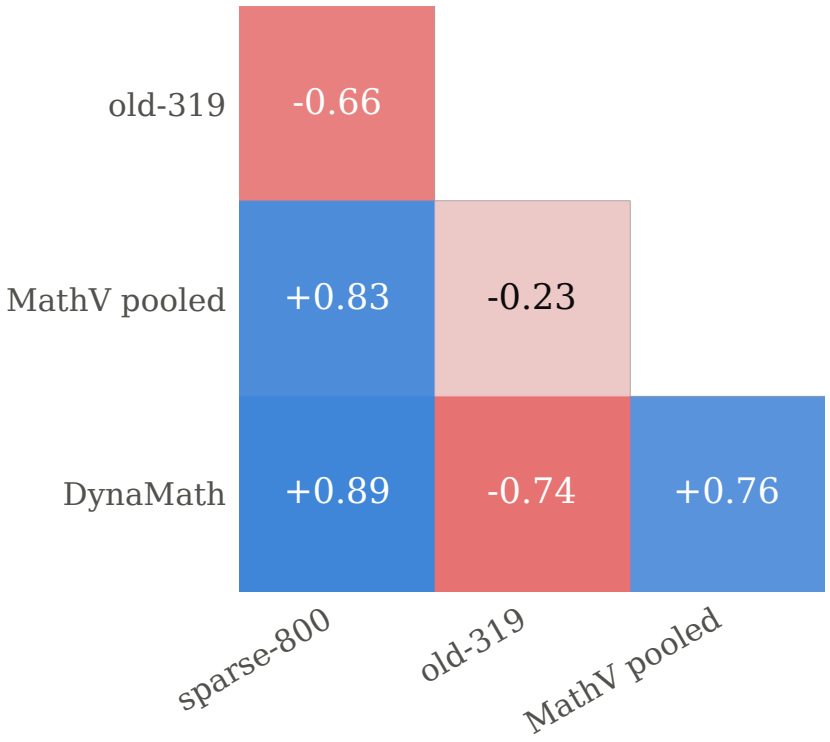

Figure 3: Spearman rank correlations between evaluation sets over the eleven trained arms, computed from the blind-judged accuracies in Table 3. sparse-800 is the strongest predictor of cross-domain accuracy (+0.89); old-319 is strongly anti-correlated with it (−0.74) and with sparse-800 itself (−0.66). The pooled in-domain metric inherits its predictive validity mechanically from its sparse-800 component (the pool contains both sets; weights 71.5/28.5).

F4 — Annealed distillation is the robust recipe; targeted distillation is not. `saf`—GRPO plus a bounded, cosine-decayed teacher-KL term—beats `grpo` by +8.6 pp on sparse-800 and posts the best DynaMath-hard score of all eleven arms (18.3%). `srpo` applies the same distillation signal but gates it to failed rollouts and never decays it, and lands at `grpo` level on every set. A global prior that fades as the policy finds its own signal works; a targeted, persistent one does not.

F5 — Pure distillation inherits its teacher's ceiling. `opd` and `reopd` [6, 7] converge to `grpo`-level performance. The 7B teacher itself scores only 27.9% on the pool; reverse-KL imitation cannot exceed what the teacher knows, while reward-driven arms can and do—the hint arms reach roughly 52% on sparse-800, far beyond their teacher. `reopd` tracks `opd` everywhere (all n.s.), consistent with its source setting: prefix replay repairs a multi-turn pathology that this single-turn setting does not have.

F6 — The gains travel. Every delivered-prior arm beats the untrained base on DynaMath by +4.8 to +7.1 pp, while plain `grpo` manages +1.7. On the easy split the hint arms' margins over base are individually significant (+6.2 to +7.8 pp, $p = 0.012$–0.035), where `grpo` moves −0.5 pp (n.s.). This is the first significant cross-domain transfer in four rounds of this project, and it belongs to the priors, not to reinforcement learning as such.

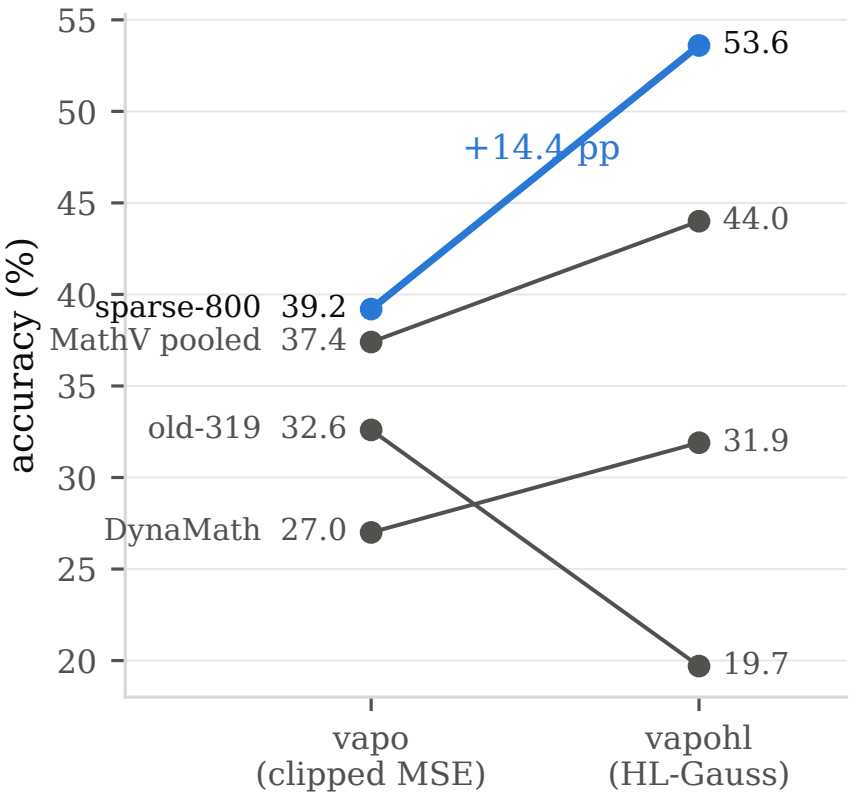

Figure 4: One change: replacing `vapo`'s clipped-MSE value loss with HL-Gauss cross-entropy (`vapohl`), everything else identical. In-domain accuracy on sparse-800 moves +14.4 pp and cross-domain DynaMath from 27.0 to 31.9, while old-319 drops—the arm stopped guessing option letters (Section 4.2).

F7 — Two priors do not compose. `hintsaf` stacks the text prior on the annealed distribution prior and lands at the weighted middle of its parents on every set (sparse-800 49.4: −0.2 pp vs. `saf` and −2.0 pp vs. `hintonly`, both n.s.). Hints and teacher-KL solve the same early bottleneck—escaping the zero-gradient region of an all-wrong batch—so stacking them is redundant rather than complementary.

6 Verification Discipline

Rounds 1–2 of this project were retracted after five silent port-fidelity bugs rewrote their conclusions, so round 4 treated verification as a deliverable in its own right. Everything in this section happened *before* conclusions were drawn.

6.1 The scoring audit that rewrote the leaderboard

Before adopting any accuracy number we audited the heuristic (pattern-extraction) scorer against roughly 340 hand-read rows, backed by a programmatic recheck of the full result set. The audit found that the heuristic penalized each arm differently—and in the *opposite* direction from round 3, because trained arms drift away from the extraction patterns rather than toward them. `uft001`'s

Hypothesis	Verdict
Answer-space guessing	Rejected: 48.9% accuracy against an 8.7% aggregate guessing ceiling, with a diverse answer distribution.
Format-reward hacking	Rejected: correct rows carry genuine multi-step reasoning (median 270 characters).
Template leakage from training	Partial: roughly 3 pp of attributable inflation; rows without a template duplicate in the training pool still score 45.9% against a 3.5% base.
Composition artifacts on old-319	Confirmed—and followed to the evaluation-validity finding of Section 4.

Table 6: Four reward-hacking hypotheses and their verdicts.

apparent collapse on old-319 (6.9%) was entirely a parser bug: the arm writes "`Final answer: A.5.`" and the extractor failed all 141 multiple-choice rows it touched. The distillation arms' terse answers were under-scored by 19–30 pp. Corrected estimates predicted the subsequent blind-judge results within about 1 pp—but raw heuristic rankings and blind-judged rankings disagreed on nearly every pairwise comparison. Only blind-judged numbers appear in this paper.

6.2 The hacking analysis

The hint arms' large sparse-800 gains invited suspicion, so four reward-hacking hypotheses were tested against them (Table 6).

6.3 Hard gates before training

Every load-bearing mechanism was gated by a test that had to pass before any training step ran: file-level MD5 identity of the teacher and student tokenizers; a critic reachability gate confirming that image features influence the value head (identical token sequences with different pixels must produce different values—observed difference 1.93); a geometric-ratio self-test for ReOPD's cut-index sampler; the hint-unlock measurement of Section 3.2 before committing to the pool; and a projection test showing that HL-Gauss targets are normalized and mean-preserving.

6.4 Engineering discipline

All framework modifications are environment-gated patches with self-tests and `.orig` backups: with the gating variables unset, the framework's behavior is bit-identical to stock. Round 4 added three such patches (hybrid distillation gating, decoupled GAE, and the HL-Gauss critic).

The gates also caught failures in flight rather than in post-mortem: dynamically injected hints bypassed the framework's overlong-prompt filter until a 2,062-token sample killed all three hint arms at the same step; the 7B teacher's unprompted answers proved too short to replay (median 9 tokens) and had to be re-sampled with a stored lead-in phrase; and swapping in the categorical value head required replacing an inner submodule of the framework's composite value-head wrapper—an error caught by a smoke test, not by reading code.

7 Limitations

Single training seed. All eleven arms were trained with seed 42. Round-3 measurements of the same GRPO recipe across three seeds put the seed-to-seed standard deviation at roughly 2.4–2.7

pp (the range reflects two scoring pipelines). The headline between-group gaps (+8 to +14 pp in-domain) clear that noise comfortably, but differences under about 5 pp—including the entire ordering *within* the delivered-prior group—should not be ranked.

The pooled in-domain metric is a summary statistic. The 71.5/28.5 weighting of sparse-800 and old-319 reflects the sizes we happened to have, not a sampling design; the two slices are not random draws from one population, so the pooled number is a weighted summary rather than an unbiased in-domain accuracy. Equal weighting would lower the metric's predictive validity ($\rho = +0.40$ vs. $+0.76$; Section 4.2). The cleanest single in-domain number is sparse-800 alone.

old-319 is retained only for comparability. On the evidence of Section 4.2 it should not be read as a generalization measure; it appears in our tables solely to keep them comparable with rounds 1–3 of this project.

sparse-800 shares its distribution with training. The held-out set is drawn from the same pool and the same template families as the training data. Template-duplicate inflation is quantified at roughly 3 pp for the strongest arm (non-duplicate rows still score 45.9% against a 3.5% base); the load-bearing evidence for transfer is DynaMath, which is programmatically rendered and shares no items with MathV360K.

The predictivity of sparse-800 is an observation, not a rule. That a held-out slice of the training distribution predicts cross-domain transfer is not guaranteed in advance—a same-distribution set could equally have favored arms that overfit the training distribution. In this setting it did not: the arms that gained most on sparse-800 also transferred best. We accordingly draw the negative lesson (a drifted held-out slice can invert method rankings) as general, and the positive one (the hard in-domain tail was the best available proxy for transfer) as specific to this setting.

DynaMath-hard is underpowered. With n = 104, even the best arm's margin over base (`saf`, 18.3%) only reaches $p = 0.09$; we draw no per-arm conclusions from the hard split alone.

The comparison spans supervision regimes. The distillation arms never see ground-truth answers, while the reward-driven arms do. This is a property of the methods rather than an oversight, but comparisons between the two families should be read with that asymmetry in mind.

Judge variance. Repeated blind-judging passes over identical outputs vary by about ±0.5 pp (`grpo`: 40.5 / 41.1 / 41.0 across three passes); we treat differences at that scale as scoring noise.

Future work. The immediate next step is multi-seed replication of the group-level result. Three corrective arms are already in flight: an off-context importance correction for hint-conditioned rollouts [5], and variants that remove advantage std-normalization, which can pathologically amplify dense terms in near-all-wrong groups [21, 22]. Beyond that, we see the most leverage in making the value of a hint prefix itself measurable—branching student rollouts from teacher trajectories at graded depths—which would unify the text and distribution priors under one credit-assignment scheme.

8 Conclusion

On a pool of visual-math problems where a 2B vision–language model almost never succeeds and GRPO receives gradient from fewer than one rollout group in six, we compared eleven training arms spanning every prior-injection family proposed for this regime. The comparison localizes what matters. A prior helps exactly when it is delivered: hint-guided exploration, an annealed global teacher-KL, and a well-parameterized value critic all clear the no-prior baseline cross-domain, while pure imitation of a weak teacher, failure-gated distillation, and a critic trained by scalar regression do not. Within the families, the active ingredients are narrower than their papers suggest: the hint arms' gains survive removing UFT's auxiliary objective, and a one-line change of value-loss parameterization moves in-domain accuracy by +14.4 pp and flips the arm across the group boundary.

The finding we most expect to outlive the specific methods, however, is the evaluation one. A held-out set that had served as this project's generalization check for four rounds ranks the eleven arms in reverse order of their true cross-domain transfer—not through noise, but through composition: a near-chance multiple-choice majority rewards arms for leaving the base model's guessing behavior intact. Any project that tracks generalization on a fixed held-out slice is exposed to this failure mode, and it only became visible here once enough arms existed to correlate the evaluation sets against a genuinely out-of-domain measure. We suggest that check—made cheap by any multi-arm comparison—as standing practice, and we document the verification protocol (blind judging, scoring audits, pre-training gates, hacking analyses) that kept the numbers in this paper attached to reality. Multi-seed replication of the group-level result and a per-item decomposition of the inverted set are the immediate next steps.

References

[1] 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.

[2] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[3] 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. *arXiv preprint arXiv:2406.17294*, 2024.

[4] Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. UFT: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*, 2025.

[5] Priyank Agrawal, Ankur Samanta, Shervin Ghasemlou, Jalaj Bhandari, Kavosh Asadi, Daniel Jiang, and Aditya Modi. Off-context GRPO: Learning to reason on hard problems using privileged information. *arXiv preprint arXiv:2607.19313*, 2026.

[6] Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

[7] Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026.

[8] Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

[9] Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[10] Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop regressing: Training value functions via classification for scalable deep RL. *arXiv preprint arXiv:2403.03950*, 2024.

[11] Zhijian Zhou, Long Li, Xuan Zhang, Zongkai Liu, Yulei Qin, Ke Li, Xing Sun, Xiaoyu Tan, Chao Qu, and Yuan Qi. Start classifying: Categorical critics for LLM reinforcement learning. *arXiv preprint arXiv:2608.02181*, 2026.

[12] DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[14] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations (ICLR)*, 2025.

[15] Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

[16] 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. *arXiv preprint arXiv:2506.10947*, 2025.

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

[18] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[19] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

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

[21] 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. *arXiv preprint arXiv:2503.20783*, 2025.

[22] Yu Wang. The dark room in the reward channel: Dense prediction rewards collapse GRPO-trained LLM agents – and the channel, not the content, decides what works. *arXiv preprint arXiv:2607.21273*, 2026.

[23] Jiawei Gu, Xuhui Jiang, Zhichao Shi, et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[24] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

[25] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.