LEARNING TO SOLVE, FORGETTING TO RETAIN: CORRECT-SET TURNOVER IN RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) improves the reasoning capabilities of large language models, yet aggregate accuracy gains can conceal a hidden cost: previously solved problems may regress as training proceeds. We characterize this phenomenon as *correct-set turnover*, capturing the coupled dynamics of solution acquisition and regression, and make retention an explicit training objective alongside acquisition. In group-relative RLVR, mastered prompts provide no reward gradient when all rollouts are correct, yet can still drift under updates driven by other prompts. This observation motivates a *timely-review principle*: previously mastered prompts should be revisited before such regression accumulates. We propose **ReMind**, a retention-aware review mechanism that tracks mastered prompts, periodically revisits them with fresh rollouts, and requeues those that regress. Through pre-rollout batch replacement, ReMind preserves the per-step rollout budget. Across 20 image-text, video, and text-only reasoning benchmarks, ReMind improves average accuracy by **3.81** and **3.44** percentage points over GRPO and DAPO, respectively, while consistently improving retention across modalities and algorithms.

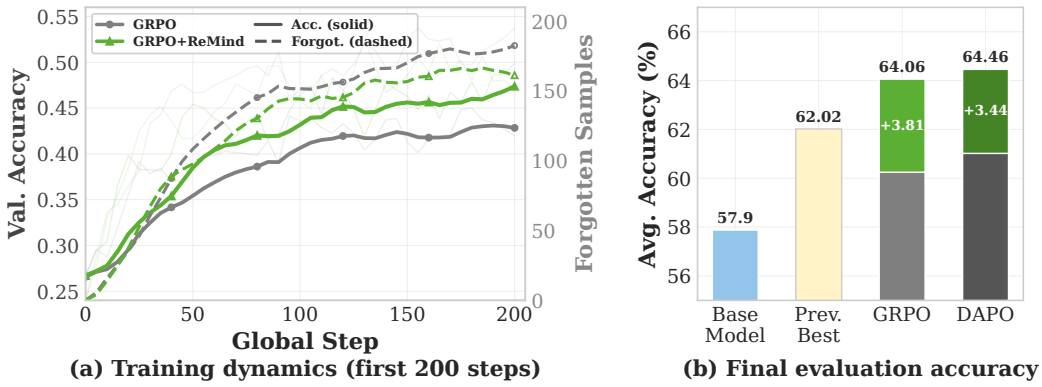


Figure 1: (a) Training dynamics during the first 200 steps: validation accuracy (solid lines, left axis) and correct-set loss (dashed lines, right axis). GRPO accumulates forgetting steadily while accuracy plateaus; ReMind reduces forgetting and sustains accuracy gains under the same rollout budget. (b) Average final accuracy across eight multimodal reasoning benchmarks. ReMind improves both GRPO and DAPO baselines.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become a standard paradigm for improving the reasoning capabilities of large language models (Guo et al., 2025; Shao et al., 2024; Yu et al., 2025). Headline accuracy gains, however, conceal a less visible cost: previously solvable problems can become unsolvable as training proceeds. RLVR has been shown to narrow the solution distribution, raising Pass@1 while reducing Pass@K (Yue et al., 2025b; Cui et al., 2025), while acquiring new reasoning skills can interfere with previously learned ones (Hu et al., 2026). Such

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.

regression occurs within the same training distribution, distinct from the cross-task setting typically studied in continual learning.

Rather than viewing forgetting only as a side effect, we frame training progress symmetrically as two opposing flows: acquisition, where new problems become solvable, and regression, where previously solved problems are lost. Of particular interest are *mastered* prompts, whose sampled rollouts are all correct at one point in training but may cease to be so after subsequent policy updates. We refer to these joint dynamics over the model's correct set as *correct-set turnover*. This view makes retention an explicit training goal alongside acquisition, shifting attention from net accuracy gains to both learning new solutions and retaining previously solved ones.

Mastered prompts have an important property under group-relative RLVR: when all rollouts are correct, their group-relative advantage vanishes and they no longer contribute a reward gradient (Shao et al., 2024; Le et al., 2026). Policy updates are instead driven by prompts with mixed outcomes, yet these shared-policy updates can still cause mastered prompts to drift (Liao et al., 2026). Paradoxically, such drift can restore learnability: once a previously mastered prompt begins producing both correct and incorrect rollouts, it regains non-zero group-relative advantages and becomes informative for optimization again. This motivates allocating a small fraction of the rollout budget to revisiting mastered prompts, so that re-emerging corrective signals can be detected and used without substantially displacing fresh training. The remaining question is *when* to spend this review budget. Inspired by the importance of review timing in spaced repetition (Ebbinghaus, 1885; Cepeda et al., 2006), we hypothesize that timely re-exposure should preserve mastered prompts more effectively than delayed review. Our controlled experiment in Figure 2 supports this hypothesis, showing larger retention gaps as review is delayed. Standard epoch-based RLVR pipelines (Shao et al., 2024) revisit prompts only passively across epochs, with no explicit record of prior mastery.

We propose **ReMind**, a lightweight retention-aware review mechanism that periodically reviews mastered samples to **remind** the model of previously solved solutions. ReMind maintains a queue of previously mastered prompts and periodically replaces a small fraction of fresh prompts with queued ones before rollout generation. Reviewed prompts pass through the standard rollout and policy-update pipeline: prompts that remain all-correct leave the queue, while those showing regression are re-enqueued for future review. This design preserves the per-step rollout budget while turning temporal changes in mastery into opportunities for corrective learning.

We evaluate ReMind across 20 benchmarks spanning image-text, video, and text-only mathematical reasoning with Qwen3-VL-8B-Instruct and Qwen2.5-7B-Math. On eight image-text benchmarks, ReMind improves GRPO, DAPO, and VCRL by 3.81, 3.44, and 1.54 percentage points on average, respectively, and outperforms replay and optimization-level baselines. These gains support compatibility with online filtering and curriculum learning. Training dynamics show lower correct-set loss with comparable acquisition coverage.

Our contributions are as follows.

- We formalize *correct-set turnover* in RLVR as the joint dynamics of acquisition and regression. Using a zero-overhead, graded mastery metric, we identify the *timely-review principle*: revisiting mastered prompts earlier improves retention, with larger retention gaps as review is delayed.
- We propose ReMind, a retention-aware review mechanism that maintains a mastery-keyed review queue and uses pre-rollout replacement to detect and repair regression without increasing the rollout budget.
- ReMind improves late-stage accuracy and retention over standard RLVR baselines and replay methods across 20 benchmarks, demonstrating robust generalizability across different modalities, models, and algorithmic backbones.

2 Related Work

RL for LLMs. RLVR has become a standard post-training method for reasoning models (Shao et al., 2024; Guo et al., 2025), with GRPO replacing the learned value model by group-relative comparison. Recent variants improve update stability and informativeness through dynamic sampling, clip-higher, and length-aware shaping (DAPO (Yu et al., 2025)), normalization bias correction (Liu et al., 2025), and diagnostics on entropy and token-level objectives (Cui et al., 2025; Dong et al.,

2026). A related filtering and curriculum line (Yu et al., 2025; Bae et al., 2025; Jiang et al., 2025) selects prompts whose rollouts are neither all correct nor all incorrect to keep useful gradients in the batch, treating uniformly correct prompts as low-value for the *next update*. RL-ZVP (Le et al., 2026) recovers learning signals from zero-variance groups, while MCPO (Liao et al., 2026) stabilizes mastered prompts during optimization. ReMind instead treats mastery temporally, revisiting previously mastered prompts as training continues.

Replay and forgetting. Experience replay reuses past data for efficiency (Lin, 1992; Mnih et al., 2015), with prioritization based on estimated learning value (Schaul et al., 2015). Recent LLM RL methods replay successful or informative trajectories along various axes, including verified solutions (Zhang et al., 2025a), difficulty and entropy buckets (Zhan et al., 2025), response reuse and early-state replay (Zhang et al., 2025b; Dou et al., 2025), and age-decayed priorities (Ma et al., 2026), but none defines priority by whether a once-mastered prompt has regressed. This differs from continual learning, where forgetting arises from sequential exposure to disjoint tasks (Kirkpatrick et al., 2017; Rebuffi et al., 2017); we study regression within a single RL training distribution. Our framing is closest to example-forgetting analysis, where forgetting is a sample-level $1\to0$ transition under single-evaluation labels (Toneva et al., 2019). This binary view is natural in supervised settings, but in group-based RLVR each prompt is already evaluated through multiple rollouts; treating these as a graded mastery estimate exposes partial regression that single-evaluation definitions cannot register, which ReMind uses to detect and repair regression during review. Recent RLVR work also reports broader regression phenomena, including cross-task capability loss (Phan et al., 2025), narrowed pass@k distributions (Yue et al., 2025b), and collapse dynamics (Hu et al., 2026), but does not turn sample-level mastery into an online training state for retention decisions.

3 Correct-Set Turnover and ReMind

3.1 Definition and Observation

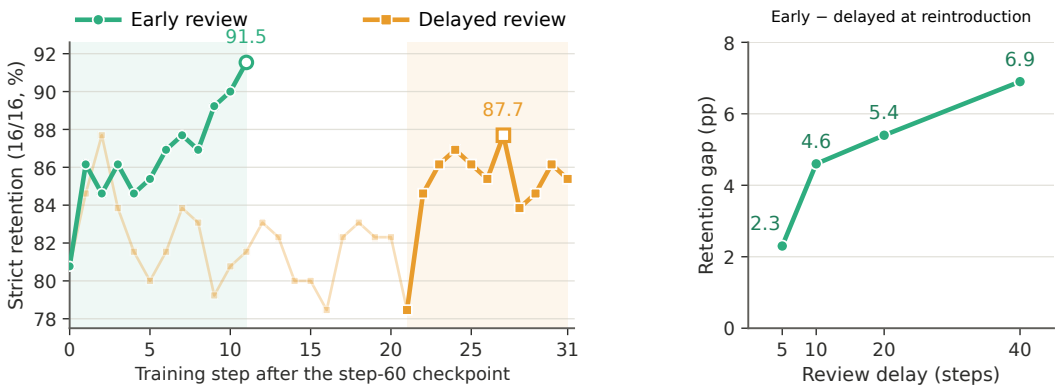
Let $\mathcal{P}$ denote the training set. For each sample $p \in \mathcal{P}$, we write $q_t(p) = \Pr[\pi_{\theta_t} \text{ solves } p]$ for the success probability under the current policy π_{θ_t}, estimated with K rollouts as $\hat{q}_t(p) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}[r_i = 1]$. A sample is *mastered* at step t if all rollouts are correct ($\hat{q}_t(p) = 1$), yielding the mastered set $M_t = \{p \in \mathcal{P} : \hat{q}_t(p) = 1\}$. Two properties of $\hat{q}_t(p)$ matter for what follows. It is *graded*: a previously mastered sample at $\hat{q}_t = 12/16$ shows partial regression that a single correct/incorrect outcome cannot distinguish from full mastery, whereas the binary signal used in prior example-forgetting work (Toneva et al., 2019) collapses this regime. It is also *free*: $\hat{q}_t(p)$ is the same per-prompt success rate GRPO computes for advantage estimation, so retention can be checked whenever a sample is included in the training batch without any auxiliary evaluation.

Comparing M_t with a later $M_{t+\Delta}$ decomposes the net change in mastery into two opposing flows: *acquisition*, where samples enter the mastered set ($p \notin M_t,\ p \in M_{t+\Delta}$), and *regression*, where samples leave it ($p \in M_t,\ p \notin M_{t+\Delta}$). We call this continual flux *correct-set turnover* and measure it with the retention rate:

$$\begin{aligned} \text{Retention}_t &= \frac{|M_t \cap M_{t+\Delta}|}{|M_t|}, \\ \text{Regression}_t &= 1 - \text{Retention}_t. \end{aligned} \tag{1}$$

Figure 1(a) illustrates turnover on the validation set. Training Qwen3-VL-8B-Instruct with standard GRPO, we track validation accuracy alongside the count of samples that were once solved but are currently wrong (correct-set loss). The two signals reveal a telling asymmetry: accuracy rises and then plateaus after roughly 100 steps, yet correct-set loss climbs steadily throughout the same window, indicating that new acquisitions are increasingly offset by regressions on previously solved samples.

We also examine whether mastered training samples regress. Across 1,275 review records of previously mastered samples, 269 (about 21%) contained both correct and incorrect rollouts, indicating partial regression from the earlier all-correct state.[1]



(a) Initial review windows (delay = 20)

(b) Delay sweep

Figure 2: Strict retention (16/16 rollouts correct) under early and delayed review. (a) Initial trajectories with a 20-step delay; shading highlights observation intervals. (b) Retention gaps at reintroduction (early minus delayed, in percentage points) increase with review delay.

3.2 The Timely-Review Principle

To see why mastered prompts can regress during training, consider which samples contribute to GRPO's reward gradient. Under binary rewards, all-correct and all-incorrect rollout groups have zero group-relative advantages. Let B_t denote the training batch and $g_t(p)$ the gradient of the GRPO reward loss for sample p. The batch reward gradient therefore comes only from groups with mixed outcomes:

$$g_t^{\mathrm{RL}} = \frac{1}{|B_t|} \sum_{\substack{p \in B_t \\ 0 < \hat{q}_t(p) < 1}} g_t(p). \tag{2}$$

The resulting updates change the shared policy, including its behavior on mastered samples. They can therefore lower a mastered sample's success probability even while that sample provides no reward gradient of its own.

Partial regression creates an opportunity to counter these changes: when a previously mastered sample produces both correct and incorrect rollouts, its nonzero group-relative advantages provide a signal for reinforcing correct responses. Under epoch-based sampling without replacement, however, a sample is revisited only in a later epoch, leaving this corrective signal unused in a single-epoch run. This motivates *review*: reintroducing previously mastered samples to generate fresh rollouts and use any resulting corrective signal. Samples that remain all-correct serve only to check retention. The next question is when to review these samples.

We compare immediate and delayed review of the same mastered prompts from a shared checkpoint. With a 20-step delay, retention is lower during most of the waiting period and improves after reintroduction (Figure 2(a)). Across delays of 5, 10, 20, and 40 steps, the retention gap relative to immediate review at the same training step grows with delay (Figure 2(b)).

These results establish a delay-sensitive retention effect: earlier review provides more timely opportunities to detect partial regression and reinforce correct responses. We refer to this as the *timely-review principle*. The next subsection operationalizes this principle with ReMind.

3.3 ReMind

ReMind implements the regular review motivated in §3.2 through a queue that periodically reintroduces previously mastered samples into the training batch before rollout generation. This provides review opportunities within the current epoch. On review steps, queued samples replace a small fraction of the regular batch and pass through the same rollout, reward, and policy-update pipeline,

[1]These records were collected during a ReMind training run; see §3.3 for the review procedure.

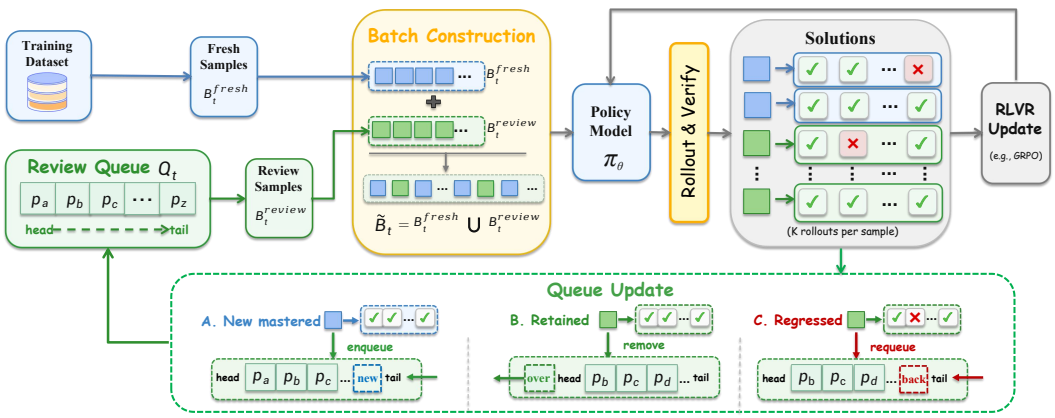


Figure 3: Overview of ReMind. The upper half combines fresh and queued review samples into a mixed batch $\widetilde{B}_t$ for rollout generation, verification, and the RLVR policy update. The lower half shows three queue-update cases: (A) newly mastered samples are enqueued, (B) retained samples are removed from the queue, and (C) regressed samples are reinserted at the tail for future review.

keeping the per-step rollout budget unchanged. Figure 3 illustrates the pipeline. Both mastery ($\hat{q}_t(p) = 1$) and regression on previously mastered samples ($\hat{q}_t(p) < 1$) are identified directly from the group rollouts GRPO already produces.

The method has three components: a FIFO review queue that stores mastered samples (§3.3.1), a batch-construction rule that mixes review and fresh samples on designated review steps (§3.3.2), and an update rule that routes regressed samples back into the queue for future repair (§3.3.3).

3.3.1 Review Queue

Following the mastery definition in §3.1, when a sample from the standard training stream satisfies $\hat{q}_t(p) = 1$, ReMind appends it to the review queue with probability p_{add}. The queue stores sample identifiers only; each subsequent review generates fresh rollouts from the current policy rather than replaying stored trajectories. FIFO ordering prioritizes samples that have waited longest since entering or re-entering the queue, providing a simple rule for allocating the review budget.

3.3.2 Batch Construction

Every f steps is designated a *review step*. On non-review steps, the batch is drawn entirely from the training sampler. On a review step, ReMind allocates a fraction r of the batch to review samples drawn from the queue Q_t, forming a mixed batch:

$$\widetilde{B}_t = B_t^{\text{fresh}} \cup B_t^{\text{review}},$$
$$|B_t^{\text{review}}| = \min\big(\lfloor r \cdot |\widetilde{B}_t| \rfloor,\ |Q_t|\big), \tag{3}$$

where B_t^{fresh} contains samples from the standard stream and B_t^{review} contains samples dequeued from the head of Q_t. This mixed batch then proceeds through rollout, reward computation, and policy update without modification. The review ratio r and period f together control the fraction of total training devoted to retention review.

3.3.3 Queue Update

After reward computation, the queue is updated according to three cases (lower panel of Figure 3). (1) Any sample from B_t^{fresh} that is mastered ($\hat{q}_t(p) = 1$) and not already in the queue is enqueued with probability p_{add}. (2) A review sample from B_t^{review} that retains full accuracy passes the current check and is not reinserted. We refer to removal after an all-correct review as graduation. (3) A review sample that regresses ($\hat{q}_t(p) < 1$) is reinserted at the queue tail for future repair. This ensures regressed samples cycle back for additional review while samples passing the current check exit.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Datasets. We train on MMFineReason-123K (Lin et al., 2026), a difficulty-filtered subset of MMFineReason-1.8M that retains only samples unsolved by Qwen3-VL-4B-Thinking (Bai et al., 2025) across four rollouts. We evaluate on eight image-text reasoning benchmarks covering math reasoning (MathVista (Lu et al.), MathVision (Wang et al., 2024), WeMath (Qiao et al., 2025), MathVerse (Zhang et al., 2024)) and general multimodal understanding (MMMU-Pro (Yue et al., 2025a), MMBench (Liu et al., 2024), MM-Star (Chen et al., 2024), ZeroBench (Roberts et al., 2025)).

Models and baselines. All image-text experiments use Qwen3-VL-8B-Instruct (Bai et al., 2025) on a shared GRPO-style backbone. We compare GRPO, DAPO, MCPO, RL-ZVP, ExGRPO, RLEP, RePO, and VCRL (Table 1).

Implementation. We build on EasyVideoR1 (Qin et al., 2026) with a batch size of 256 and $K = 8$ rollouts per prompt unless otherwise specified. For ReMind, we set the review ratio $r = 0.10$, review period $f = 5$, and enqueue probability $p_{\text{add}} = 0.25$, so that review samples constitute $\sim$2% of total training prompts. Full training hyperparameters and baseline descriptions are in Appendix H.

Table 1: Main results on eight image-text reasoning benchmarks. Best and second-best scores are **bold** and underlined; △ rows show gains over the corresponding base method.

Method	Math Reasoning					General Multimodal					Avg
	Vista	Vision	WeMath	Verse	Avg	MMMU-P	MMB	MM-S	ZeroB	Avg	
Base Model	73.80	47.37	54.10	54.61	57.47	51.75	89.79	71.83	19.76	58.28	57.88
ExGRPO	77.30	55.46	62.67	56.89	63.08	55.49	90.44	72.00	19.01	59.24	61.16
RLEP	78.50	54.23	62.48	58.91	63.53	55.38	90.45	72.27	19.61	59.43	61.48
RePO	78.00	53.77	63.52	59.39	63.67	54.95	90.45	72.60	23.50	60.38	62.02
MCPO	76.50	53.75	62.10	59.12	62.87	54.35	90.07	71.13	23.35	59.73	61.30
RL-ZVP	77.40	54.39	62.38	58.76	63.23	56.22	90.46	71.90	23.65	60.56	61.89
GRPO	76.20	48.82	56.57	59.52	60.28	55.78	90.29	72.20	22.60	60.22	60.25
GRPO + ReMind	79.10	56.81	66.38	61.52	65.95	58.78	90.41	73.40	**26.05**	62.16	64.06
△	+2.90	+7.99	+9.81	+2.00	+5.67	+3.00	+0.12	+1.20	+3.45	+1.94	+3.81
DAPO	77.60	52.78	57.43	59.55	61.84	56.85	90.53	72.77	20.66	60.20	61.02
DAPO + ReMind	**79.85**	57.32	68.29	60.37	66.46	57.89	**91.22**	**74.70**	**26.05**	62.47	64.46
△	+2.25	+4.54	+10.86	+0.82	+4.62	+1.04	+0.69	+1.93	+5.39	+2.27	+3.44
VCRL	77.75	56.81	64.76	61.31	65.16	58.25	90.53	72.93	25.00	61.68	63.42
VCRL + ReMind	79.45	**58.50**	**68.48**	**63.00**	**67.36**	**59.54**	90.60	74.53	25.60	**62.57**	**64.96**
△	+1.70	+1.69	+3.72	+1.69	+2.20	+1.29	+0.07	+1.60	+0.60	+0.89	+1.54

4.2 MAIN RESULTS

ReMind improves image-text reasoning across training algorithms. Starting from the same GRPO backbone, MCPO and RL-ZVP improve the overall average from 60.25 to 61.30 and 61.89, respectively (Table 1), supporting the value of explicitly handling high-accuracy and zero-variance prompts. ReMind further raises the average to 64.06 (+3.81 over GRPO), outperforming MCPO and RL-ZVP by 2.76 and 2.17 points, respectively, as well as RePO, the strongest trajectory-replay baseline, by 2.04 points.[2] Adding ReMind to DAPO and VCRL further improves their averages to 64.46 and 64.96, supporting its compatibility with dynamic sampling and curriculum learning. Together, these results are consistent with ReMind's temporal view that the utility of mastered prompts can re-emerge as training continues.

The gains extend to video understanding and text math reasoning. Across six benchmarks in each domain (Table 2), ReMind improves the video average over GRPO by 3.21 points and exceeds ExGRPO on text math by 1.23 points. Video-Holmes shows the largest video gain (+6.04); on MLVU, ReMind partially recovers GRPO's drop below the base model. Benchmark details are in Appendix G.

[2]Three additional independent GRPO+ReMind training runs yield 64.18 ± 0.03 overall (mean $\pm$ sample standard deviation); Appendix A reports the per-seed results.

Table 2: Generalization of ReMind to video understanding and text math reasoning. The base models are Qwen3-VL-8B-Instruct for video and Qwen2.5-7B-Math (Yang et al., 2024) for text math. Best results within each block are in **bold**.

(a) Video understanding

Method	MVB	Motion	MLVU	Holmes	LVR	VMQA	**Avg**
Base Model	69.50	56.10	**78.40**	45.30	77.30	47.80	62.40
GRPO	70.75	57.01	75.80	44.80	78.61	50.33	62.88
ReMind	**71.22**	**61.44**	77.37	**50.84**	**82.49**	**53.19**	**66.09**

(b) Text math reasoning

Method	AIME24	AIME25	AMC	MATH	Minerva	Olympiad	**Avg**
Base Model	12.50	10.20	48.50	80.40	32.70	41.00	37.55
SFT	22.20	22.30	52.80	82.60	**40.80**	43.70	44.07
SFT + RL	25.80	23.10	62.70	87.20	39.70	50.40	48.15
ExGRPO	32.30	25.70	65.60	87.60	40.10	57.00	51.38
ReMind	**33.85**	**26.25**	**67.85**	**90.80**	38.97	**57.93**	**52.61**

4.3 Ablation Studies

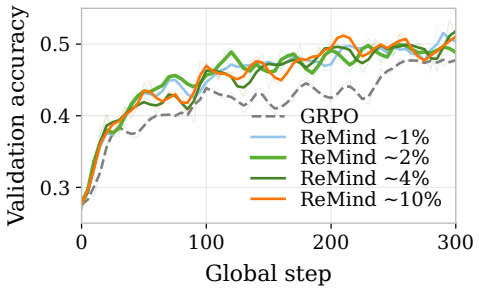


Figure 4: Review budget sensitivity.

ReMind outperforms GRPO at all tested review budgets. All configurations outperform GRPO (62.83–64.06 vs. 60.25; Table 3). Even 1% review avoids GRPO's mid-training plateau (Figure 4). Increasing the budget to 4% or 10% yields no further gain, possibly because more review displaces fresh training signal. At a matched 2% budget, burst scheduling ($r = 0.10$, $f = 5$) outperforms smooth ($r = 0.02$, $f = 1$) by 0.36 points. Accumulating reviews across several steps may yield a more diverse cohort for identifying regressed samples.

Table 3: Full results for review budget and schedule sensitivity. The upper block compares budget settings; the lower block compares burst and smooth schedules at matched ∼2% budget. Shaded rows are the default; benchmark abbreviations follow Table 1.

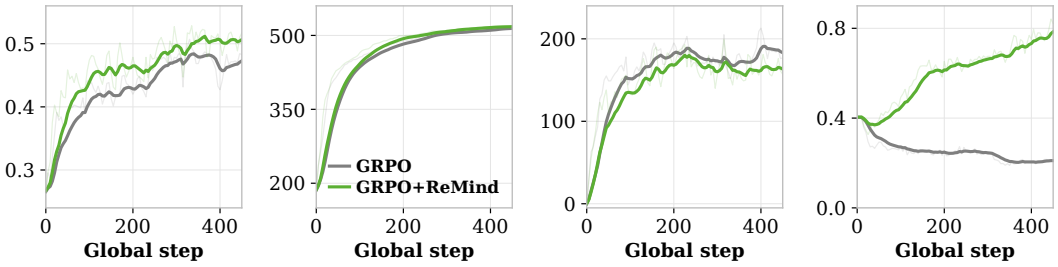
Setting	r	f	**Math Reasoning**					**General Multimodal**					**Avg**
			Vista	Vision	WeMath	Verse	Avg	MMMU-P	MMB	MM-S	ZeroB	Avg	
Budget ablation													
∼1%	0.05	5	78.45	55.20	63.33	60.15	64.28	57.76	90.68	73.10	25.15	61.67	62.98
∼2% (default)	0.10	5	79.10	56.81	66.38	61.52	65.95	58.78	90.41	73.40	26.05	62.16	64.06
∼4%	0.20	5	76.50	55.46	66.00	60.05	64.50	57.58	90.84	73.20	25.90	61.88	63.19
∼10%	0.10	1	76.25	55.46	65.24	60.37	64.33	57.43	90.78	72.27	24.85	61.33	62.83
Schedule ablation (matched ∼2% budget)													
Burst	0.10	5	79.10	56.81	66.38	61.52	65.95	58.78	90.41	73.40	26.05	62.16	64.06
Smooth	0.02	1	78.45	57.27	64.86	61.17	65.44	57.58	90.75	74.03	25.45	61.95	63.70

Regression-aware review outperforms random prompt review. We compare ReMind with two simple prompt-review strategies under the same approximately 2% review budget. *Random-seen* samples uniformly from all previously seen prompts, while *random-mastered* samples from previously mastered prompts without regression-aware queue updates. All strategies generate fresh rollouts under the current policy.

Random-seen and random-mastered reach 62.92 and 62.40 on average, respectively, versus GRPO's 60.25, supporting the value of prompt review across selection strategies. ReMind reaches 64.06, outperforming random-seen and random-mastered by 1.14 and 1.66 points, respectively. Table 6 in Appendix B reports the full eight-benchmark comparison.

Sensitivity to the mastery criterion. To assess sensitivity to how mastered samples are identified, we vary the mastery criterion and report full eight-benchmark results in Appendix C.2. The $K = 16$ all-correct criterion reaches 64.66 on average, compared with 64.06 for $K = 8$ and 60.25 for GRPO. At $K = 8$, the looser six-of-eight threshold scores 63.57, still above GRPO but below the all-correct criterion. This fixed-K comparison supports conservative mastery labeling, while gains over GRPO persist across the tested criteria.

4.4 Training Dynamics



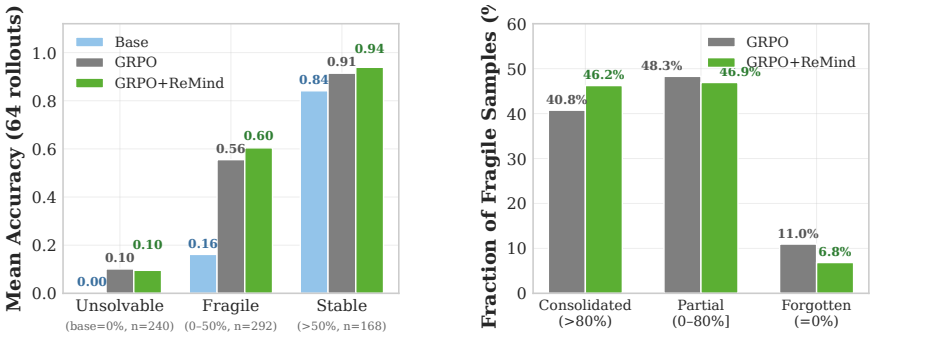
(a) Accuracy **(b) Acquisition** **(c) Correct-set loss** **(d) Entropy**

Figure 5: Training dynamics over 450 steps: (a) validation accuracy, (b) ever-solved prompts, (c) correct-set loss (count of previously solved prompts currently answered incorrectly), and (d) policy entropy. Faint lines show recorded values; solid lines show exponential moving averages.

Higher accuracy accompanies lower forgetting. Figure 5(a–c) shows similar acquisition coverage for both methods (∼500 ever-solved samples), while ReMind maintains higher accuracy and reduces correct-set loss from roughly 190 to 155 samples. First-solve and first-mastery timing, together with matched fresh-prompt exposure, further support that review preserves acquisition (Appendix F).

ReMind maintains higher policy entropy. In panel (d), GRPO's entropy declines as forgetting increases between steps 50 and 150, while ReMind recovers from the initial drop and maintains entropy near 0.7–0.8. Retaining more previously mastered solutions may help sustain a broader output distribution, with higher entropy as a by-product.

4.5 Further Analysis



(a) Accuracy by base-model difficulty **(b) Sample fate for fragile tier (n=292)**

Figure 6: Retention by difficulty (64 rollouts per prompt); full breakdown in Appendix D.

Retention gains concentrate on fragile samples that are solvable but unstable. Figure 6 stratifies 700 validation samples into three tiers by base-model accuracy: *unsolvable* (0%, n=240), *fragile* (0–50%, n=292), and *stable* (>50%, n=168). ReMind's improvement over GRPO is largest on the fragile tier (+4.9 pp) and negligible on unsolvable samples, consistent with the premise that review can only retain previously acquired knowledge. Within the fragile tier, ReMind consolidates 46.2% of samples to above 80% accuracy (vs. 40.8% under GRPO) and reduces complete forgetting from 11.0% to 6.8%, shifting fragile knowledge toward stable mastery rather than catastrophic loss.

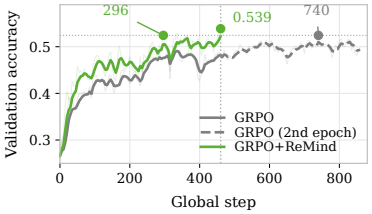


Figure 7: One-epoch ReMind versus two-epoch GRPO.

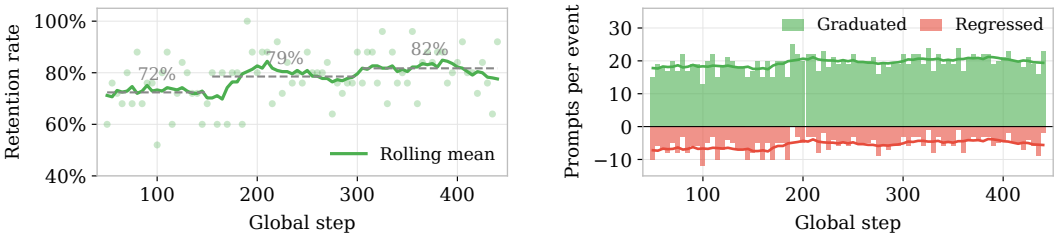
One epoch of ReMind outperforms two epochs of GRPO. A natural alternative to ReMind's timely review is simply training longer. Figure 7 compares one-epoch ReMind with two-epoch GRPO on the same data. Second-epoch GRPO fluctuates around 0.50 and finishes at 0.497, below ReMind's 0.539. ReMind also matches the two-epoch peak at step 296 versus 740, using about 2.5× fewer training steps. This comparison over full training runs complements the controlled early-versus-delayed review study in §3.2, supporting the value of timely, selective review over deferring re-exposure to another epoch.

The review mechanism adds negligible overhead. Table 4 reports wall-clock time for training, validation, and review; review accounts for only 3% of the total. Only a lightweight FIFO prompt queue is needed, with no additional model replicas, optimizer states, or stored historical trajectories, making ReMind easy to integrate into RLVR training loops.

Table 4: Wall-clock breakdown.

Component	Time	Share
Training	9.9 h	78%
Validation	2.5 h	19%
Review	0.3 h	**3%**
Total	12.7 h	100%

About 21% of prompt reviews produce mixed outcomes. We logged reviews of previously mastered samples to identify learning signals. In the lifecycle audit, 269 of 1,275 prompt-review records (21%) contain mixed outcomes, yielding nonzero group-relative advantages. These samples return to the queue; all-correct samples leave it. Figure 8 shows improving retention while some samples continue to regress. Peak queue occupancy is 8%, with no evictions (Appendix E).



(a) Review retention **(b) Graduation and regression**

Figure 8: 79 review events: (a) retention rates and phase means; (b) graduated and regressed counts (regressions below zero). Curves show centered nine-event means.

Most graduates pass later 16/16 audits, but failures remain. For simplicity, ReMind stops tracking samples after they pass review and leave the queue. To test whether they regress later, we evaluate these samples under ReMind and vanilla GRPO at matched later checkpoints, using 16 rollouts per prompt. Among samples that graduated by step 200, 11.5% no longer achieve all-correct results under ReMind, versus 15.3% under GRPO—a 25% relative reduction. Samples that graduated by step 260 show a similar pattern (Appendix E). Thus, review improves later retention, but some samples still regress after leaving the queue. We also tested uniformly returning graduated prompts to the review queue under the same ∼2% total review budget; this variant does not improve the eight-benchmark average (63.78 vs. 64.06; Appendix E.1).

5 Conclusion

We identify persistent *correct-set turnover* during accuracy plateaus in RLVR, and show that timely review can effectively counter regression. ReMind operationalizes the timely-review principle with a lightweight, self-correcting review queue; with only 3% additional cost, it yields consistent accuracy gains across image, video, and text reasoning with different modalities, models, and algorithmic backbones. More broadly, our results suggest that a meaningful share of the headroom in RLVR is not only about learning more, but also about losing less. Residual turnover persists even with review, suggesting that regression can remain persistent during RLVR optimization; we hope this work encourages the community to treat sample-level retention as a primary design axis in reasoning RL, alongside reward design, exploration, and curriculum that make forgetting visible and addressable.

Ethics Statement

This work proposes a training-time review mechanism for reinforcement learning with verifiable rewards; it does not introduce new data collection, human annotation, or model deployment, and all experiments use publicly available models and benchmarks. ReMind is designed to improve training stability and does not alter the model's safety properties beyond what the underlying RLVR training already provides. Because the method prioritizes retaining previously mastered samples, practitioners should monitor generalization when applying it to narrow training distributions to avoid overfitting to dataset-specific patterns.

Use of AI Assistants

We used AI Assistants to assist with writing polishing, including grammar checking and sentence-level rewording. All technical content, experimental design, and scientific claims were produced by the authors. The AI assistant was not used for data analysis, code generation, or result interpretation.

References

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. *arXiv preprint arXiv:2504.03380*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Nicholas J Cepeda, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3): 354, 2006.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos, 2025. URL https://arxiv.org/abs/2507.07966.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning?, 2025. URL https://arxiv.org/abs/2505.21374.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Yiming Dong, Kun Fu, Haoyu Li, Xinyuan Zhu, Yurou Liu, Lijing Shao, Jieping Ye, and Zheng Wang. Probing RLVR training instability through the lens of objective-level hacking. *arXiv preprint arXiv:2602.01103*, 2026.

Shihan Dou, Muling Wu, Jingwen Xu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. Improving RL exploration for LLM reasoning through retrospective replay. *arXiv preprint arXiv:2504.14363*, 2025.

Hermann Ebbinghaus. *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot, 1885.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models, 2025. URL https://arxiv.org/abs/2501.02955.

Sihan Hu, Xiansheng Cai, Yuan Huang, Zhiyuan Yao, Linfeng Zhang, Pan Zhang, Youjin Deng, and Kun Chen. Emergent slow thinking in llms as inverse tree freezing, 2026. URL https://arxiv.org/abs/2509.23629.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. VCRL: Variance-based curriculum reinforcement learning for large language models. *arXiv preprint arXiv:2509.19803*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kiXFIESZKv.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024b. URL https://arxiv.org/abs/2311.17005.

Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. Repo: Replay-enhanced policy optimization. *arXiv preprint arXiv:2506.09340*, 2025.

Zhaokang Liao, Yingguo Gao, Yi Yang, Yongheng Hu, and Jingting Ding. MCPO: Mastery-consolidated policy optimization for large reasoning models. *arXiv preprint arXiv:2604.16972*, 2026. URL https://arxiv.org/abs/2604.16972.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods, 2026. URL https://arxiv.org/abs/2601.21821.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3–4):293–321, 1992.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

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.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Weiyu Ma, Yongcheng Zeng, Yan Song, Xinyu Cui, Jian Zhao, Xuhui Liu, and Mohamed Elhoseiny. Freshness-aware prioritized experience replay for LLM/VLM reinforcement learning. *arXiv preprint arXiv:2604.16918*, 2026.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Hoang Phan, Xianjun Yang, Kevin Yao, Jingyu Zhang, Shengjie Bi, Xiaocheng Tang, Madian Khabsa, Lijuan Liu, and Deren Lei. Beyond reasoning gains: Mitigating general capabilities forgetting in large reasoning models. *arXiv preprint arXiv:2510.21978*, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Chuanyu Qin, Chenxu Yang, Qingyi Si, Naibin Gu, Dingyu Yao, Zheng Lin, Peng Fu, Nan Duan, and Jiaqi Wang. Easyvideor1: Easier rl for video understanding, 2026. URL https://arxiv.org/abs/2604.16893.

Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos, 2025. URL https://arxiv.org/abs/2506.05349.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning, 2017. URL https://arxiv.org/abs/1611.07725.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

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

Mariya Toneva, Alessandro Sordoni, Rémi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

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

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

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

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025a.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025b.

Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek F. Wong, and Yu Cheng. ExGRPO: Learning to reason from experience, 2025. URL https://arxiv.org/abs/2510.02245.

Hongzhi Zhang, Jia Fu, Jingyuan Zhang, Kai Fu, Qi Wang, Fuzheng Zhang, and Guorui Zhou. RLEP: Reinforcement learning with experience replay for llm reasoning. *arXiv preprint arXiv:2507.07451*, 2025a.

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

Yuheng Zhang, Wenlin Yao, Changlong Yu, Yao Liu, Qingyu Yin, Bing Yin, Hyokun Yun, and Lihong Li. Improving sampling efficiency in RLVR through adaptive rollout and response reuse. *arXiv preprint arXiv:2509.25808*, 2025b.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. URL https://arxiv.org/abs/2406.04264.

A Independent Training Runs

We repeat GRPO+ReMind training with three additional independent seeds in the main image-text setting. Table 5 reports all eight benchmark scores for each run, together with the math, general multimodal, and overall averages. We compute the mean and sample standard deviation across Seeds 1–3 using the unrounded averages. The original GRPO+ReMind run is listed separately and is not included in these statistics. The three runs achieve overall averages of 64.1575, 64.1600, and 64.2100, respectively, giving 64.18 ± 0.03 across runs. Their math average is 66.36 ± 0.27 and general multimodal average is 61.99 ± 0.24. All three overall averages are close to the original 64.06 result and exceed the reported GRPO reference by 3.91–3.96 points. The base-model and GRPO rows are reference results, not repeated-run estimates.

B Full Review-Strategy Results

Table 6 gives the eight-benchmark breakdown for §4.3. The three review strategies use the same approximately 2% review budget and generate current-policy rollouts for the selected prompts. Random-seen samples uniformly from previously seen prompts, while random-mastered restricts

Table 5: Full image-text evaluation for three additional independent GRPO+ReMind training runs. Benchmark abbreviations follow Table 1. Math and General each average four benchmarks; Overall averages all eight. The shaded rows report means and sample standard deviations over Seeds 1–3 only, computed before rounding.

Method / run	Math Reasoning					General Multimodal					Overall
	Vista	Vision	WeMath	Verse	Avg	MMMU-P	MMB	MM-S	ZeroB	Avg	
Base Model	73.80	47.37	54.10	54.61	57.47	51.75	89.79	71.83	19.76	58.28	57.88
GRPO	76.20	48.82	56.57	59.52	60.28	55.78	90.29	72.20	22.60	60.22	60.25
GRPO + ReMind	79.10	56.81	66.38	61.52	65.95	58.78	90.41	73.40	26.05	62.16	64.06
Seed 1	78.75	57.48	66.86	61.08	66.04	58.54	90.93	73.87	25.75	62.27	64.16
Seed 2	78.75	57.19	69.24	60.78	66.49	58.33	90.51	73.63	24.85	61.83	64.16
Seed 3	78.95	57.48	68.29	61.45	66.54	58.48	90.66	74.57	23.80	61.88	64.21
Mean (Seeds 1–3)	78.82	57.38	68.13	61.10	66.36	58.45	90.70	74.02	24.80	61.99	64.18
Std. (Seeds 1–3)	0.12	0.17	1.20	0.34	0.27	0.11	0.21	0.49	0.98	0.24	0.03

selection to previously mastered prompts without regression-aware queue updates. Both improve the overall average over GRPO, with random-seen exceeding random-mastered by 0.52 points; restricting review to previously mastered prompts alone provides no advantage in this comparison. ReMind achieves the highest average, with its largest advantage over either random strategy on ZeroBench.

Table 6: Full results for prompt-review strategies. All methods use Qwen3-VL-8B-Instruct; benchmark abbreviations follow Table 1. Best results in each column are bold; the shaded row is ReMind.

Method	Vista	Vision	WeMath	Verse	MMMU-P	MMB	MM-S	ZeroB	Avg
GRPO	76.20	48.82	56.57	59.52	55.78	90.29	72.20	22.60	60.25
Random-seen	78.65	55.89	64.57	58.50	**58.94**	90.79	**73.57**	22.46	62.92
Random-mastered	76.95	**56.96**	63.81	60.03	58.01	**91.07**	72.43	19.91	62.40
ReMind	**79.10**	56.81	**66.38**	**61.52**	58.78	90.41	73.40	**26.05**	**64.06**

C Additional Sensitivity Analysis

C.1 Review Budget and Schedule

Table 3 in §4.3 reports the full per-benchmark results for the budget and schedule configurations; we expand the analysis here.

Budget analysis. Across the four budget levels, math reasoning performance varies more than general multimodal performance (Math Avg spans 64.28–65.95 vs. General Avg 61.33–62.16), confirming that the review mechanism primarily benefits reasoning-intensive tasks where correct-set turnover is more frequent. The 10% budget ($f = 1$) underperforms the 4% budget despite allocating more capacity to review: at every step, 10% of the batch consists of previously mastered samples, reducing the diversity of fresh training signal. This suggests a trade-off between retention and acquisition that is well balanced at the 2% default.

Schedule analysis. The smooth schedule (reviewing 2% every step) achieves a slightly lower General Avg (61.95 vs. 62.16) and Math Avg (65.44 vs. 65.95) compared to burst. Both schedules improve over GRPO, with burst review achieving a 0.36-point higher overall average at the matched review budget.

Training curves. Figure 4 compares validation accuracy under the four review budgets. All configurations separate from GRPO early in training, complementing the final benchmark results in Table 3.

C.2 Mastery Criterion

Table 7 reports full results for the mastery-criterion comparison in §4.3. We compare the default 8/8 all-correct criterion, a 16/16 all-correct setting, and a looser threshold of at least six correct rollouts out of eight. At fixed $K = 8$, the all-correct criterion exceeds the looser threshold by 0.49 points on average. The $K = 16$ all-correct setting achieves the highest overall average (64.66), and every criterion improves over GRPO (60.25).

Table 7: Full mastery-criterion sensitivity results with Qwen3-VL-8B-Instruct. Benchmark abbreviations follow Table 1. Best results in each column are bold; the shaded row is the default criterion.

Method / criterion	Vista	Vision	WeMath	Verse	MMMU-P	MMB	MM-S	ZeroB	Avg
GRPO	76.20	48.82	56.57	59.52	55.78	90.29	72.20	22.60	60.25
GRPO + ReMind, 8/8	79.10	56.81	66.38	61.52	58.78	90.41	73.40	26.05	64.06
GRPO + ReMind, 16/16	79.35	**57.88**	**68.10**	**62.40**	**58.83**	90.60	73.90	**26.20**	**64.66**
GRPO + ReMind, $\geq$6/8	**79.65**	56.07	64.19	61.28	57.66	**90.83**	**74.20**	24.70	63.57

D Per-Difficulty Forgetting Breakdown

Table 8 provides a fine-grained breakdown of the per-difficulty analysis summarized in §4.5. We roll out each of the 700 validation samples 64 times under the base model (Qwen3-VL-8B-Instruct), vanilla GRPO, and GRPO+ReMind, and stratify samples into seven bins by the base model's mean accuracy.

Observations. The improvement from ReMind over GRPO is non-uniform across difficulty levels. ReMind yields its largest gains in the $(30, 70\%]$ base-success range (+6.4–6.7 pp over GRPO), while base-to-final degradation is most frequent in the $(0, 10\%]$ bin. Samples unsolved by the base model in 64 rollouts ($= 0\%$) show little difference between the two methods.

Across all 700 samples, ReMind reduces the fraction of regressed samples from 13.0% to 10.0%, with reductions in most difficulty bins. For example, regression falls from 24.1% to 16.8% in the $(0, 10\%]$ bin and from 21.8% to 14.5% in the $(30, 50\%]$ bin. In the $(90, 100\%]$ bin, the regression rate increases slightly from 12.6% to 13.8%, while both methods maintain near-perfect mean accuracy.

Mastery consolidation. Among the 292 fragile-tier samples ($0 <$ base accuracy $\leq 50\%$), GRPO consolidates 119 (40.8%) to above 80% accuracy, while ReMind consolidates 135 (46.2%). Conversely, GRPO completely forgets 32 samples (11.0%), compared with 20 (6.8%) under ReMind. This consolidation effect explains much of ReMind's advantage on the downstream benchmarks: by preventing fragile knowledge from decaying to zero and instead reinforcing it toward stable mastery, the review mechanism converts partial solutions into reliable performance.

E Queue Lifecycle and Post-Graduation Audits

Review-event dynamics. Figure 8 in §4.5 tracks 79 review events, containing 1,975 prompt-review records: 1,541 pass verification and 434 regress. These event-trace statistics are reported separately from the 1,275-record lifecycle audit below. The per-event retention rate rises from 72% in steps 50–150 to 82% in steps 300–440, while the mean number of regressed samples per event falls from 6.9 to 4.6. Late-stage reviews still contain 4–5 regressed samples per event, showing that regression persists throughout the observed run.

Queue lifecycle. We log queue updates during training to examine repeated review, graduation, and capacity usage. A reviewed prompt that regresses is re-enqueued for another review, while a prompt that passes the current-policy all-correct check graduates and leaves the queue. Table 9 summarizes the lifecycle audit: 269 of 1,275 prompt-review records contain mixed correct and incorrect rollouts (21.1%) and trigger re-enqueueing. The denominator counts review records, including repeat reviews, rather than distinct prompts. Of the 1,006 graduated prompts, 65 had previously regressed before passing a later review; median queue residence among graduates is 71 training steps. These records

Table 8: Fine-grained per-difficulty analysis. Each row corresponds to a bin of validation samples grouped by the base model's mean accuracy over 64 rollouts. **Mean Acc.** is the average accuracy within each bin; **Regressed** reports the percentage of samples whose accuracy is lower than that of the base model.

Base Accuracy Bin	n	Mean Acc.			Δ vs. GRPO		Regressed (%)	
		Base	GRPO	ReMind	Δ(Acc.)	Δ(Regr.)	GRPO	ReMind
$= 0\%$	240	0.000	0.101	0.096	−0.006	0.0	0.0	0.0
$(0, 10\%]$	137	0.047	0.443	0.479	+0.036	−7.3	24.1	16.8
$(10, 30\%]$	100	0.187	0.630	0.689	+0.059	−4.0	20.0	16.0
$(30, 50\%]$	55	0.403	0.699	0.763	+0.064	−7.3	21.8	14.5
$(50, 70\%]$	43	0.591	0.835	0.902	+0.067	−7.0	16.3	9.3
$(70, 90\%]$	38	0.806	0.859	0.883	+0.024	−2.6	21.1	18.4
$(90, 100\%]$	87	0.981	0.978	0.981	+0.004	+1.1	12.6	13.8
All	700	0.269	0.486	0.510	+0.024	−3.0	13.0	10.0

show that the queue supports repeated review of unstable prompts rather than treating the initial mastery observation as a permanent label. Peak occupancy was only 8% of the 20,000-prompt capacity, and no capacity-driven evictions occurred.

Table 9: Queue lifecycle audit for ReMind. Review counts refer to prompt-review records; graduation counts refer to distinct prompts.

Statistic	Value
Prompt-review records	1,275
Partial regression and re-enqueueing	269 (21.1%)
Graduated prompts	1,006
Graduated prompts with a prior regression	65
Median queue residence of graduates (steps)	71
Queue capacity (prompts)	20,000
Peak occupancy (% of capacity)	8%
Capacity-driven evictions	0

Post-graduation retention. We audit cohorts of prompts that graduated by training steps 200 and 260 at later checkpoints, using 16 rollouts per prompt. Strict retention requires all 16 rollouts to be correct; strict non-retention is the fraction of audited prompts that do not meet this criterion. For each cohort, we compare ReMind and vanilla GRPO at the same audit checkpoint. Table 10 shows that ReMind lowers strict non-retention from 15.3% to 11.5% for the by-step-200 cohort and from 16.4% to 12.3% for the by-step-260 cohort, approximately a 25% relative reduction in both cases. Graduation confirms an all-correct outcome at the review checkpoint; it does not guarantee permanent retention. The remaining 11.5–12.3% strict non-retention shows that later regression persists among graduated prompts even as ReMind improves their retention relative to GRPO.

Table 10: Post-graduation audits with 16 rollouts per prompt. Entries are strict non-retention rates (%, lower is better); the two methods are evaluated at the same later checkpoint for each cohort.

Graduated cohort	GRPO	GRPO + ReMind
By step 200	15.3	**11.5**
By step 260	16.4	**12.3**

E.1 Additional Review of Graduated Prompts

We evaluate a variant that periodically re-enqueues graduated prompts for review and training, using uniform rather than risk-based selection. The total review budget remains approximately 2%, so these additional reviews share the budget allocated to the review queue. Table 11 reports the full eight-benchmark evaluation with Qwen3-VL-8B-Instruct. This variant reaches 63.78 overall, compared

with 64.06 for standard ReMind and 60.25 for GRPO. It improves some benchmark scores but provides no overall gain over ReMind in this comparison.

Table 11: Additional review of graduated prompts at the same approximately 2% total review budget as standard ReMind. Benchmark abbreviations follow Table 1. Math and General average their four respective benchmarks; Overall averages all eight.

Method	Math Reasoning					General Multimodal					Overall
	Vista	Vision	WeMath	Verse	Avg	MMMU-P	MMB	MM-S	ZeroB	Avg	
Base Model	73.80	47.37	54.10	54.61	57.47	51.75	89.79	71.83	19.76	58.28	57.88
GRPO	76.20	48.82	56.57	59.52	60.28	55.78	90.29	72.20	22.60	60.22	60.25
GRPO + ReMind	79.10	56.81	66.38	61.52	65.95	58.78	90.41	73.40	26.05	62.16	64.06
w/ additional review	77.95	56.68	67.05	61.37	65.76	58.52	90.61	74.10	23.95	61.80	63.78

A possible explanation is that uniform additional review of graduated prompts diverts budget toward prompts that remain stable. In the by-step-200 audit, 88.5% of graduated prompts retain all-correct performance across 16 rollouts; allocating additional reviews to these prompts can reduce opportunities to revisit regressed prompts already in the queue. This interpretation is consistent with the fixed-budget result, but the aggregate scores do not directly measure budget displacement or changes in post-graduation retention. An adaptive schedule that prioritizes graduates at higher risk of regression is a direction for future work.

F Acquisition Timing and Fresh-Prompt Exposure

First-solve and first-mastery timing. To complement the cumulative ever-solved counts in §4.4, we examine when prompts are first solved and first meet the all-correct mastery criterion. Table 12 reports the median and 90th percentile of the recorded first-solve and first-mastery steps. With ReMind, median first-solve time decreases from 163 to 158 steps and median first-mastery time from 192 to 189 steps. The corresponding 90th percentiles also decrease, from 278 to 272 and from 286 to 280 steps. Thus, neither reported timing quantile indicates delayed acquisition under review.

Table 12: Per-prompt acquisition timing in training steps (lower is earlier). The 90th percentile is denoted by p90.

Metric	Median		p90	
	GRPO	ReMind	GRPO	ReMind
First solve	163	158	278	272
First mastery	192	189	286	280

Matched fresh-prompt exposure. Because review replaces approximately 2% of training prompts, we also compare acquisition after both methods have seen the same number of fresh prompts. At this matched exposure, ReMind has solved 242 more prompts and mastered 368 more prompts than GRPO. This comparison complements the timing statistics by aligning the amount of fresh-prompt exposure, and both analyses support that the default review budget preserves acquisition in this setting.

G Cross-Modal Evaluation Benchmarks

Table 13 summarizes the six video understanding benchmarks used in §4.2. They cover general perception, long-video comprehension, video reasoning, and STEM knowledge. For text math reasoning, we evaluate on AIME 2024/2025, AMC (Li et al., 2024a), MATH-500 (Hendrycks et al., 2021), Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024).

Table 13: Video understanding benchmarks used in the cross-modal evaluation (§4.2).

Benchmark	Samples	Metric
General Video Understanding		
MVBench (Li et al., 2024b)	3,586	Accuracy
MotionBench (Hong et al., 2025)	3,715	Accuracy
Long Video Understanding		
MLVU (Zhou et al., 2025)	502	Accuracy
Video Reasoning		
Video-Holmes (Cheng et al., 2025)	1,837	Accuracy
LongVideo-Reason (Chen et al., 2025)	851	Accuracy
STEM Knowledge		
VideoMathQA (Rasheed et al., 2025)	2,100	Accuracy

H Experimental Setup Details

Training data. MMFineReason-123K (Lin et al., 2026) is derived from the MMFineReason-1.8M corpus via difficulty-based filtering. Each candidate sample is rolled out four times with Qwen3-VL-4B-Thinking (Bai et al., 2025), and only samples on which the model fails every attempt are retained, yielding a training set of 123K challenging multimodal reasoning problems.

Evaluation benchmarks. We evaluate on eight multimodal reasoning benchmarks spanning visually grounded mathematics, broad subject knowledge, and fine-grained visual understanding. MathVista (Lu et al.), MathVision (Wang et al., 2024), WeMath (Qiao et al., 2025), and MathVerse (Zhang et al., 2024) target mathematical reasoning in visual contexts, from general problem solving to competition-level diagrams and structured difficulty tiers. MMMU-Pro (Yue et al., 2025a) and MMBench (Liu et al., 2024) test broad multi-discipline knowledge and general multimodal competence. MM-Star (Chen et al., 2024) stresses fine-grained visual discrimination, and ZeroBench (Roberts et al., 2025) serves as a hard stress test for the tail of the reasoning distribution.

Baselines. All baselines use the same base policy and GRPO-style training framework, with method-specific objectives and sampling rules. GRPO (Shao et al., 2024) is the pure on-policy reference. DAPO (Yu et al., 2025) augments GRPO with dynamic sampling and online filtering. MCPO (Liao et al., 2026) regularizes policy drift on mastered prompts through hinge-KL regularization and reweights majority-correct prompts. RL-ZVP (Le et al., 2026) constructs entropy-guided advantages for zero-variance groups, retaining the GRPO formulation for groups with mixed outcomes. VCRL (Jiang et al., 2025) uses a variance-based curriculum for prompt selection. We combine ReMind with both DAPO and VCRL to evaluate its compatibility with dynamic sampling and curriculum learning, using the default approximately 2% review budget. ExGRPO (Zhan et al., 2025) reuses successful trajectories from historical replay buffers. RLEP (Zhang et al., 2025a) replays trajectories from a far-future checkpoint. RePO (Li et al., 2025) collects early on-policy rollouts and revisits them asynchronously. We also report the base model (Qwen3-VL-8B-Instruct) without RL post-training as a reference point.

Implementation details. Our implementation is built on the EasyVideoR1 (Qin et al., 2026) framework. The maximum context length is 8192 tokens, split into a prompt budget of 4096 and a response budget of 4096, applied consistently at training and evaluation time. We train with a learning rate of 1×10^{-6} and a batch size of 256, sampling $K = 8$ rollouts per prompt at temperature 1.0. Clipping thresholds are set to $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$; the GRPO backbone omits the standard KL penalty and entropy regularization. For ReMind, we set the review ratio $r = 0.10$, review period $f = 5$, and enqueue probability $p_{\text{add}} = 0.25$. Under these settings, review samples constitute approximately 2% of total training prompts on average. All experiments are run on 4 compute nodes, each with 8 NVIDIA H200 140GB GPUs.

I Training-Time Breakdown

Table 4 in §4.5 reports timings for the 12.7 h training run. Standard training includes rollout generation and policy updates, and validation is performed periodically. Review overhead is measured as the marginal step-time increase on review steps relative to non-review steps, covering review rollouts and queue management. Review requires only a prompt queue, without additional model replicas, optimizer states, or stored historical trajectories.

J Limitations

ReMind inherits the standard RLVR assumption that the verifiable reward is correct; because mastery status directly governs which samples enter the review queue, systematic reward errors (e.g., false positives from imprecise answer extraction) would be more consequential here than in methods that do not condition on mastery. Our experiments cover 7–8B-scale models on mathematical and multimodal reasoning; whether the optimal review budget or the magnitude of retention gains transfers to substantially larger models or other domains such as code generation remains to be verified. ReMind reduces but does not eliminate correct-set regression; our current design uses a fixed review schedule, and adaptive scheduling that responds to the model's regression dynamics could further narrow this gap.