LEARNING FROM THE NEAR FUTURE: TEMPORAL SELF-DISTILLATION FOR RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) is a core post-training recipe for reasoning models, yet pure on-policy learning can be inefficient when useful trajectories are difficult to discover or exploration narrows. Existing self-guided approaches largely reuse capability already available to the current or earlier learner. We instead ask whether learning can also make use of capabilities that emerge later in training: *can a model learn from its own future self?* We introduce **temporal self-distillation**, in which a policy receives guidance from a stronger later checkpoint of itself. We hypothesize that the most useful temporal teacher need not be the strongest one: a teacher must provide sufficiently new capability while remaining compatible enough for that capability to be readily transferred, motivating a **near-future** regime. We study this principle through two complementary mechanisms. **Near-Future Policy Optimization (NPO)** performs off-policy behavioral transfer using verified future-self trajectories, while **Near-Future Policy Distillation (NPD)** performs on-policy token-level transfer on learner-generated trajectories. We further introduce **AutoNPO**, which adaptively determines when temporal guidance is useful and how far to roll back, turning future-self guidance into a repeated self-bootstrap process. Across eight image-text benchmarks, NPO improves GRPO from 60.25 to 62.84 and AutoNPO reaches 63.15, with consistent gains on text-only and video reasoning. Under NPD, a near-future teacher reaches 63.23 after continued RL versus 61.92 with a far-future teacher, despite lower immediate post-distillation performance. Together, these results suggest that effective temporal self-distillation depends not simply on teacher strength, but on a balance between newly acquired capability and learner compatibility.

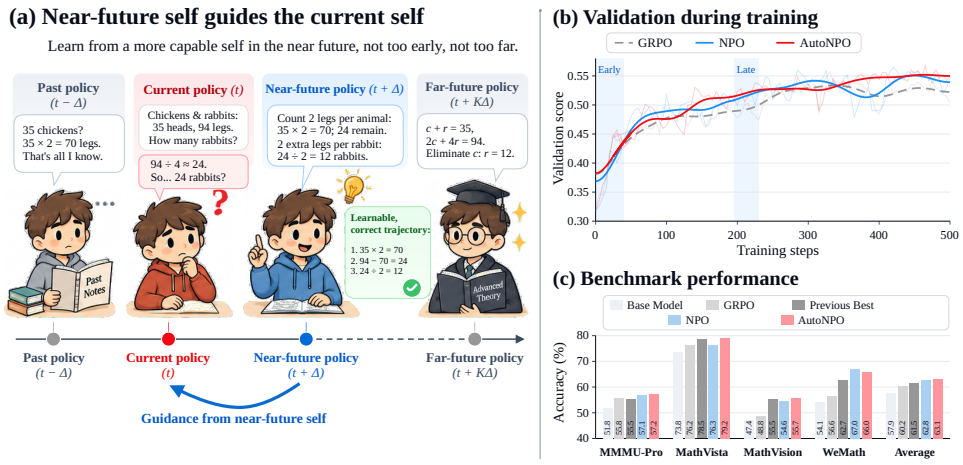


Figure 1: **(a)** Guidance from a near-future self. **(b)** Validation curves: thin/thick lines show raw/smoothed scores; shading marks manual interventions. **(c)** Representative image-text results.

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.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become a core post-training recipe for reasoning models (Guo et al., 2025; Shao et al., 2024; Yu et al., 2025). Yet pure on-policy learning can be inefficient: correct trajectories are scarce early on, while later the rollout distribution may narrow and performance can plateau (Yue et al., 2025b; Zhang et al., 2025). Existing methods therefore supplement on-policy learning with signals from stronger external models (Yan et al., 2025; Huang et al., 2025; Ma et al., 2025; Fu et al., 2025) or from experience accumulated during training (Zhan et al., 2025; Zhang et al., 2025). External guidance can provide capabilities beyond the learner's reach but may be difficult to absorb, whereas historical experience remains closer but is limited by what earlier policies had discovered.

A related direction makes learning more self-guided, reusing successful traces (Gulcehre et al., 2023; Zelikman et al., 2022) or deriving supervision from the current model conditioned on solutions or feedback (Zhao et al., 2026; Hübotter et al., 2026). Such methods largely look *backward or inward*. Training, however, naturally produces increasingly capable yet closely related policies, with later checkpoints acquiring solutions unavailable to earlier ones. This raises a natural question: *can a model learn from its own future self?* We call this setting **temporal self-distillation**.

The central problem is then *which future self to learn from*. A teacher that is too close may contribute little new capability, while one that is too far may be difficult to absorb. We characterize this tension through **signal quality**, which captures useful capability beyond the current policy, and **guidance cost**, which captures transfer difficulty. As checkpoint distance Δ grows, capability gain may increase, but so can transfer cost. Our pilot study reveals an intermediate regime where these effects are best balanced, suggesting that a moderately stronger future self may provide more useful guidance than either a very close or a much stronger distant teacher.

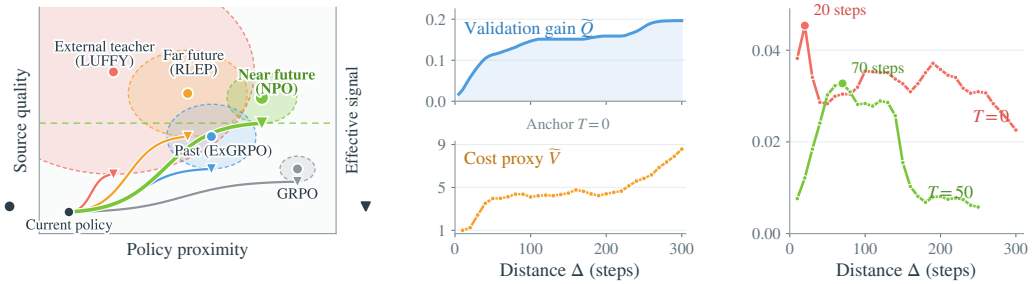
We study this principle through two complementary transfer mechanisms. **Near-Future Policy Optimization (NPO)** performs off-policy behavioral transfer by bringing verified trajectories from a future policy back to an earlier learner while preserving the RL objective. **Near-Future Policy Distillation (NPD)** builds on on-policy distillation (Agarwal et al., 2024; Lu & Lab, 2025), providing token-level teacher distributions on learner-generated trajectories. Together, they decouple *temporal teacher selection* from the *mechanism of knowledge transfer*, allowing us to test whether the near-future advantage extends beyond off-policy trajectory transfer.

While NPO and NPD assume a chosen temporal teacher, practical training also requires deciding when future guidance is useful and how far to look. We therefore introduce **AutoNPO**, which automatically determines when to intervene and how far to roll back. AutoNPO turns temporal self-distillation into a repeated self-bootstrap process: later-acquired capability strengthens an earlier learner, which continues training and may produce an even stronger future guide. Thus, improvement itself becomes a reusable learning signal. Although the intervention remains hand-designed, this recursive structure points toward more self-directed improvement.

Empirically, NPO improves the eight-benchmark image-text average from 60.25 to 62.84, while AutoNPO reaches 63.15, with consistent gains on text-only and video reasoning. Under NPD, a near-future teacher reaches 63.23 after continued RL versus 61.92 for a far-future teacher, despite lower immediate post-distillation performance. This cross-mechanism result suggests that the near-future advantage is not merely an artifact of reduced off-policy trajectory shift.

Our contributions are as follows.

- We introduce **temporal self-distillation**, in which a model reuses capabilities acquired by later versions of itself, and frame temporal teacher selection through a trade-off between additional capability and learner compatibility.
- We instantiate this principle through **NPO** for off-policy behavioral transfer and **NPD** for on-policy distributional transfer, and further introduce **AutoNPO** to adaptively determine when and how far temporal guidance should be applied.
- Across image-text, text-only, and video reasoning, temporal self-distillation consistently improves RLVR, while NPD shows that the near-future advantage persists beyond off-policy trajectory transfer.



(a) Quality–proximity trade-off **(b) Quality and cost** **(c) Effective guidance score**

Figure 2: **Why the near future can be a useful guide.** **(a)** Conceptual comparison of guidance sources by signal quality and learner compatibility. **(b)** Along a GRPO training trajectory, farther checkpoints can provide stronger capability while also increasing source–learner discrepancy. **(c)** Guidance scores for two anchor policies, showing that the preferred temporal distance can shift over training. Details are provided in Appendix A.6.

2 Motivation: Why Learn from the Near Future?

Temporal self-distillation introduces a natural question: *how far into the future should the learner look?* A teacher that is too close may contribute little new capability, while one that is too far may be difficult for the learner to absorb. We characterize this trade-off through two properties of a guidance source: **signal quality** Q, which captures useful capability beyond the current learner, and **guidance cost** V, which captures the difficulty of transferring that capability back.

Here, Q and V are general conceptual quantities whose concrete realization depends on the transfer mechanism. In mixed-policy RL, Q can be estimated from additional verified successes, while V is closely related to off-policy variance and can be approximated using source–learner discrepancy measures such as token-level KL divergence, log-ratio variance, or token log-probability gap.

Checkpoint distance Δ naturally controls this trade-off. Looking farther ahead gives the policy more opportunity to acquire new capabilities, potentially increasing Q, but also more opportunity to diverge from the learner, increasing V. This motivates the **near-future hypothesis**: the most useful temporal teacher need not be the strongest one, but may lie at an intermediate distance where additional capability remains readily transferable.

Figure 2(b,c) provides a pilot study along a GRPO training run. As Δ increases, later checkpoints provide stronger capability but also become more mismatched with the learner. Among the candidates examined, the preferred teacher lies at an intermediate temporal distance, and this distance shifts with training stage: from $\Delta = 20$ for the early anchor $T = 0$ to $\Delta = 70$ for the later anchor $T = 50$. This suggests that temporal teacher selection should be learner-dependent rather than fixed.

These observations motivate the methods that follow. **NPO** and **NPD** test whether the near-future preference persists under behavioral and distributional transfer, respectively, while **AutoNPO** adapts when to intervene and how far along the learning trajectory to look.

3 Near-Future Policy Optimization

Having motivated temporal teacher selection, we now present the primary off-policy realization of temporal self-distillation: **Near-Future Policy Optimization (NPO)**. NPO transfers verified behavior from a nearby future policy to an earlier learner through trajectory-level mixed-policy RL, while preserving the underlying RLVR objective.

3.1 Off-Policy Trajectory Transfer

Consider an anchor checkpoint $\pi^{(t)}$. Starting from this checkpoint, we continue training for Δ steps to obtain a future policy $\pi_F = \pi^{(t+\Delta)}$. We then restore the anchor checkpoint and initialize the

replayed learner as $\pi_L \leftarrow \pi^{(t)}$. The future policy remains frozen and serves only as a source of verified guidance behavior during the replayed segment.

Before replay begins, we roll out π_F offline on each prompt x, verify its responses, and retain one correct trajectory o'_x when available (optionally with length or entropy filtering). Prompts for which the future policy produces no verified-correct response are excluded from the guidance cache. The cache is constructed once and reused throughout the NPO segment, so the future policy does not need to be queried during each replay update.

At each replay step, the current learner samples an on-policy group of n trajectories,

$$o_i \sim \pi_L(\cdot \mid x), \qquad i = 1, \ldots, n, \tag{1}$$

and we compute the group pass-rate

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} r(x, o_i). \tag{2}$$

Guidance is injected only when the learner is still struggling with the prompt. If $\hat{p}(x) \leq \tau_{\text{gate}}$ and a cached trajectory o'_x exists, we replace one rollout-group slot with the verified future-self trajectory; otherwise the group remains fully on-policy. Writing the final slot as $\tilde{o}_n$, the NPO group is

$$\mathcal{G}_{\text{NPO}}(x) = \{o_1, \ldots, o_{n-1}, \tilde{o}_n\}, \qquad \tilde{o}_n = \begin{cases} o'_x & \text{if } \hat{p}(x) \leq \tau_{\text{gate}} \text{ and } o'_x \text{ exists,} \\ o_n & \text{otherwise.} \end{cases} \tag{3}$$

Group-relative advantages are computed over $\mathcal{G}_{\text{NPO}}(x)$ in the same way as standard GRPO,

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{n})}{\text{std}(\{r_j\}_{j=1}^{n})}, \tag{4}$$

and the learner is updated using the usual clipped objective,

$$\mathcal{L}_{\text{NPO}}(\theta) = \mathbb{E}_{x, \mathcal{G}_{\text{NPO}}(x)} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{|o_i|} \sum_{\ell=1}^{|o_i|} \min \left(\rho^q_{i,\ell}(\theta) A_i,\ \text{clip}\big(\rho^q_{i,\ell}(\theta), 1-\epsilon, 1+\epsilon\big) A_i \right) \right], \tag{5}$$

where

$$\rho^q_{i,\ell}(\theta) = \frac{\pi_\theta(o_{i,\ell} \mid x, o_{i,<\ell})}{q_i(o_{i,\ell} \mid x, o_{i,<\ell})}. \tag{6}$$

For ordinary rollout slots, q_i is the learner policy that generated the trajectory; for the guidance slot, $q_i = \pi_F$. The formulation therefore admits the standard source-policy importance correction for the off-policy guidance trajectory. In practice, Section 4.4 shows that omitting this correction has only a marginal effect for the near-future guides considered here.

NPO therefore changes the *source* of one rollout slot rather than the learning rule itself: the reward, verifier, group-relative advantage construction, and clipped RLVR objective remain unchanged. In this sense, NPO implements temporal self-distillation through off-policy behavioral transfer: the future policy contributes verified behavior rather than token-level teacher probabilities.

3.2 Manual Temporal Interventions

We first study two manually chosen intervention points to probe when near-future transfer is useful. Both use the same mechanism in Section 3.1; they differ only in when the anchor is chosen and how the future guide is obtained.

Early-Stage Bootstrapping. Early in training, many rollout groups contain no correct trajectory and therefore provide little or no useful group-relative learning signal. We run a short scout continuation from the base checkpoint, use its final checkpoint as the future guide, and then restore the base model for replay with NPO. Because the guide is only slightly ahead, it can expose verified solutions that the fresh learner does not yet sample reliably while remaining close to the restarted policy. This accelerates learning through the sparse-reward regime.

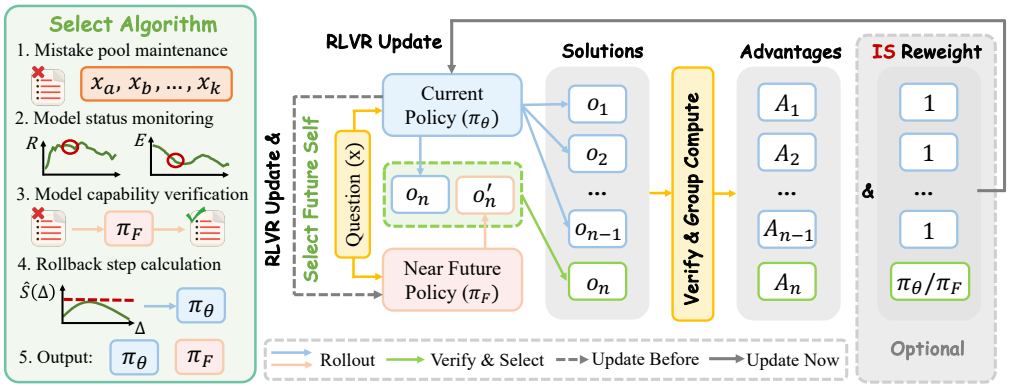


Figure 3: **Overview of NPO and AutoNPO. Right:** NPO (§3.1) transfers a verified trajectory from a near-future policy π_F into one slot of the learner's rollout group when the learner is still struggling with the prompt. **Left:** AutoNPO (§3.3) uses online training signals, a mistake pool, a capability probe, and an estimated quality–cost score $\hat{S}$ to decide *when* to intervene and *how far* to roll back.

Late-Stage Plateau Intervention. Later in training, the validation curve may flatten and a stable subset of prompts can remain unsolved. We continue training beyond the observed plateau, take a stronger nearby checkpoint as the future guide, and then restore and replay the plateaued segment with NPO. The future guide contributes verified-correct trajectories on prompts the earlier learner still fails, helping the replayed policy escape the observed plateau and reach a higher final performance level.

Takeaway. The same transfer rule addresses sparse early signals and limited later exploration, motivating adaptive choices of *when* to intervene and *how far* to roll back.

3.3 AutoNPO: Adaptive Temporal Selection

The pilot study in Section 2 suggests that the appropriate temporal distance is learner-dependent and can change over the course of training. **AutoNPO** operationalizes this observation by adaptively deciding *when* retrospective guidance is useful and *how far* along the training trajectory to roll back.

To decide *when*, AutoNPO maintains a lightweight mistake pool of recently failed prompts. Reward stagnation together with declining policy entropy raises a warning; if this pattern persists, a confirmation probe tests whether the current policy has acquired new capability on earlier failures. An intervention is triggered only when the probe indicates that sufficiently useful guidance has emerged.

To decide *how far*, AutoNPO ranks the available rollback distances according to the quality–cost trade-off introduced in Section 2. We operationalize this trade-off using

$$\hat{S}(\Delta) = \frac{\hat{Q}(\Delta)}{\hat{V}(\Delta)}, \tag{7}$$

where $\hat{Q}(\Delta)$ estimates the additional verified capability available from the future guide and $\hat{V}(\Delta)$ is a KL-based proxy for the guidance cost induced by source–learner mismatch. AutoNPO selects

$$\Delta^\star = \arg\max_{\Delta} \hat{S}(\Delta). \tag{8}$$

The current policy then generates verified guidance for the selected failed prompts, training rolls back to $t - \Delta^\star$, and the segment is replayed with NPO. Once replay catches up to step t, the controller enters a short cooldown and re-arms, allowing the strengthened learner to continue training and potentially serve as an even stronger future guide at a later intervention.

AutoNPO therefore turns a fixed temporal intervention into an adaptive self-bootstrap process: capability acquired later in training can be reused to reshape an earlier learner, whose continued optimization may in turn produce stronger future guidance. Implementation details and full pseudocode are provided in Appendix C

Table 1: Multimodal reasoning results on Qwen3-VL-8B-Instruct (%, higher is better). OPSD and SDPO use self-distillation; OPD uses a frozen Qwen3-VL-32B-Instruct teacher. Blue rows denote NPO variants. **Bold** marks the best result per column; *Avg.* is the unweighted mean over eight benchmarks.

Method	MMMU-Pro	MathVista	MathVision	ZeroBench	WeMath	MMBench	MM-Star	MathVerse	Avg.
Qwen3-VL-8B-Instruct	51.75	73.80	47.37	19.76	54.10	89.79	71.83	54.61	57.88
Distillation methods									
OPSD	53.58	75.10	47.53	21.06	54.95	89.70	71.57	58.03	58.94
SDPO	52.64	74.00	47.27	25.15	52.19	90.03	72.21	58.44	58.99
OPD (32B teacher)	53.42	76.95	48.57	21.71	58.19	90.15	72.00	53.67	59.33
RLVR methods									
GRPO (pure on-policy)	55.78	76.20	48.82	22.60	56.57	90.29	72.20	59.52	60.25
LUFFY (external teacher)	54.23	73.80	54.00	20.51	52.38	89.49	69.47	55.58	58.68
ExGRPO (historical replay)	55.49	77.30	55.46	19.01	62.67	90.44	72.00	56.89	61.16
RLEP (far future)	55.38	78.50	54.23	19.61	62.48	90.45	72.27	58.91	61.48
NPO (early only)	56.85	76.60	54.31	**26.35**	62.76	90.41	70.30	59.38	62.12
NPO (early + late)	57.07	76.30	54.61	24.85	**66.95**	90.30	**72.67**	**60.00**	62.84
AutoNPO	**57.24**	**79.20**	**55.72**	24.70	66.00	**90.63**	72.63	59.11	**63.15**

4 EXPERIMENTS

We organize the experiments around three questions: (i) whether near-future off-policy transfer improves RLVR across training stages and domains; (ii) whether the same temporal advantage persists under on-policy distillation; and (iii) whether the observed gains are consistent with the proposed quality–cost view.

4.1 EXPERIMENTAL SETUP

Data and Benchmarks. We use the MMFineReason-123K data preparation of Yang et al. (2026), a difficulty-filtered subset of MMFine-Reason-1.8M (Lin et al., 2026) on which Qwen3-VL-4B-Thinking (Bai et al., 2025) fails all four rollouts. We evaluate on eight image-text reasoning benchmarks, six text-only reasoning benchmarks, and six video reasoning benchmarks; the full benchmark list is provided in Appendix A.1. We report accuracy (%) and the unweighted mean *Avg.*.

Models and Baselines. The image-text experiments use Qwen3-VL-8B-Instruct (Bai et al., 2025) as the base policy. For RLVR, we compare against GRPO (Shao et al., 2024), LUFFY (Yan et al., 2025), ExGRPO (Zhan et al., 2025), and RLEP (Zhang et al., 2025), covering pure on-policy learning, external-teacher guidance, historical replay, and longer-distance future guidance. We also include the distillation baselines OPSD (Zhao et al., 2026), SDPO (Hübotter et al., 2026), and OPD (Agarwal et al., 2024; Lu & Lab, 2025). OPD uses a frozen Qwen3-VL-32B-Instruct teacher to supervise the student's own rollouts at the token level.

Implementation Details. Our NPO implementation is built on EasyVideoR1 (Qin et al., 2026). The default GRPO configuration uses an 8192-token context, batch size 256, $n = 8$ rollouts per prompt, sampling temperature 1.0, learning rate 1×10^{-6}, and clipping thresholds $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. NPO injects future-self guidance when $\hat{p}(x) \leq \tau_{\text{gate}}$, with $\tau_{\text{gate}} = 0.6$ by default. The distillation experiments use token-level Jensen–Shannon divergence on student-generated responses. Full training and teacher configurations are given in Appendix A.1.

4.2 OFF-POLICY TEMPORAL SELF-DISTILLATION: NPO AND AUTONPO

Image-text reasoning. Table 1 evaluates NPO and AutoNPO on eight image-text reasoning benchmarks. Pure on-policy GRPO reaches an average of 60.25. Applying NPO only in the early stage raises this to 62.12, showing that near-future transfer is already useful when correct on-policy trajectories are scarce. Adding a late-stage intervention further improves the average to 62.84, while AutoNPO reaches 63.15 without manually selected intervention points. These results match the two intended roles of NPO: near-future guidance accelerates early learning and remains useful when the policy later approaches a plateau.

Table 2: **Text (top) and video (bottom) reasoning.** Text methods continue from the same LUFFY policy. **Bold** marks the better score; *Avg.* is computed within each domain. Full comparisons are in Appendix A.2.

Method	AIME 2024	AIME 2025	AMC	MATH-500	Minerva Math	OlympiadBench	Avg.
GRPO	24.8	17.8	**67.5**	88.4	38.6	55.3	48.7
NPO	**34.6**	**25.7**	66.9	**89.2**	**40.4**	**58.2**	**52.5**
Method	**TempCompass**	**MotionBench**	**MLVU-Dev**	**VideoMathQA**	**Charades-STA**	**VideoMME-v2**	**Avg.**
GRPO	75.98	57.01	75.80	50.33	57.65	33.03	58.30
NPO	**76.47**	**59.17**	**76.40**	**52.71**	**58.06**	**33.81**	**59.44**

Text-only and video reasoning. We next test whether the same off-policy temporal transfer generalizes beyond image-text reasoning (Table 2). For text reasoning, we follow the experimental settings of LUFFY (Yan et al., 2025) and ExGRPO (Zhan et al., 2025) to ensure comparability. Starting from the same LUFFY policy, NPO reaches an average of 52.5, compared with 48.7 for on-policy GRPO, and also exceeds continual LUFFY and ExGRPO in the full comparison reported in Appendix A.2. For video reasoning, NPO improves over GRPO on all six benchmarks, raising the average from 58.30 to 59.44. The gains across text, image-text, and video reasoning indicate that the benefit is not specific to a single modality or benchmark family.

4.3 On-Policy Temporal Self-Distillation: NPD

We next ask whether the near-future advantage depends on off-policy trajectory transfer. To isolate this question, we study an on-policy realization of temporal self-distillation, which we call **Near-Future Policy Distillation (NPD)**. NPD uses the standard on-policy distillation objective but selects a near-future checkpoint as the teacher. Unlike NPO, where the future policy generates the transferred trajectory, NPD evaluates the teacher on student-generated responses and transfers token-level distributions on the learner's own rollout distribution.



Figure 4: **Continued RL after NPD.** Thin lines show raw evaluations; thick lines show trailing means over five evaluations.

We compare a near-future teacher with a far-future teacher while keeping the student initialization, distillation objective, and subsequent GRPO stage fixed. Both teachers are frozen checkpoints from the same GRPO training process, and each distills its behavior into the original Qwen3-VL-8B-Instruct student. We then select the best checkpoint from each distillation run and continue both students using the same GRPO configuration. Implementation details are given in Appendix A.4.

The near-future teacher leads to stronger downstream learning. After subsequent RL, the near-guided student reaches an eight-benchmark average of 63.23, compared with 61.92 for the far-guided student, and performs better on seven of the eight benchmarks (Table 5). Interestingly, the near-guided student starts the subsequent RL stage from a lower validation score, yet improves substantially more during continued training (Figure 4).

Because NPD applies teacher supervision on the learner's own rollout distribution, this advantage cannot be attributed solely to reduced off-policy trajectory shift. More importantly, the result shows that higher immediate post-distillation performance does not necessarily imply better learnability under subsequent RL: the near-future teacher produces a student that starts weaker, but learns more effectively afterward.

4.4 Analysis and Discussion

Source-level evidence for the quality–cost view. Figure 5 compares several guidance sources in terms of verified success and source–learner mismatch. Current-policy and historical-replay trajectories remain close to the learner but add little new success, whereas the externally guided

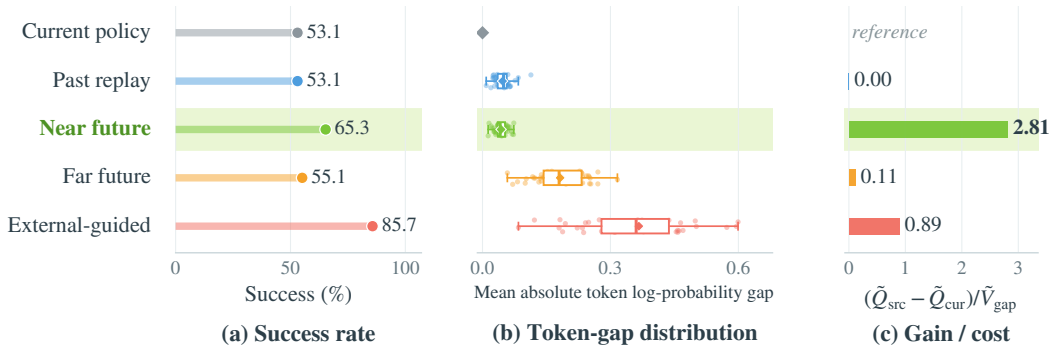


Figure 5: **Source-level evidence for the quality–cost trade-off.** **(a)** Verified success $\tilde{Q}$ across guidance sources. **(b)** Source–learner mismatch measured by mean absolute token log-probability gap $\tilde{V}_{\text{gap}}$. **(c)** Incremental verified success per unit mismatch cost. Near-future guidance provides the largest measured usable-gain ratio among the compared sources.

policy provides substantially higher raw success at a much larger token-level discrepancy. The near-future source occupies an intermediate regime: its verified success rate rises from 53.1% for the current policy to 65.3%, while its token-level gap remains close to historical replay. By contrast, the longer-distance future source incurs greater discrepancy without a commensurate increase in success. The resulting incremental gain-per-cost proxy is highest for the near-future source (2.81), compared with 0.89 for the externally guided policy and 0.11 for the longer-distance future source. This source-level evidence is consistent with the quality–cost hypothesis in Section 2: the most useful guide need not be either the closest or the strongest one.

The same qualitative pattern is observed under two different mismatch proxies. Figure 2 uses a KL-based cost proxy along the checkpoint sequence, whereas Figure 5 measures absolute token log-probability gaps across source types. Together, these analyses support treating source–learner mismatch as a general guidance cost rather than tying it to a single metric.

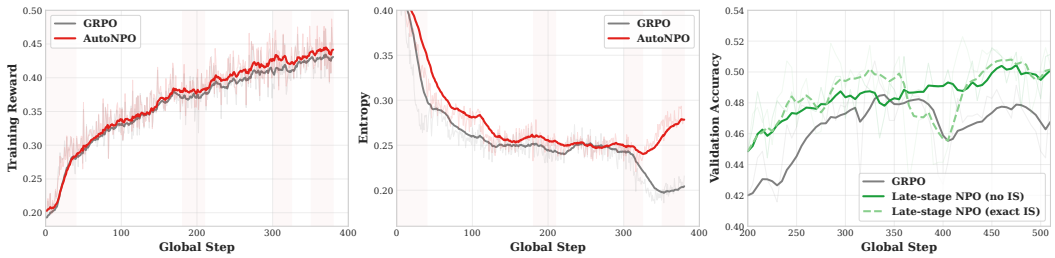


Figure 6: Training dynamics of (Auto)NPO compared with vanilla GRPO. **Red shaded regions** in panels (a) and (b) highlight AutoNPO's intervention windows. **(a)** Training reward dynamics: AutoNPO stays above GRPO throughout, with the gap widening step-wise after each intervention window. **(b)** Generation entropy: GRPO undergoes a steady entropy collapse, while AutoNPO's interventions actively re-expand exploration and preserve a substantially higher-entropy policy. **(c)** IS-correction ablation on the late-stage intervention: the two NPO variants with and without exact importance-sampling correction both clearly outperform GRPO, indicating that IS correction is marginal for NPO. Thick lines are smoothed curves; thin lines show raw logged values.

Training dynamics. Figure 6 shows how AutoNPO changes the training process through localized retrospective interventions. AutoNPO improves more rapidly in the early stage and remains above GRPO throughout training; later intervention windows produce additional step-wise improvements rather than a uniform shift across the entire run. At the same time, GRPO's generation entropy steadily declines, consistent with a narrowing rollout distribution, whereas AutoNPO maintains higher entropy and re-expands after intervention windows. This behavior suggests that retrospective

transfer does more than replay successful answers: it can reopen useful regions of the learner's rollout distribution for subsequent on-policy exploration.

The late-stage comparison in Figure 6 also shows that the NPO gain is largely preserved when the source-policy importance correction is omitted. We therefore retain the correction in the formal off-policy formulation, but do not treat it as a critical ingredient of near-future transfer.

Ablations. The ablations in Appendix A.3 (Table 4) test the main design choices behind adaptive temporal transfer. Restricting the rollback search range or fixing a single temporal distance reduces performance relative to AutoNPO, consistent with the learner-dependent teacher selection motivated in Section 2. Guidance gating is also important: overly restrictive gating misses useful transfer opportunities, whereas overly permissive gating replaces too much on-policy experience. Finally, omitting the source-policy importance correction changes performance only marginally for the near-future guides considered here.

Cost analysis. To separate algorithmic gains from simply spending more rollout compute, we compare AutoNPO and GRPO under the same 500 GPU-hour budget, including AutoNPO's rollback and guidance costs. AutoNPO reaches 53.43% best validation accuracy, compared with 50.71% for GRPO with $n = 8$ and 51.14% for GRPO with twice as many rollouts ($n = 16$). Their corresponding eight-benchmark averages are 63.15, 60.25, and 61.14. Thus, the AutoNPO advantage remains even when the on-policy baseline is given substantially more rollout capacity. Full benchmark results are reported in Appendix A.5.

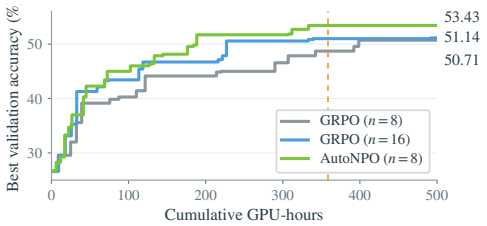


Figure 7: Best validation accuracy within 500 GPU-hours. The dashed line marks AutoNPO's first intervention.

5 Related Work

RLVR and trajectory-source methods. GSPO and SAPO refine the on-policy RLVR update (Zheng et al., 2025; Gao et al., 2025). LUFFY adds external teacher traces (Yan et al., 2025), while ExGRPO and RLEP reuse historical or stronger-policy experience (Zhan et al., 2025; Zhang et al., 2025). NPO selects a nearby later checkpoint to balance additional capability against source–learner mismatch.

Self-distillation and self-guided learning. ReST and STaR learn from successful model generations (Gulcehre et al., 2023; Zelikman et al., 2022). Context distillation uses privileged context (Snell et al., 2022), while on-policy distillation supplies teacher predictions on student-generated responses (Agarwal et al., 2024; Lu & Lab, 2025). OPSD and SDPO condition these predictions on reference solutions or feedback (Zhao et al., 2026; Hübotter et al., 2026). Temporal self-distillation asks *which later checkpoint should teach an earlier learner*. We study this choice through NPO's off-policy trajectories and NPD's on-policy token targets. Further comparisons appear in Appendix B.

6 Conclusion and Future Work

We introduce **temporal self-distillation**: a policy can learn from a stronger later version of itself, provided the teacher is sufficiently capable yet still close enough to be readily learned from. We capture this trade-off through $\mathcal{S} = Q/V$ and instantiate it through two complementary mechanisms: **NPO** for off-policy behavioral transfer and **NPD** for on-policy distributional transfer. Across image-text, text-only, and video reasoning, near-future guidance consistently improves learning, and its advantage persists across both transfer settings.

More broadly, **AutoNPO** turns temporal self-distillation into an adaptive self-bootstrap process by deciding when to intervene and how far back to revisit. A natural next step is to let sufficiently capable models reason over their own learning histories and decide *when, where, and how* to improve themselves, moving toward more self-directed forms of recursive self-improvement.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, analysis of reviewer feedback, planning of follow-up experiments, and LaTeX formatting. The authors are responsible for the scientific claims, experimental results, and final content of the paper.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

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.

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.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning, 2025. URL https://arxiv.org/abs/2506.19767.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query, 2017. URL https://arxiv.org/abs/1705.02101.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

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.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling, 2025. URL https://arxiv.org/abs/2507.01679.

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

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*, 2024.

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

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.

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, 2024a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024b. URL https://arxiv.org/abs/2403.00476.

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.

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*, 2024.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, 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.

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.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL https://arxiv.org/abs/2601.19897.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. Trust-region adaptive policy optimization, 2025. URL https://arxiv.org/abs/2512.17636.

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.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026. URL https://arxiv.org/abs/2604.03128.

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, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025b.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

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*, 2025.

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.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

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

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 Additional Experimental Results

A.1 Experimental Details

Training data. We use the MMFineReason-123K data preparation of Yang et al. (2026), derived from the MMFineReason-1.8M corpus (Lin et al., 2026) via difficulty-based filtering. Each training 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 kept. This criterion removes trivial examples and concentrates training on problems where improvement is still possible, matching the regime where NPO is designed to help.

Image-text benchmarks. For evaluation, we use eight multimodal reasoning benchmarks that cover visually grounded mathematics, broad subject knowledge, and fine-grained visual understanding. MathVista (Lu et al., 2024), 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., 2024a) 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.

Models and baselines. All image-text experiments use Qwen3-VL-8B-Instruct (Bai et al., 2025) as the base policy. NPO and the trajectory-based RL baselines use a GRPO-style group-based RLVR backbone. GRPO (Shao et al., 2024) is the pure on-policy reference. LUFFY (Yan et al., 2025) imports trajectories from a stronger external teacher, representing the high-quality but high-shift regime. ExGRPO (Zhan et al., 2025) reuses successful trajectories from historical replay buffers, and RLEP (Zhang et al., 2025) replays trajectories from a far-future model. We also report the Base LLM without post-training as a reference point for overall RL gains. The distillation baselines are OPSD (Zhao et al., 2026), SDPO (Hübotter et al., 2026), and OPD (Agarwal et al., 2024; Lu & Lab, 2025); their implementations are described below.

Text and video extensions. For text-only reasoning, we follow the matched continuation setting of prior mixed-policy RLVR work (Yan et al., 2025; Zhan et al., 2025). All methods continue from the same LUFFY policy; the training data, verifier, and evaluation protocol are held fixed. The six text-only benchmarks are AIME 2024, AIME 2025, AMC (Li et al., 2024), MATH-500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). The six video benchmarks are TempCompass (Liu et al., 2024b), MotionBench (Hong et al., 2025), MLVU-Dev (Zhou et al., 2025), VideoMathQA (Rasheed et al., 2025), Charades-STA (Gao et al., 2017), and VideoMME-v2 (Fu et al., 2026). These benchmarks cover temporal comparison, motion understanding, long-video QA, video-grounded mathematics, temporal localization, and general video QA.

GRPO-based implementation details. Our NPO implementation is built on the EasyVideoR1 (Qin et al., 2026) framework. The default GRPO-based configuration uses an 8192-token context, split into a prompt budget of 4096 and a response budget of 4096, at training and evaluation time. The default learning rate is 1×10^{-6} with a batch size of 256, and each prompt is sampled $n = 8$ times at temperature 1.0. Clipping thresholds are set to $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$, and we omit both the KL penalty and entropy regularization from the objective. For all NPO variants, the guidance gate in Section 3.1 uses $\tau_{\text{gate}} = 0.6$, implemented as `mix_policy_accuracy_threshold`; a future-self trajectory is injected only when $\hat{p}(x) \leq \tau_{\text{gate}}$, leaving confident groups untouched. The GRPO-based experiments use 4 compute nodes, each with 8 NVIDIA H200 140GB GPUs.

Distillation baselines. The three distillation baselines learn from token-level teacher distributions evaluated along student-generated responses. We use Jensen–Shannon divergence over the teacher's top-256 tokens, with the remaining probability mass collected into a tail category. Teacher targets are detached from the gradient, and the update contains no additional outcome-reward PPO loss. For OPSD and SDPO, we follow the experimental setting of Yang et al. (2026): MMFineReason-123K, batch size 256, a rollout group size of eight, a learning rate of 1×10^{-5}, and maximum prompt and

Table 3: **Full text and video reasoning results.** (a) Text continuation methods start from the same LUFFY policy; GRPO is the on-policy continuation baseline. (b) Video results cover six benchmarks; gain is the difference between NPO and GRPO. Averages are computed within each panel.

(a) Text reasoning

Method	AIME 2024	AIME 2025	AMC	MATH-500	Minerva Math	OlympiadBench	Avg.
LUFFY	29.4	23.1	65.6	87.6	37.5	57.2	50.1
↪ Continual LUFFY	30.7	22.5	66.2	86.8	**41.2**	55.3	50.4
↪ GRPO	24.8	17.8	**67.5**	88.4	38.6	55.3	48.7
↪ ExGRPO	32.3	**25.7**	65.6	87.6	40.1	57.0	51.4
↪ **NPO**	**34.6**	**25.7**	66.9	**89.2**	40.4	**58.2**	**52.5**

(b) Video reasoning

Benchmark	Base	GRPO	NPO	Gain
TempCompass	75.69	75.98	**76.47**	+0.49
MotionBench	56.14	57.01	**59.17**	+2.16
MLVU-Dev	73.32	75.80	**76.40**	+0.60
VideoMathQA	47.81	50.33	**52.71**	+2.38
Charades-STA	56.00	57.65	**58.06**	+0.41
VideoMME-v2	32.50	33.03	**33.81**	+0.78
Avg.	56.91	58.30	**59.44**	**+1.14**

response lengths of 4,096 tokens each. The sampling temperature is 1.0 for all three baselines; the external-teacher OPD baseline uses a learning rate of 1×10^{-6}.

OPSD. Following the multimodal reference-context setup of Yang et al. (2026), the self-teacher receives the question, image, and a reference solution from the MMFineReason data. These reference traces are generated by Qwen3-VL-235B-A22B-Thinking and verified as correct. Our implementation uses a frozen copy of the initial 8B model as the teacher and supplies the reference trace directly, without a separate hint-generation stage. When a reference trace is unavailable, the implementation falls back to the final ground-truth answer. The student generates its response from the original question and image, without the reference context.

SDPO. The current student also serves as the self-teacher, conditioned on feedback from its own sampled attempts. For each prompt, the implementation samples eight candidate responses, verifies them, and selects the shortest correct and shortest incorrect response when available as teacher-side context. A separate student rollout provides the response used for distillation. Missing feedback components are left empty. The teacher follows the current student parameters, with stop-gradient applied during teacher scoring; dataset reference traces are not supplied to this teacher.

OPD. A frozen Qwen3-VL-32B-Instruct model supervises the 8B student on the same image, question, and student-generated prefixes. The teacher receives no additional reference solution. The configuration uses a batch size of 256, one student rollout per prompt, and maximum prompt and response lengths of 16,384 and 4,096 tokens, respectively. Table 1 reports the distilled student directly, without subsequent RL training.

A.2 Full Text and Video Results

Table 3 retains the complete continuation comparison for text reasoning and the base-model reference for video reasoning. In the text experiment, GRPO denotes on-policy continuation from the LUFFY starting policy. AIME is the American Invitational Mathematics Examination, and AMC is the American Mathematics Competitions; Minerva Math and OlympiadBench are written out in Table 2.

A.3 Ablation Discussion

Table 4 separates four design choices in AutoNPO. The manual interventions isolate stage effects: early guidance improves the average from 60.25 to 62.12 and is strongest on ZeroBench, where correct on-policy samples are scarce, while late-stage guidance reaches 62.84 and gives the largest

Table 4: Ablation study on AutoNPO design choices. *Avg.* is the unweighted mean over the eight multimodal reasoning benchmarks.

Method	MMMU-Pro	MathVista	MathVision	ZeroBench	WeMath	MMBench	MM-Star	MathVerse	Avg.
GRPO	55.78	76.20	48.82	22.60	56.57	90.29	72.20	59.52	60.25
NPO, early-stage only	56.85	76.60	54.31	**26.35**	62.76	90.41	70.30	59.38	62.12
NPO, early + late-stage	57.07	76.30	54.61	24.85	**66.95**	90.30	72.67	60.00	62.84
AutoNPO	**57.24**	**79.20**	**55.72**	24.70	66.00	90.63	72.63	59.11	**63.15**
Search range $\Delta_{\max} = 20$	57.09	77.90	53.29	23.95	67.24	90.60	72.53	59.90	62.81
Search range $\Delta_{\max} = 80$	55.98	76.60	52.58	23.05	65.62	90.83	72.07	59.26	62.00
Fixed $\Delta = 40$	56.97	76.85	55.39	25.60	63.24	**90.89**	72.77	**60.15**	62.73
Gate $\tau_{\text{gate}} = 0.3$	55.38	77.75	52.99	23.50	60.86	90.45	72.53	58.87	61.54
Gate $\tau_{\text{gate}} = 0.8$	56.38	76.75	54.62	24.10	64.19	90.08	**73.10**	56.92	62.02
Late-stage NPO w/o exact IS	56.50	76.60	55.03	24.40	66.10	90.30	71.90	59.99	62.60

Table 5: **Benchmark results for the distillation extension.** All students start from Qwen3-VL-8B-Instruct. The step-40 and step-520 teachers are frozen 8B GRPO checkpoints. The 32B-teacher row reports pure OPD without subsequent RL and is included as a reference. **Bold** marks the best result per column, and *Avg.* is the unweighted mean across eight benchmarks.

Method	MMMU-Pro	MathVista	MathVision	ZeroBench	WeMath	MMBench	MM-Star	MathVerse	Avg.
Qwen3-VL-8B-Instruct	51.75	73.80	47.37	19.76	54.10	89.79	71.83	54.61	57.88
GRPO	55.78	76.20	48.82	22.60	56.57	90.29	72.20	59.52	60.25
OPD (32B teacher)	53.42	**76.95**	48.57	21.71	58.19	90.15	72.00	53.67	59.33
OPD (far, step 520) + GRPO	56.50	75.40	54.84	**25.15**	62.57	90.49	71.67	58.74	61.92
OPD (near, step 40) + GRPO	**58.57**	76.15	**56.64**	24.70	**65.24**	**90.96**	**72.77**	**60.78**	**63.23**

gains on WeMath, where the vanilla run plateaus. AutoNPO combines these effects without manual scheduling and reaches 63.15.

The controller ablations show that the gain is not explained by simply injecting more off-policy trajectories. Changing the rollback search range from $\Delta_{\max} = 20$ to 80 weakens performance at longer distances, consistent with the near-future trade-off. The gate sweep changes the default $\tau_{\text{gate}} = 0.6$ to 0.3 or 0.8: the lower gate misses many learnable prompts, while the higher gate replaces too many on-policy samples. Exact importance-sampling correction is also empirically marginal for near-future guidance: late-stage NPO reaches 62.84 with exact IS and 62.60 without it, and the training curves in Figure 6(c) remain close. Dropping IS saves memory and compute because guidance log-probabilities no longer need to be stored or recomputed; this simplification should not be read as a general claim for far-off-policy sources.

A.4 Distillation Extension

OPD and GRPO settings. Both students start from Qwen3-VL-8B-Instruct. The frozen teachers are the step-40 and step-520 checkpoints of the vanilla GRPO run. OPD uses MMFineReason data, a batch size of 256, one student rollout per prompt, temperature 1.0, and learning rate 3×10^{-7}. The loss is token-level Jensen–Shannon divergence over the teacher's top-256 tokens, with the remaining probability mass grouped into a tail category; no outcome-reward PPO loss is applied during OPD. The best checkpoints from the near- and far-teacher OPD runs are selected for subsequent RL, at OPD steps 95 and 165, respectively. Both then use the same GRPO data and configuration: batch size 256, eight rollouts per prompt, temperature 1.0, learning rate 10^{-6}, clipping thresholds 0.2 and 0.28, and a maximum of 500 updates. The data order is fixed, and both runs use seed 1. Both stages allow 16,384 prompt tokens and 4,096 response tokens. The OPD durations differ; the matched configuration applies to the subsequent GRPO stage.

Validation curves. Figure 4 reports accuracy on 700 validation prompts, evaluated every five updates with one response per prompt, temperature 0.7, top-p 0.8, top-k 20, and presence penalty 1.5. Before RL, the near- and far-guided students score 39.14% and 45.14%; their observed peaks are 52.14% at step 380 and 51.43% at step 435, giving improvements of 13.00 and 6.29 percentage points. These validation scores are separate from the benchmark averages in Table 5. The figure

Table 6: Benchmark results for the compute comparison (%). Bold marks the best result; Avg. is the mean over eight benchmarks.

Method	MMMU-Pro	MathVista	MathVision	ZeroBench	WeMath	MMBench	MM-Star	MathVerse	Avg.
GRPO ($n = 8$)	55.78	76.20	48.82	22.60	56.57	90.29	72.20	**59.52**	60.25
GRPO ($n = 16$)	55.67	74.75	53.78	23.50	62.19	90.16	69.57	59.50	61.14
AutoNPO	**57.24**	**79.20**	**55.72**	**24.70**	**66.00**	**90.63**	**72.63**	59.11	**63.15**

shows steps 0–400 with trailing means over up to five evaluations; peaks use all available evaluations, through steps 500 and 490 for the near and far runs, respectively.

A.5 Compute Comparison

A.6 Figure Details

Figure 2. Panel (a) is a conceptual illustration of the quality–cost trade-off. Dots indicate source quality and policy proximity, shaded disks represent guidance cost, and triangles illustrate the useful signal after accounting for that cost. Their positions and sizes are schematic rather than measured comparisons between methods.

Panels (b,c) use checkpoints from a vanilla GRPO run with Qwen3-VL-8B-Instruct. For an anchor at step T, the quality proxy $\widetilde{Q}(\Delta)$ is the positive validation-score gain over that anchor, using the logged `val.reward_score`. The validation sequence is smoothed with an eight-observation window, and the gains with a three-observation window. Panel (b) then takes a cumulative maximum to display the gain envelope. We use validation improvement here as a proxy for additional capability.

The cost proxy is $\widetilde{V}(\Delta) = \exp((\Delta))$, where $d_T(\Delta)$ is the empirical token KL estimated on trajectories generated by the anchor and scored by later checkpoints. The exponential transform sets the scale of the heuristic cost; it is not a direct measurement of gradient variance. Panel (b) uses anchor $T = 0$. Panel (c) plots $\widetilde{Q}/\widetilde{V}$ for anchors $T = 0$ and $T = 50$, using the smoothed gains without the cumulative-maximum operation. The marked points are the highest scores among the sampled distances, at 20 and 70 steps, respectively.

B Additional Related Work

RLVR and trajectory sources. Beyond the methods discussed in the main text, Prefix-RFT supplies expert prefixes (Huang et al., 2025), while ReLIFT, SRFT, and TRAPO combine RL with supervised updates (Ma et al., 2025; Fu et al., 2025; Su et al., 2025). RePO and ExGRPO reuse successful historical trajectories (Li et al., 2025; Zhan et al., 2025), and RLEP draws experience from a stronger seed policy (Zhang et al., 2025). These methods motivate enriching the learner's available signals. Our focus is the temporal choice of source: how much additional optimization should a teacher receive before guiding an earlier learner?

Self-distillation and temporal teacher selection. ReST and STaR bootstrap from successful generations (Gulcehre et al., 2023; Zelikman et al., 2022); context distillation uses privileged context (Snell et al., 2022). On-policy distillation evaluates teacher predictions on student-generated responses (Agarwal et al., 2024; Lu & Lab, 2025), with OPSD and SDPO using reference solutions or feedback to inform those predictions (Zhao et al., 2026; Hübotter et al., 2026). Self-distillation also supports continual learning (Shenfeld et al., 2026). Temporal self-distillation instead obtains its teacher through further optimization along the learner's training path. NPO and NPD examine this teacher choice through trajectory-level and token-level transfer, respectively.

C AutoNPO Implementation and Pseudocode

At step t, the current policy $\pi^{(t)}$ can serve as a future guide for an earlier checkpoint $\pi^{(t-\Delta)}$. The controller first confirms that it has acquired useful new capability, then selects a rollback distance

using $\hat{\mathcal{S}}(\Delta) = \hat{Q}(\Delta)/\hat{V}(\Delta)$. The details below complement the overview in Section 3.3; Algorithm 1 gives the complete procedure.

Mistake pool. AutoNPO maintains a lightweight mistake pool $\mathcal{B}$ containing prompts with low recent group accuracy together with the steps at which they were failed. The pool serves both as the probe set for detecting newly acquired capability and as the evaluation set for comparing candidate rollback distances. After an intervention is selected, the relevant subset also defines the prompts on which future-self guidance is prepared.

Trigger. The controller uses a coarse-to-fine two-stage check. A zero-cost warning stage monitors training reward and policy entropy, and activates when reward stagnation coincides with continued entropy decline, a practical indicator of reduced exploration. If this pattern persists for several probe intervals, a confirmation stage rolls out the current policy once on a small subset of $\mathcal{B}$. The resulting pass-rate determines whether the current policy has acquired enough new capability to serve as a useful guide; the same rollout also supplies the estimates of $\hat{Q}(\Delta)$ used below.

Rollback distance. Once an intervention is confirmed, AutoNPO selects

$$\Delta^\star = \arg\max_{\Delta \in \mathcal{D}} \frac{\hat{Q}(\Delta)}{\hat{V}(\Delta)}, \tag{9}$$

where $\mathcal{D}$ is the set of available checkpoint distances. For each candidate Δ, let $\mathcal{B}_\Delta$ denote the subset of mistake-pool prompts that failed during the segment beginning at $t - \Delta$. We estimate

$$\hat{Q}(\Delta) = \text{pass-rate}\big(\pi^{(t)}; \mathcal{B}_\Delta\big), \tag{10}$$

and use an exponential proxy based on the per-token KL between $\pi^{(t)}$ and $\pi^{(t-\Delta)}$ for $\hat{V}(\Delta)$. Because all candidate $\hat{Q}(\Delta)$ values are obtained from the same confirmation rollout, checkpoint ranking requires no additional inference beyond the probe itself.

Execution. After choosing $\Delta^\star$, the current policy $\pi^{(t)}$ acts as the future guide for the earlier checkpoint $\pi^{(t-\Delta^\star)}$. We roll out $\pi^{(t)}$ on prompts in $\mathcal{B}_{\Delta^\star}$, verify the outputs, and cache one correct trajectory per prompt when available. Training then restores checkpoint $t - \Delta^\star$ and replays the intervened segment using the NPO mechanism. During replay, cached guidance is injected only for prompts in $\mathcal{B}_{\Delta^\star}$; all other prompts retain fully on-policy rollout groups.

Once the replayed run catches back up to step t, the controller enters a short cooldown and then re-arms. The resulting policy can therefore later serve as a stronger future guide for another intervention, allowing temporal self-transfer to repeat as training progresses. Algorithm 1 gives the full procedure.

D Broader Impact

Temporal self-distillation obtains guidance from the model's own later checkpoints, reducing dependence on stronger external teachers for post-training. Improved reasoning can support useful applications, but can also enable more effective misuse, including misleading content. Verifier-based training may reinforce harmful behavior if the verifier rewards it. Deployment therefore requires safety evaluation and safeguards appropriate to the application, alongside the reasoning improvements studied here.

Algorithm 1 AutoNPO: adaptive intervention for near-future policy optimization

Require: dataset $\mathcal{D}$, base policy $\pi^{(0)}$, mistake threshold τ_{err}, confirmation pass-rate threshold τ_{pass}, probe size N_{probe}, probe interval T_{probe}, cooldown length T_{cool}, cost proxy $\hat{V}(\cdot)$

$\mathcal{B} \leftarrow \emptyset$, `alert` $\leftarrow 0$, `in_retro` $\leftarrow$ False, $\mathcal{G} \leftarrow \emptyset$, $t_{\text{cool}} \leftarrow 0$
for $t = 1, 2, \ldots$ **do**
  sample batch $\{x\}$ from $\mathcal{D}$
  if `in_retro` **then for** each x with $\mathcal{G}[x]$ defined, schedule $\mathcal{G}[x]$ to replace one rollout slot
  end if
  roll out with $\pi^{(t)}$, compute rewards and group-relative advantages
  update $\pi^{(t+1)}$ via the clipped objective in Eq. (5)
  update reward EMA and entropy history
  for each prompt x in the batch **do**
    if mean group accuracy of $x < \tau_{\text{err}}$ **then** record prompt x in $\mathcal{B}$ with timestamp t
    end if
  end for
  if not `in_retro` and $t \geq t_{\text{cool}}$ and $t \bmod T_{\text{probe}} = 0$ **then**
    // Stage 1: zero-cost warning
    `alert` $\leftarrow$ `alert` $+ 1$ if reward stagnates and entropy declines, else `alert` $\leftarrow 0$
    if `alert` $\geq m$ **then**
      // Stage 2: confirmation rollout
      sample N_{probe} prompts from $\mathcal{B}$; roll out with $\pi^{(t)}$
      $\hat{p} \leftarrow$ pass-rate on the probe sample
      for each saved checkpoint at distance Δ: $\hat{Q}(\Delta) \leftarrow$ pass-rate of $\pi^{(t)}$ on $\mathcal{B}_\Delta$
      if $\hat{p} > \tau_{\text{pass}}$ **then**
        $\Delta^\star \leftarrow \arg\max_\Delta \hat{Q}(\Delta)/\hat{V}(\Delta)$ ▷ Sec. 3.3
        $\mathcal{G} \leftarrow$ verified-correct re-rollout of $\pi^{(t)}$ on $\mathcal{B}_{\Delta^\star}$
        load model weights and dataloader state from checkpoint $t - \Delta^\star$
        $t_{\text{resume}} \leftarrow t$; `in_retro` $\leftarrow$ True
      end if
      `alert` $\leftarrow 0$
    end if
  end if
  if `in_retro` and current step $\geq t_{\text{resume}}$ **then**
    `in_retro` $\leftarrow$ False; $\mathcal{G} \leftarrow \emptyset$; $t_{\text{cool}} \leftarrow t + T_{\text{cool}}$
  end if
end for