AsyncOPD: How Stale Can On-Policy Distillation Be?

Wonjun Kang[*1] **Kevin Galim**[*1] **Seunghyuk Oh**[1] **Minjun Kang**[2] **Sanghyun Park**[2]
Donghoon Kim[1] **Minjae Lee**[1] **Minseo Kim**[1,3] **Rishabh Tiwari**[3] **Yuchen Zeng**[4]
Hyung Il Koo[†1,2] **Kangwook Lee**[†5,6]

[1] FuriosaAI [2] Ajou University [3] UC Berkeley [4] Microsoft Research
[5] KRAFTON [6] Ludo Robotics

Code: `https://github.com/furiosa-ai/async-opd`

Abstract

On-policy distillation (OPD) trains a student on its own rollouts guided by teacher feedback and is becoming increasingly important for large language model (LLM) post-training. Like reinforcement learning (RL), however, OPD faces an on-policy systems bottleneck, as rollouts can dominate training time for reasoning workloads. Asynchronous training pipelines can alleviate this bottleneck by decoupling rollout generation from learner updates, but doing so introduces stale-policy data. While prior work has studied stale data in asynchronous RL, its effects in OPD remain underexplored. We present the first systematic study of staleness in asynchronous OPD, focusing on a practical setting where teacher feedback is implemented through local KL losses and full-vocabulary teacher logits are too expensive to store or transfer, necessitating finite teacher-score caches. We first show that KL direction changes the stale-data problem: teacher-weighted forward KL is more robust to stale rollouts, whereas student-weighted reverse KL is vulnerable. Second, for this vulnerable reverse-KL case, we study whether methods designed to stabilize asynchronous RL can mitigate OPD staleness. In our experiments, they do not improve over a simpler OPD-specific surrogate: recomputing the reverse-KL signal under the current student at learner time. Third, we analyze how finite teacher-score caches create a bias-variance tradeoff for sparse and sampled reverse-KL OPD estimators. This motivates multi-sample Monte Carlo (MC), which preserves MC correctability while reducing one-sample variance. Finally, we present and open-source **AsyncOPD**, a fully asynchronous OPD training pipeline built from these estimator choices. Experiments show that AsyncOPD improves training throughput by $1.6\times$ to $3.8\times$ over strict synchronous training while reaching comparable accuracy.

1 Introduction

On-policy distillation (OPD) [1, 7, 22] and reinforcement learning (RL) [28, 30] have become central post-training methods for improving large language model (LLM) reasoning [8], including mathematics [19] and coding [29]. OPD trains a student on its own rollouts using dense token-level feedback from a teacher [14], whereas RL learns from reward feedback on rollouts. OPD provides an effective and efficient route for LLM post-training, especially for smaller student models [26]. Recent work shows that OPD is not limited to distilling large teachers into small students: it also supports on-policy self-distillation [2, 34] and multi-teacher distillation from domain-specialized teachers comparable in size to the student [3, 23].

[*]Equal Contribution. [†]Corresponding Authors. Email: `kangwj1995@furiosa.ai`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

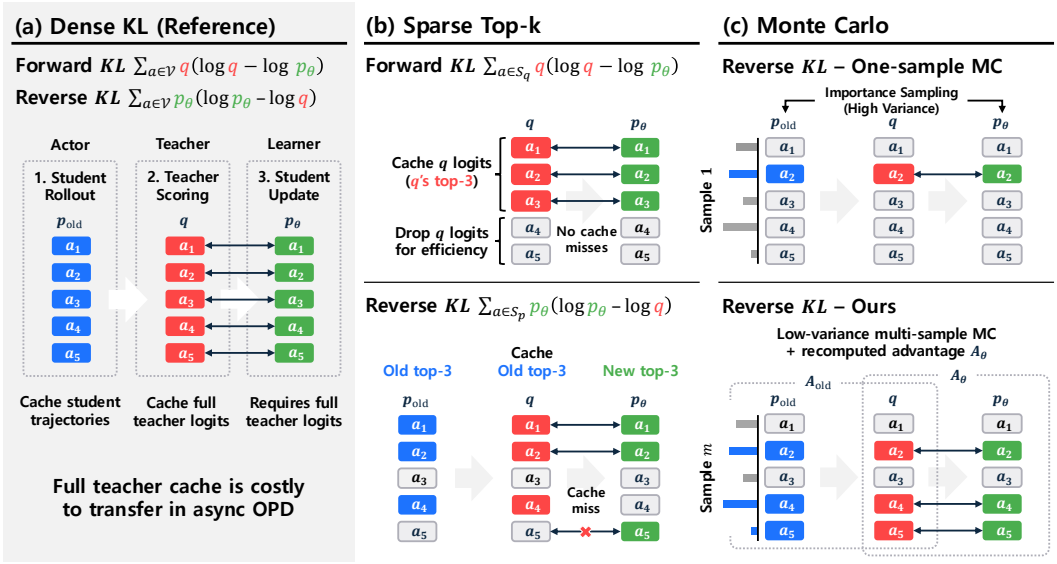


Figure 1: Estimator design for asynchronous OPD. (a) Dense KL is the full-vocabulary reference, but full teacher-logit caches are costly to store or transfer in asynchronous OPD. (b) Sparse top-k exposes a support mismatch under staleness: forward KL is teacher-supported, but reverse KL is student-supported and may require actions outside the cached teacher-scored support. (c) One-sample Monte Carlo is correctable in expectation by importance sampling, but has high variance; our estimator recomputes A_θ at learner time and uses multi-sample MC to reduce variance.

OPD and RL inherit an on-policy systems bottleneck: each learner update must wait for fresh rollouts from the model being trained [6, 10]. For reasoning tasks, these rollouts are long and expensive, so synchronous training often waits on generation rather than updating the model, leaving learners underutilized. Asynchronous RL [5, 16] relieves this bottleneck by decoupling rollout generation from learner updates: rollout workers keep generating data while the learner updates on earlier rollouts, improving training efficiency and hardware utilization [20, 25, 36]. A similar pipeline can be applied to OPD by running student rollout, teacher scoring, and learner updates in parallel [21].

However, asynchronous execution introduces stale-policy data, and learning from such data can degrade model quality [4]. This creates a trade-off: more aggressive asynchrony improves training throughput, but it also increases the policy lag between rollout and learning. Prior work on asynchronous RL therefore studies how to stabilize learning from stale-policy data [5, 12, 35]. However, it remains underexplored whether these ideas and stale-data solutions transfer to OPD, because practical implementations of OPD expose a different feedback interface. Teacher feedback is often implemented through local KL losses, which require teacher scores over actions at student-visited prefixes. Since full-vocabulary teacher logits are expensive to store or transfer, especially in an asynchronous pipeline, teacher scores are usually cached only on a finite set of actions (Fig. 1). Once the learner receives a teacher-scored cache, it can recompute current-student log probabilities on cached actions, but it cannot recover teacher scores for actions that were never scored. This raises three questions that structure our study: (i) how asynchronous OPD behaves under staleness, (ii) whether asynchronous RL ideas and stale-data solutions transfer to OPD, and (iii) how finite teacher-score caches shape OPD estimator design.

First, we study how KL direction shapes staleness. Under asynchronous OPD with cached teacher scores, the same stale rollout cache can affect different KL objectives differently. As illustrated in Fig. 1, forward KL is teacher-weighted and is more robust to stale rollouts, whereas reverse KL is student-weighted and becomes vulnerable when current-student actions fall outside the scored cache. We therefore focus on reverse-KL OPD in the remainder of the staleness analysis.

Second, focusing on the reverse-KL case, we ask whether methods designed to stabilize asynchronous RL can also mitigate OPD staleness. This comparison is natural because reverse KL in OPD admits an RL-style policy-gradient surrogate, where the teacher-student log-ratio acts as a token-level advantage. We therefore evaluate PPO-style clipping [18], decoupled PPO [5], and M2PO [35].

In our experiments, they do not improve over a simpler OPD-specific surrogate: recomputing the reverse-KL token-level advantage under the current student at learner time without clipping.

Third, we return to the teacher-cache constraint and study the resulting bias-variance tradeoff for sparse and sampled reverse-KL OPD implementations. Stale student top-k supports provide deterministic coverage but are support-mismatched because they may omit actions required by the current top-k objective, and reweighting inside the stale support cannot recover the missing teacher scores. One-sample Monte Carlo (MC) avoids this fixed-support mismatch through importance-correctable samples from the stale rollout policy, but suffers from high variance. This motivates multi-sample MC, which caches and teacher-scores multiple stale-policy samples at each decoding step, preserving MC correctability while reducing one-sample variance.

Finally, we instantiate these findings in **AsyncOPD**, a fully asynchronous OPD pipeline that overlaps student rollout, teacher scoring, and learner updates. On Qwen3-Base models, AsyncOPD improves training throughput by $1.6\times$ to $3.8\times$ over strict synchronous training while maintaining comparable accuracy. Our contributions are:

- We provide the first systematic study of staleness in asynchronous OPD through the lens of an OPD-specific teacher-cache constraint.
- We show that KL direction changes the stale-data problem: forward KL is comparatively robust to stale rollouts, whereas reverse KL is vulnerable because it is student-weighted (Section 4).
- We identify that the most effective reverse-KL policy-gradient surrogate uses the advantage recomputed at learner time without clipping, and that advanced asynchronous RL surrogates do not improve over this choice (Section 5).
- We show that stale student top-k supports are support-mismatched, while one-sample MC remains correctable but high-variance; this motivates multi-sample MC (Section 6).
- We present and open-source **AsyncOPD**, a fully asynchronous OPD training pipeline, and demonstrate improved training efficiency while maintaining OPD quality (Section 7).

2 Related Works

On-Policy Distillation On-policy distillation (OPD) trains a student on its own rollouts while using a teacher to provide dense token-level feedback on the visited prefixes [14, 22]. GKD [1] introduced a token-level KL formulation, while MiniLLM [7] studied a sequence-level reverse-KL variant. Li et al. [13] study token-level OPD training dynamics and recipes for unstable configurations. TIP [24] characterizes per-token importance through student entropy and teacher-student divergence. G-OPD [27] interprets token-level OPD as dense KL-constrained RL and extends it with reward scaling. These works clarify OPD as an effective post-training objective, but assume rollouts, teacher scoring, and learner updates stay synchronized.

Asynchronous RL In synchronous RL pipelines, training often waits for the longest rollout in a batch to finish, leaving learner resources idle. Asynchronous RL improves hardware utilization by decoupling rollout generation from learner updates. Async RLHF [16] overlaps generation and learning so that new samples are produced while the learner trains on earlier ones. StreamRL [36] further disaggregates the RLHF pipeline into streaming stages. AReaL [5] fully decouples rollout workers from training workers for continuous asynchronous execution. Laminar [20] uses fine-grained weight synchronization for trajectory-level asynchrony. However, asynchronous RL must learn from stale-policy data. Decoupled PPO [5] stabilizes asynchronous RL training by separating the behavior policy for stale rollouts from the proximal policy that anchors PPO [18] updates. M2PO [35] stabilizes stale updates with second-moment importance-weight constraints, and A-3PO [12] reduces decoupled PPO overhead through staleness-aware interpolation.

Asynchronous OPD VeRL [21] implements step-off OPD schedulers that overlap student rollout, teacher scoring, and learner update by fixing rollout lag to one or two learner steps. These schedulers establish the practical feasibility of asynchronous OPD, but leave open how OPD estimators behave under stale teacher-scored caches. KDFlow [31] improves systems efficiency for LLM distillation by decoupling teacher inference from learner training and transmitting teacher hidden states, but targets synchronous OPD and leaves asynchronous execution as future work. We study this missing asynchronous OPD regime directly and build AsyncOPD from the resulting estimator choices.

3 Preliminaries: On-Policy Distillation

OPD setup At each decoding timestep, we view the visited prefix s as the local state and the next token a as the action. Let $q(a \mid s)$ denote the teacher policy and $p_\theta(a \mid s)$ denote the current student policy. Following prior work on token-level OPD [13, 14, 27], we apply local losses to generated output tokens and analyze the resulting objectives at a fixed prefix state s. OPD can be defined with different divergences; forward and reverse KL are two standard choices [1].

Forward-KL OPD At a fixed prefix s, forward-KL OPD is teacher-weighted:

$$D_F(\theta; s) = \mathrm{KL}(q(\cdot \mid s) \,\|\, p_\theta(\cdot \mid s)) = \textstyle\sum_{a \in \mathcal{V}} q(a \mid s) \left(\log q(a \mid s) - \log p_\theta(a \mid s)\right). \tag{1}$$

At a fixed prefix s, the gradient is $\nabla_\theta D_F(\theta; s) = -\sum_{a \in \mathcal{V}} q(a \mid s) \nabla_\theta \log p_\theta(a \mid s)$.

Reverse-KL OPD At the same prefix, reverse-KL OPD is student-weighted:

$$D_R(\theta; s) = \mathrm{KL}(p_\theta(\cdot \mid s) \,\|\, q(\cdot \mid s)) = -\textstyle\sum_{a \in \mathcal{V}} p_\theta(a \mid s) \left(\log q(a \mid s) - \log p_\theta(a \mid s)\right). \tag{2}$$

Differentiating and using $\mathbb{E}_{a \sim p_\theta(\cdot \mid s)}[\nabla_\theta \log p_\theta(a \mid s)] = \nabla_\theta \sum_a p_\theta(a \mid s) = 0$ gives

$$\begin{aligned} \nabla_\theta D_R(\theta; s) &= \textstyle\sum_{a \in \mathcal{V}} p_\theta(a \mid s) \left(\log p_\theta(a \mid s) - \log q(a \mid s) + 1\right) \nabla_\theta \log p_\theta(a \mid s) \\ &= -\textstyle\sum_{a \in \mathcal{V}} p_\theta(a \mid s) \left(\log q(a \mid s) - \log p_\theta(a \mid s)\right) \nabla_\theta \log p_\theta(a \mid s). \end{aligned} \tag{3}$$

Viewing Eq. (3) as a policy-gradient estimator and $A = \log q(a \mid s) - \log p(a \mid s)$ as the advantage term connects reverse-KL OPD to standard RL training machinery, and practical implementations typically use PPO-style surrogates. Given behavior-policy samples $a \sim p_{\mathrm{beh}}$, define $\rho_\theta(a, s) = p_\theta(a \mid s)/p_{\mathrm{beh}}(a \mid s)$ and $\bar{\rho}_\theta(a, s) = \mathrm{clip}(\rho_\theta(a, s), 1 - \epsilon, 1 + \epsilon)$. The PPO-style local surrogate uses these ratios with a frozen behavior-time signal $A_{\mathrm{beh}}(a, s)$, where $\mathrm{sg}(\cdot)$ denotes stop-gradient:

$$L_{\mathrm{PPO}}(\theta; A_{\mathrm{beh}}) = -\mathbb{E}_{a \sim p_{\mathrm{beh}}} \left[\min\left(\rho_\theta \,\mathrm{sg}(A_{\mathrm{beh}}), \bar{\rho}_\theta \,\mathrm{sg}(A_{\mathrm{beh}})\right)\right]. \tag{4}$$

Sparse and sampled implementations The dense objectives above are full-vocabulary references. Practical OPD instead evaluates local KL losses on finite supports or sampled actions [13, 14], trading computation against support coverage and estimator variance. **Sparse top-k** implementations choose a support $S(s)$ and evaluate the corresponding restricted KL after renormalizing teacher and student distributions on $S(s)$. **Monte Carlo (MC)** implementations draw actions from a proposal distribution and estimate the corresponding local gradient; for reverse KL, this yields the student-sampled policy-gradient estimator. Details are in Section A.

4 Forward- and Reverse-KL OPD Under Staleness

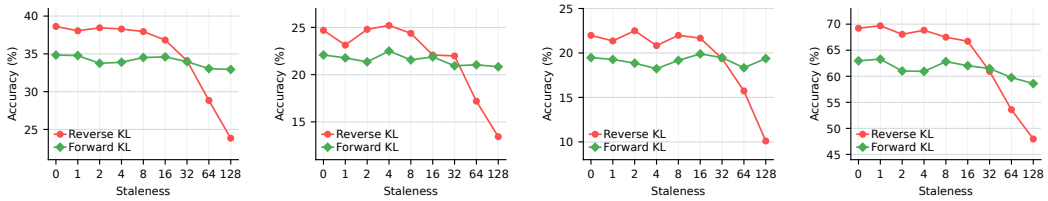
Asynchronous OPD has both prefix-level and action-level staleness. Once a rollout is generated, its visited prefixes are fixed, so an action-level estimator cannot change which states the learner sees. We therefore focus on the action-level staleness that estimator design can directly address.

4.1 Asynchronous OPD Setup

Asynchronous OPD is a cached-data pipeline: rollout first selects prefixes and actions, teacher scoring then annotates those actions, and the learner updates the student later. Unlike synchronous OPD, these stages are separated in time, so the visited prefixes, action cache, teacher scores, and update policy may be tied to different student versions. Figure 1 summarizes this cached-teacher setting and the estimator contrasts induced by the three-stage cache pipeline.

Teacher-cache constraint Full-vocabulary teacher logits allow dense KL computation, but caching and transferring them is prohibitively expensive, especially in an asynchronous pipeline. We therefore focus on sparse top-k supports and MC samples as the sparse and sampled cases.

Stage 1: Student rollout A rollout actor samples trajectories from a stale student p_{old}, which fixes the visited prefixes s. At each prefix, it stores cached actions $C_{\mathrm{old}}(s)$ together with their rollout-time log probabilities under p_{old}, such as a sampled token or a top-k support.



(a) Average (b) AIME24 (c) AIME25 (d) AMC

Figure 2: Accuracy comparison under staleness for forward- and reverse-KL OPD. Reverse KL starts higher at zero staleness but degrades faster as staleness grows; forward KL is flatter across the sweep.

Stage 2: Teacher scoring Let $C_{\text{score}}(s)$ denote the teacher-scored cache; it may come from the rollout cache $C_{\text{old}}(s)$ or be selected by the teacher at scoring time. Once teacher scoring is complete, teacher logits are available only on $C_{\text{score}}(s)$.

Stage 3: Student update By learner update time, the student has moved to the current policy p_θ. The learner can recompute $\log p_\theta(a \mid s)$ for $a \in C_{\text{score}}(s)$, but the current local OPD objective may place mass on actions outside this teacher-scored cache, such as the current student top-k support or current student-sampled actions. Thus the learner can update current student probabilities on cached actions, but cannot recover teacher signals for missing actions without additional teacher access.

Experimental setup Unless otherwise stated, we train a Qwen3-4B-Base student using a Qwen3-30B-A3B-Instruct-2507 teacher [26]. The training data is DeepMath [9], filtered to 57,630 math problems with difficulty level at least 6, and we report final-checkpoint Avg@32 accuracy on AIME24 [32], AIME25 [33], and AMC [15]. Experimental details are provided in Section B; dataset and metric details are in Section C.

4.2 Forward KL vs. Reverse KL Under Staleness

The KL direction fixes the action weighting: forward KL weights actions by the teacher q, whereas reverse KL weights them by the student p_θ. With cached teacher scores, this weighting difference becomes a support-ownership difference (Fig. 1). Under a scored-cache restriction, this makes forward KL less exposed to stale student action choices: it does not need to convert stale student-sampled actions into a current-student expectation. Reverse KL instead depends on student-weighted action terms, so the same asynchronous cache creates a different action-level staleness problem.

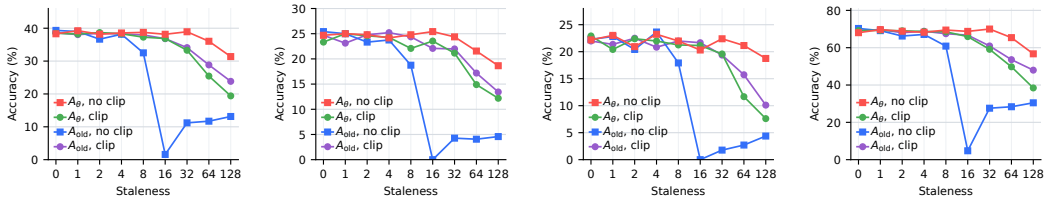
Experimental results Figure 2 compares representative practical OPD implementations from prior work: sparse top-k forward KL [21] and PPO-style reverse-KL surrogates [14, 27]. Reverse KL starts higher at zero staleness, but as staleness increases it drops faster and is eventually overtaken by forward KL. We therefore focus the rest of the staleness analysis on how to make reverse-KL OPD robust under larger rollout staleness.

Finding 1. Forward KL is teacher-weighted and robust to rollout staleness, whereas reverse KL is student-weighted and vulnerable to rollout staleness.

Two axes of reverse-KL staleness The cache analysis above suggests a possible mechanism for this gap: because reverse KL is weighted by the current student, stale teacher-scored caches may fail to cover actions needed by the current reverse-KL objective. In addition, reverse-KL policy-gradient updates can be instantiated with multiple stale-data surrogates, including PPO-style and asynchronous-RL variants. We therefore split the reverse-KL analysis into a **policy-gradient surrogate axis**, studied in Section 5, and a **cached-support axis**, studied in Section 6.

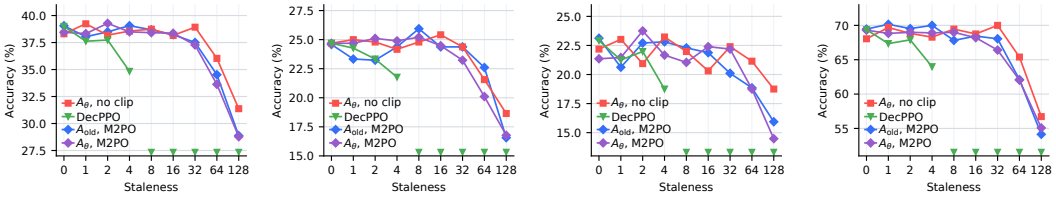
5 Reverse-KL: Policy-Gradient Surrogates Under Staleness

Reverse-KL OPD admits several policy-gradient surrogate choices under stale rollouts. This section compares which choices remain effective under staleness.



(a) Average (b) AIME24 (c) AIME25 (d) AMC

Figure 3: Accuracy comparison under staleness for the advantage-and-clipping ablation. Recomputing A_θ at learner time and avoiding clipping gives the most stable performance across the sweep, while clipping mainly helps the frozen A_{old} baseline.



(a) Average (b) AIME24 (c) AIME25 (d) AMC

Figure 4: Accuracy comparison under staleness for advanced asynchronous RL surrogates. Decoupled PPO [5] and M2PO [35] do not consistently improve over the simpler OPD/IS surrogate that recomputes A_θ without clipping; Decoupled PPO is clipped for readability because of low accuracy.

5.1 Policy-Gradient Surrogate Choices

PPO-style objective In the PPO-style surrogate in Eq. (4), the advantage is computed under the behavior policy and then held fixed during the learner update. In stale reverse-KL OPD, the behavior policy is the rollout student, so a mechanical PPO-style adaptation sets $p_{\text{beh}} = p_{\text{old}}$ and uses the rollout-time reverse-KL advantage $A_{\text{old}}(a, s) = \log q(a \mid s) - \log p_{\text{old}}(a \mid s)$ as A_{beh}, together with the clipped old-to-current ratio. The unclipped variant simply drops the clipped term.

Exact importance-sampling identity In contrast, rewriting the reverse-KL objective (Eq. (2)) by importance sampling suggests a different surrogate choice. With the current reverse-KL advantage $A_\theta(a, s) = \log q(a \mid s) - \log p_\theta(a \mid s)$, and assuming p_{old} has support wherever p_θ does, the current reverse-KL objective admits the exact old-to-current importance-sampling (IS) identity

$$D_R(\theta; s) = -\mathbb{E}_{a \sim p_\theta}\left[A_\theta(a, s)\right] = -\mathbb{E}_{a \sim p_{\text{old}}}\left[\rho_\theta(a, s) A_\theta(a, s)\right]. \tag{5}$$

For the policy-gradient update, the advantage is used as a stop-gradient weight; the derivative of the omitted A_θ term cancels by the score-function identity, as in Eq. (3). Thus the IS view points to the opposite surrogate choice from the mechanical PPO adaptation: recompute A_θ under the current student and use the old-to-current ratio without clipping, with A_θ treated as a stop-gradient advantage.

A two-by-two surrogate ablation The PPO-style adaptation and the OPD/IS identity suggest different surrogate choices. We therefore ablate the advantage (A_{old} versus A_θ) and whether to clip the ratio, with $\text{sg}(\cdot)$ denoting stop-gradient:

$$L_{\text{old}}^{\text{clip}}(\theta) = -\mathbb{E}_{a \sim p_{\text{old}}}\left[\min\left(\rho_\theta\, \text{sg}(A_{\text{old}}), \bar{\rho}_\theta\, \text{sg}(A_{\text{old}})\right)\right], \quad L_{\text{old}}^{\text{noclip}}(\theta) = -\mathbb{E}_{a \sim p_{\text{old}}}\left[\rho_\theta\, \text{sg}(A_{\text{old}})\right], \tag{6}$$

$$L_\theta^{\text{clip}}(\theta) = -\mathbb{E}_{a \sim p_{\text{old}}}\left[\min\left(\rho_\theta\, \text{sg}(A_\theta), \bar{\rho}_\theta\, \text{sg}(A_\theta)\right)\right], \quad L_\theta^{\text{noclip}}(\theta) = -\mathbb{E}_{a \sim p_{\text{old}}}\left[\rho_\theta\, \text{sg}(A_\theta)\right]. \tag{7}$$

Here $L_{\text{old}}^{\text{clip}}$ is the PPO-style adaptation, while L_θ^{noclip} is the OPD/IS surrogate.

Advanced asynchronous RL surrogates Decoupled PPO [5] and M2PO [35] are asynchronous RL surrogates designed to improve robustness to stale-policy updates. We evaluate whether these previously unstudied asynchronous RL surrogates also help OPD under staleness.

5.2 Experimental Results

Fig. 3 and Table 1a compare the four combinations of A_{old} versus A_θ and clipping versus no clipping. The best variant is the OPD/IS choice: A_θ without clipping. The PPO-style baseline, A_{old} with clipping, remains a strong stale-surrogate baseline. Clipping helps A_{old} by limiting stale, large-ratio updates, but hurts A_θ: recomputing A_θ already reduces the high-percentile ρ_θ tail at staleness 64 (Fig. 5), so clipping removes useful signal. Likewise, Fig. 4 and Table 1a show that advanced asynchronous RL surrogates such as decoupled PPO and M2PO do not outperform A_θ without clipping, which becomes our reference surrogate below.

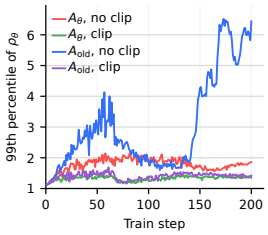


Figure 5: A_θ reduces the p99 ρ_θ tail under no clip.

Finding 2. The most effective reverse-KL correction is to recompute A_θ at learner time without clipping; advanced asynchronous RL surrogates such as decoupled PPO and M2PO do not improve over it.

6 Reverse-KL: Cached Supports Under Staleness

Having fixed A_θ without clipping as the reference surrogate, we now ask which cached actions provide the teacher scores needed to evaluate it, and how to improve this cached-support estimator. This cached-support axis is specific to OPD because teacher scoring is local and expensive: the teacher cache determines which actions have teacher scores available to the learner.

Sparse top-k: stale-support biased Although sparse top-k is biased relative to the dense reverse-KL objective, it is a practical low-variance approximation on the current student support $S_\theta(s) = \text{TopK}(p_\theta(\cdot \mid s), k)$. Under asynchronous rollout reuse, however, teacher scores are cached on the rollout-time support $S_{\text{old}}(s) = \text{TopK}(p_{\text{old}}(\cdot \mid s), k)$, which may miss actions in the current support $S_\theta(s)$. Reweighting within S_{old} cannot recover these missing teacher scores, so stale sparse top-k remains a support-mismatched approximation, not an exact correction of the current top-k objective.

One-sample MC: correctable but high variance Sampled-token MC instead caches an action drawn from a behavior distribution: $a \sim p_{\text{old}}(\cdot \mid s)$ together with $\log p_{\text{old}}(a \mid s)$. When the behavior policy covers the current policy support, exact old-to-current IS gives an unbiased estimator of the current reverse-KL fixed-prefix gradient. Thus one-sample MC is action-level correctable in expectation, but the resulting IS estimator can have high variance. This proposal-sampling structure is the key contrast with stale top-k, whose actions come from a deterministic stale support.

6.1 Proposed Solution: Multi-Sample MC

Multi-sample MC: correctable with reduced variance We propose multi-sample MC (Fig. 8), which, at each decoding timestep of a student rollout, draws multiple local next-token samples from the behavior policy without rolling them out into additional trajectories. It reduces one-sample MC variance by caching these local samples and averaging their IS-corrected gradients.

Multi-sample MC is especially natural in asynchronous OPD. In RL for LLM post-training, branching a prefix into multiple actions is expensive because each branch typically requires a full continuation before the reward or advantage can be evaluated. In synchronous OPD, sparse top-k already provides

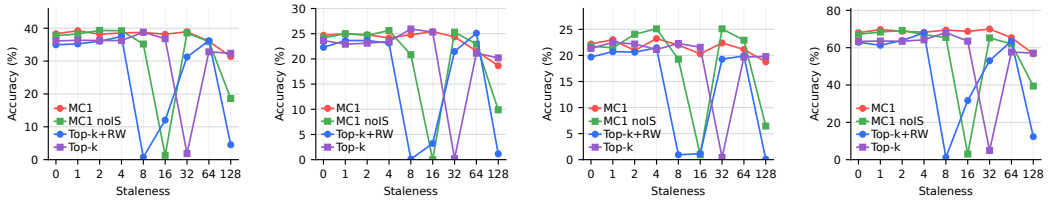
Table 1: Staleness-sensitivity slopes. Entries fit accuracy against $\log_2(\text{staleness} + 1)$; more negative values indicate stronger degradation with staleness.

(a) Policy-gradient surrogates

Benchmark	A_{old} (clip)	A_{old}	A_θ (clip)	A_θ	A_{old} (M2PO)	A_θ (M2PO)
AIME24	-1.44	-3.99	-1.64	-0.69	**-0.66**	-1.00
AIME25	-1.42	-3.69	-1.88	**-0.38**	-0.78	-0.81
AMC	-3.06	-8.00	-4.06	**-1.12**	-1.75	-1.72

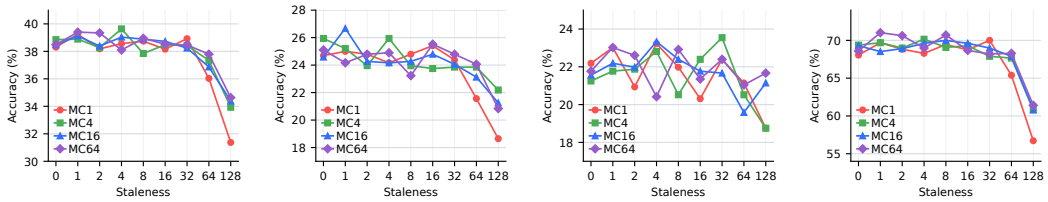
(b) Multi-sample MC

Benchmark	MC1	MC4	MC16	MC64
AIME24	-0.69	-0.42	-0.46	**-0.34**
AIME25	-0.38	-0.23	-0.24	**-0.11**
AMC	-1.12	-0.87	**-0.74**	-0.84



(a) Average (b) AIME24 (c) AIME25 (d) AMC

Figure 6: Accuracy comparison under staleness for sampled MC versus stale top-k. Top-k+RW denotes reweighting on the stale top-k support. Old-to-current IS corrects sampled MC in expectation, whereas reweighting cannot repair the missing teacher scores induced by stale top-k supports.

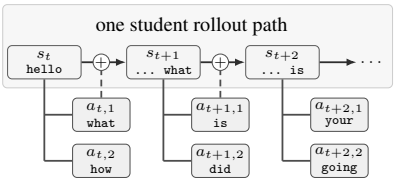


(a) Average (b) AIME24 (c) AIME25 (d) AMC

Figure 7: Accuracy comparison under staleness for multi-sample MC. Increasing the number of samples improves large-staleness behavior.

a low-variance approximation and one-sample MC provides an unbiased sampled gradient estimator, so there is little motivation to cache multiple sampled actions per prefix. Under asynchronous OPD, this tradeoff changes: sparse top-k becomes the stale fixed-support approximation analyzed above, and one-sample MC remains correctable but high-variance, making multi-sample MC a natural cached-support estimator for asynchronous OPD.

Concretely, at each visited timestep t with prefix s_t, rollout samples $a_{t,1}, \ldots, a_{t,m} \sim p_{\text{old}}(\cdot \mid s_t)$ and caches their rollout log probabilities and teacher scores. For notational simplicity, write $s = s_t$ and $a_i = a_{t,i}$ below. At learner time, we recompute $A_\theta(a_i, s)$ and use the averaged unclipped old-to-current IS surrogate $\widehat{L}_m^{\text{MC}}(\theta; s) = -\frac{1}{m}\sum_{i=1}^{m} \rho_\theta(a_i, s)\,\text{sg}(A_\theta(a_i, s))$. By linearity, the gradient has the same expectation as the one-sample MC estimator; averaging independent behavior-policy samples reduces the Monte Carlo variance. We measure this variance reduction at large staleness in Section E.



example: $m = 2$ local samples per timestep

Figure 8: Multi-sample MC ($m = 2$).

6.2 Experimental Results

Sparse top-k vs. one-sample MC Figure 6 compares one-sample MC and sparse top-k, with and without old-to-current reweighting. For one-sample MC, IS substantially improves robustness as staleness increases. For sparse top-k, the same reweighting does not improve performance, since it cannot recover missing current-support actions. As a result, one-sample MC with IS is the strongest of the four methods, consistent with the support-correctability analysis above. We include an additional ablation disentangling MC sample count from IS in Section F.

One-sample MC vs. multi-sample MC Figure 7 and Table 1b show that multi-sample MC improves one-sample MC at large staleness: $m = 4$ already gives a clear jump, while $m \in \{4, 16, 64\}$ performs similarly.

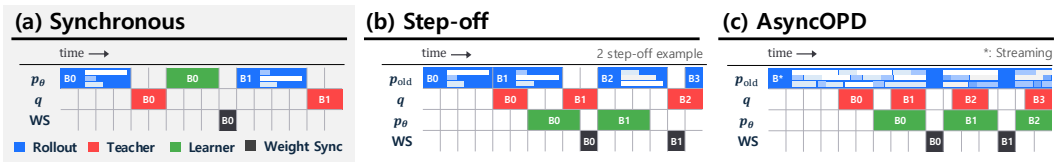


Figure 9: Scheduler comparison for synchronous OPD, step-off scheduling, and AsyncOPD. Synchronous OPD is barriered; step-off scheduling [21] overlaps stages but keeps gated rollout batches, while AsyncOPD streams rollout data to reduce long-tail waiting.

Table 2: AsyncOPD scheduler results for Qwen3-Base models. Train tok/s is training throughput; parentheses show speedup over the matched strict-sync baseline. Overlap is concurrent OPD-stage activity. Avg@32 is final AIME24. AsyncOPD achieves the highest throughput and overlap in all matched settings while maintaining comparable final accuracy.

Student	Scheduler	MC64			MC1		
		Train tok/s (× sync)	Overlap	Avg@32	Train tok/s (× sync)	Overlap	Avg@32
Qwen3-1.7B-Base	Strict sync	8.7k (1.00×)	0.81	8.85	8.6k (1.00×)	0.82	**8.65**
	Two-step-off	14.2k (1.64×)	1.49	8.23	18.5k (2.15×)	1.65	8.12
	AsyncOPD (ours)	**23.4k (2.70×)**	**2.13**	**9.38**	**28.1k (3.28×)**	**2.31**	8.44
Qwen3-4B-Base	Strict sync	9.5k (1.00×)	0.81	**25.00**	8.1k (1.00×)	0.82	23.33
	Two-step-off	12.4k (1.30×)	1.64	23.85	12.9k (1.60×)	1.65	**23.54**
	AsyncOPD (ours)	**15.8k (1.66×)**	**2.19**	**25.00**	**16.4k (2.03×)**	**2.27**	23.44
Qwen3-8B-Base	Strict sync	7.5k (1.00×)	0.81	26.56	6.5k (1.00×)	0.84	28.44
	Two-step-off	11.6k (1.55×)	1.76	26.56	9.6k (1.47×)	1.63	**28.85**
	AsyncOPD (ours)	**14.5k (1.94×)**	**2.24**	**28.65**	**10.6k (1.63×)**	**2.17**	27.50

Finding 3. One-sample MC is more effective than stale sparse top-k; multi-sample MC further improves this estimator by reducing one-sample variance while preserving MC correctability.

7 AsyncOPD: Fully Asynchronous OPD

AsyncOPD is our fully asynchronous OPD system. Following AReaL [5], it overlaps rollout, teacher scoring, and learner updates.

Scheduler The step-off scheduler family was originally implemented in VeRL [21]: a k-step-off run fixes rollout lag to k learner updates, but still waits for complete rollout batches. AsyncOPD streams examples instead: workers pause only for weight sync, preserve in-flight prefixes, teacher scoring consumes completed items, and the learner updates once a scored batch is ready (Fig. 9).

Experimental setup The main comparison uses Qwen3-{1.7B,4B,8B}-Base students with the Qwen3-30B-A3B-Instruct-2507 teacher. All runs use the same reverse-KL estimator: current-policy A_θ, no clipping, old-to-current IS, and either MC64 or MC1. We compare strict sync, two-step-off, and AsyncOPD for 100 training iterations on the same 8-GPU node; all AsyncOPD runs use $\tau = 4$. Section J gives GPU allocation, queue-depth, and scheduler details.

Experimental Results Table 2 reports training throughput, pipeline overlap (average concurrent OPD-stage activity), and final AIME24 Avg@32 for the Qwen3-Base students. AsyncOPD achieves the highest throughput and overlap in every matched comparison. In MC64, it reaches up to 2.7× the strict-sync throughput while achieving the best or tied-best final accuracy. MC1 shows the same trend: AsyncOPD delivers the highest throughput (up to 3.3× strict-sync) and overlap for every student, with competitive final accuracy. Train-time accuracy curves are reported in Section J.

8 Conclusion

We present the first systematic study of staleness in asynchronous on-policy distillation (OPD). Our results show that KL direction shapes the stale-data problem: forward KL remains robust to stale rollouts, whereas reverse KL is more vulnerable because it is student-weighted. In reverse-KL OPD, the most effective policy-gradient surrogate uses the current advantage recomputed at learner time without clipping; advanced asynchronous RL surrogates do not improve over this choice. We also find that stale student top-k supports are support-mismatched, whereas one-sample Monte Carlo (MC) remains correctable but high-variance. This contrast motivates multi-sample MC, which preserves MC correctability while reducing one-sample variance. Finally, we present and open-source **AsyncOPD**, a fully asynchronous OPD training pipeline built from these estimator choices, improving training efficiency while maintaining OPD quality.

Limitations and Future Work We study sparse and Monte Carlo OPD estimators, not dense full-vocabulary KL in the asynchronous setting. Although dense KL avoids cached-support mismatch, it is difficult to implement efficiently when rollout, teacher scoring, and learner updates are decoupled. KDFlow [31] suggests one path by transmitting teacher hidden states and recomputing student logits, but only for synchronous OPD. Extending this approach to asynchronous OPD while handling stale rollouts and preserving throughput is an important future direction. Our experiments are also limited to a single 8-GPU node by available resources, not by the pipeline itself; scaling to larger multi-node clusters remains future work.

Acknowledgments and Disclosure of Funding

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 04-26-03-0081, Energy-Efficient Training–Inference System Optimization for Reinforcement Learning-Based Post-Training). This work was also supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2026-RS-2023-00255968) grant funded by the Korea government (MSIT). This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

Competing interests. The authors declare no competing interests beyond the affiliations and funding disclosed in this paper.

References

[1] R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. R. Garea, M. Geist, and O. 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`.

[2] R. Agrawal, J. Fein-Ashley, and P. Rashidinejad. Reinforcement learning from rich feedback with distributional dagger. *arXiv preprint arXiv:2606.05152*, 2026.

[3] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[4] F. Devvrit, L. Madaan, R. Tiwari, R. Bansal, S. S. Duvvuri, M. Zaheer, I. S. Dhillon, D. Brandfonbrener, and R. Agarwal. The art of scaling reinforcement learning compute for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=FMjeC9Msws`.

[5] W. Fu, J. Gao, X. Shen, C. Zhu, Z. Mei, C. He, S. Xu, G. Wei, J. Mei, W. JIASHU, T. Yang, B. Yuan, and Y. Wu. AREAL: A large-scale asynchronous reinforcement learning system for language reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=X9diEuva9R`.

[6] W. Gao, Y. Zhao, D. An, T. Wu, L. Cao, S. Xiong, J. Huang, W. Wang, S. Yang, W. Su, et al. Rollpacker: Mitigating long-tail rollouts for fast, synchronous rl post-training. *arXiv preprint arXiv:2509.21009*, 2025.

[7] Y. Gu, L. Dong, F. Wei, and M. Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=5h0qf7IBZZ`.

[8] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[9] Z. He, T. Liang, J. Xu, Q. Liu, X. Chen, Y. Wang, L. Song, D. Yu, Z. Liang, W. Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

[10] M. Kim, M. Lee, S. Oh, K. Galim, D. Kim, C. Hooper, H. Singh, A. Gholami, H. I. Koo, and W. Kang. Efficientrollout: System-aware self-speculative decoding for rl rollouts. *arXiv preprint arXiv:2606.18967*, 2026.

[11] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[12] X. Li, S. Wu, and Z. Shen. A-3po: Accelerating asynchronous llm training with staleness-aware proximal policy approximation. *arXiv preprint arXiv:2512.06547*, 2025.

[13] Y. Li, Y. Zuo, B. He, J. Zhang, C. Xiao, C. Qian, T. Yu, H.-a. Gao, W. Yang, Z. Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

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

[15] Mathematical Association of America. American Mathematics Competitions – AMC. `https://maa.org/`, 2023. Accessed 2026-04-03.

[16] M. Noukhovitch, S. Huang, S. Xhonneux, A. Hosseini, R. Agarwal, and A. Courville. Faster, more efficient RLHF through off-policy asynchronous learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=FhTAG591Ve`.

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[19] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[20] G. Sheng, Y. Tong, B. Wan, W. Zhang, C. Jia, X. Wu, Y. Wu, X. Li, C. Zhang, Y. Peng, et al. Laminar: A scalable asynchronous rl post-training framework. *arXiv preprint arXiv:2510.12633*, 2025.

[21] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[22] M. Song and M. Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

[23] B. Xiao, B. Xia, B. Yang, B. Gao, B. Shen, C. Zhang, C. He, C. Lou, F. Luo, G. Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

[24] Y. Xu, H. Sang, Z. Zhou, R. He, Z. Wang, and A. Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

[25] R. Yan, Y. Jiang, T. Wu, J. Gao, Z. Mei, W. Fu, H. Mai, W. Wang, Y. Wu, and B. Yuan. Areal-hex: Accommodating asynchronous rl training over heterogeneous gpus. *arXiv preprint arXiv:2511.00796*, 2025.

[26] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[27] W. Yang, W. Liu, R. Xie, K. Yang, S. Yang, and Y. Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

[28] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[29] A. Zeng, X. Lv, Z. Hou, Z. Du, Q. Zheng, B. Chen, D. Yin, C. Ge, C. Huang, C. Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

[30] K. Zhang, Y. Zuo, B. He, Y. Sun, R. Liu, C. Jiang, Y. Fan, K. Tian, G. Jia, P. Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

[31] S. Zhang, X. Zhang, T. Zhang, B. Hu, Y. Chen, and J. Xu. Kdflow: A user-friendly and efficient knowledge distillation framework for large language models. *arXiv preprint arXiv:2603.01875*, 2026.

[32] Y. Zhang and T. Math-AI. AIME 2024. `https://huggingface.co/datasets/Maxwell-Jia/AIME_2024`, 2024. Hugging Face dataset; accessed 2026-04-03.

[33] Y. Zhang and T. Math-AI. AIME 2025. `https://huggingface.co/datasets/yentinglin/aime_2025`, 2025. Hugging Face dataset; accessed 2026-04-03.

[34] S. Zhao, Z. Xie, M. Liu, J. Huang, G. Pang, F. Chen, and A. Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

[35] H. Zheng, J. Zhao, and B. Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=IIgl5MWelz`.

[36] Y. Zhong, Z. Zhang, X. Song, H. Hu, C. Jin, B. Wu, N. Chen, Y. Chen, Y. Zhou, C. Wan, et al. Streamrl: Scalable, heterogeneous, and elastic rl for llms with disaggregated stream generation. *arXiv preprint arXiv:2504.15930*, 2025.

A Sparse and Monte Carlo Reverse-KL Implementations

A.1 Sparse Top-k Reverse-KL OPD

The dense reverse-KL objective in Eq. (2) sums over the full vocabulary. A sparse top-k implementation instead evaluates reverse KL on a finite student support

$$S_\theta(s) = \mathrm{TopK}(p_\theta(\cdot \mid s), k)\,. \tag{8}$$

For any support S, define the restricted normalizers $Z_p^S(s) = \sum_{u\in S} p_\theta(u \mid s)$ and $Z_q^S(s) = \sum_{u\in S} q(u \mid s)$, and the renormalized distributions

$$\tilde{p}_\theta^S(a \mid s) = \frac{p_\theta(a \mid s)\mathbf{1}[a \in S]}{Z_p^S(s)}, \qquad \tilde{q}^S(a \mid s) = \frac{q(a \mid s)\mathbf{1}[a \in S]}{Z_q^S(s)}. \tag{9}$$

The sparse reverse-KL objective is

$$\begin{aligned} D_R^S(\theta; s) &= \mathrm{KL}\big(\tilde{p}_\theta^S(\cdot \mid s) \,\|\, \tilde{q}^S(\cdot \mid s)\big) \\ &= -\sum\nolimits_{a\in S} \tilde{p}_\theta^S(a \mid s)\left(\log \tilde{q}^S(a \mid s) - \log \tilde{p}_\theta^S(a \mid s)\right). \end{aligned} \tag{10}$$

In practice, when $S = S_\theta(s)$, we treat the selected top-k support as fixed during the local update.

Table 3: Experimental settings.

Setting	Value
Student	Qwen3-4B-Base for main staleness/estimator experiments; AsyncOPD experiments also use Qwen3-{1.7B,4B,8B}-Base and thinking-disabled Qwen3-{1.7B,4B,8B} [26].
Teacher	Qwen3-30B-A3B-Instruct-2507 [26]
Training data	DeepMath [9] filtered to 57,630 math problems with difficulty level at least 6
Prompt / response lengths	2,048 prompt tokens / 16,384 response tokens
Training horizon	200 training iterations; in the common setup each iteration contains four mini-batch optimizer updates
Optimization	batch size 256, mini-batch size 64
Optimizer	AdamW-style optimizer with learning rate 3×10^{-6}, constant schedule, weight decay 0.01
PPO clipping	$\epsilon = 0.2$ for clipped PPO-style surrogates
M2PO budget	0.01
Rollout generation	temperature 1.0, top-$p = 1.0$, no top-k truncation
Evaluation generation	32 samples per problem, temperature 1.0
Implementation	vLLM rollout and teacher scoring; PyTorch FSDP learner training
Hardware	Single 8×B200 node
GPU allocation	One teacher GPU; strict sync time-shares seven GPUs between rollout and training; stale-cache sweeps and AsyncOPD use concurrent rollout and trainer GPU pools

A.2 Monte Carlo Reverse-KL OPD

Let $A_\theta(a, s) = \log q(a \mid s) - \log p_\theta(a \mid s)$. From Eq. (3), the dense reverse-KL gradient can be written as

$$\nabla_\theta D_R(\theta; s) = -\mathbb{E}_{a \sim p_\theta(\cdot|s)} \left[A_\theta(a, s) \nabla_\theta \log p_\theta(a \mid s) \right]. \tag{11}$$

A one-sample current-policy Monte Carlo estimator is therefore

$$\widehat{g}_{\mathrm{MC}}(s, a) = -A_\theta(a, s) \nabla_\theta \log p_\theta(a \mid s), \qquad a \sim p_\theta(\cdot \mid s). \tag{12}$$

With m independent samples $a_i \sim p_\theta(\cdot \mid s)$, the corresponding multi-sample estimator averages the same local term:

$$\widehat{g}_m(s) = -\frac{1}{m} \sum\nolimits_{i=1}^{m} A_\theta(a_i, s) \nabla_\theta \log p_\theta(a_i \mid s). \tag{13}$$

B Experimental Details

This section details the experimental setup. Unless explicitly stated otherwise, experiments use the common setup in Table 3 and report final-checkpoint Avg@32 accuracy. Our implementation uses vLLM [11] for rollout generation and teacher scoring, PyTorch FSDP [17] for learner training, and runs each experiment on a single 8×B200 node. Individual experiments take roughly 1–12 hours, depending on the setting. Asset URLs, license names, and versions are summarized in Table 5.

Constructing the staleness axis. The main text uses staleness as an experimental control over how old the cached rollout data is when the learner updates on it. In all staleness plots and tables in Sections 4 to 6, staleness is measured in train-batch steps. One train-batch step is one logical rollout batch consumed by the learner for a training iteration. The sweep value k is therefore the target number of train-batch steps by which the consumed cache trails the current learner; equivalently, it is the target cache depth in logical rollout batches. A value $k = 0$ is synchronous: rollout, teacher scoring, training, and weight synchronization occur in strict sequence. For $k > 0$, the run first generates exactly k rollout batches with the initial student snapshot before the first learner update. Training then consumes the oldest available generated batch; after each learner update and weight synchronization, a new rollout batch is generated with the latest student snapshot whenever needed to restore the target cache depth.

Table 4: Evaluation datasets.

Dataset	Problems
AIME 2024	30
AIME 2025	30
AMC 2023	40

This protocol is the operational source of the prefix- and action-level staleness discussed in the main text: the consumed prefixes and cached actions come from an older rollout student, while the update is applied to the current student. For a consumed rollout batch, let t_{roll} be the train-batch index of the student snapshot used for generation and t_{train} be the train-batch index at learner time. The staleness used in the plots is

$$\Delta_{\mathrm{batch}} = t_{\mathrm{train}} - t_{\mathrm{roll}}.$$

All examples in a logical batch share the same rollout snapshot and therefore share the same Δ_{batch}. Under the controlled cache protocol, Δ_{batch} ramps as $0, 1, 2, \ldots$ while the initial cache is drained and then plateaus at k. Thus a 64-batch target cache depth is plotted as staleness 64. A train-batch step can contain multiple mini-batch optimizer updates; in the common setup, $B = 256$ and $B_{\mathrm{mini}} = 64$, so each train-batch step contains $M = B/B_{\mathrm{mini}} = 4$ optimizer updates. This conversion is useful for implementation accounting, but it is not the staleness axis used in the plots.

We use k as the x-axis because it is the controlled train-batch staleness intervention shared across methods. The sweep covers $k \in \{0, 1, 2, 4, 8, 16, 32, 64, 128\}$ across the forward-KL, reverse-KL / PPO-style, M2PO / DecPPO, top-k, and Monte-Carlo support-size variants; apart from the estimator choice and k, these runs share the common model, data, batch-size, generation, and evaluation settings in Table 3.

C Datasets and Metrics

Training data. We filter the DeepMath dataset [9] to retain 57,630 math problems with difficulty level greater than or equal to 6, and use this filtered subset as the training data.

Evaluation datasets. Table 4 lists the evaluation datasets used: AIME 2024 [32], AIME 2025 [33], and AMC 2023 [15]. AIME24 is evaluated every 20 steps. The remaining datasets are only evaluated for the final checkpoint.

Accuracy metric. Evaluation samples 32 responses per problem. For a dataset D, the reported Avg@32 is the mean per-problem pass rate,

$$\mathrm{Avg@32}(D) = 100 \cdot \frac{1}{|D|} \sum_{i \in D} \frac{c_i}{32}, \tag{14}$$

where c_i is the number of sampled responses judged correct for problem i. Paper tables and plots use Avg@32 unless noted otherwise.

D Existing Asset Licenses

Table 5 lists the reused assets.

E Multi-Sample MC Variance at Large Staleness

We measure how multi-sample MC reduces the variance of the old-to-current IS surrogate at large staleness. At timestep t with prefix s_t, local MC actions $a_{t,1}, \ldots, a_{t,m}$ are sampled iid with replacement from $p_{\mathrm{old}}(\cdot \mid s_t)$ (duplicates are allowed), and the learner evaluates

$$\widehat{L}_m^{\mathrm{MC}}(\theta; s_t) = -\frac{1}{m} \sum_{i=1}^{m} \rho_\theta(a_{t,i}, s_t)\, \mathrm{sg}(A_\theta(a_{t,i}, s_t)), \qquad \rho_\theta(a, s) = \frac{p_\theta(a \mid s)}{p_{\mathrm{old}}(a \mid s)}. \tag{15}$$

Table 5: Existing assets used in this work, with source URLs, license names, and versions.

Category	Asset	URL	License name	Version
Dataset	DeepMath	https://huggingface.co/datasets/zwhe99/DeepMath-103K	MIT	2025
Dataset	AIME 2024	https://huggingface.co/datasets/Maxwell-Jia/AIME_2024	MIT	2024
Dataset	AIME 2025	https://huggingface.co/datasets/yentinglin/aime_2025	Not specified	2025
Dataset	AMC 2023 / American Mathematics Competitions	https://huggingface.co/datasets/math-ai/amc23	Not specified	2023
Model	Qwen3-4B-Base	https://huggingface.co/Qwen/Qwen3-4B-Base	Apache-2.0	Not specified
Model	Qwen3-1.7B-Base	https://huggingface.co/Qwen/Qwen3-1.7B-Base	Apache-2.0	Not specified
Model	Qwen3-8B-Base	https://huggingface.co/Qwen/Qwen3-8B-Base	Apache-2.0	Not specified
Model	Qwen3-4B	https://huggingface.co/Qwen/Qwen3-4B	Apache-2.0	Not specified
Model	Qwen3-1.7B	https://huggingface.co/Qwen/Qwen3-1.7B	Apache-2.0	Not specified
Model	Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B	Apache-2.0	Not specified
Model	Qwen3-30B-A3B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507	Apache-2.0	Qwen3-2507
Software	vLLM	https://github.com/vllm-project/vllm	Apache-2.0	0.16.0
Software	PyTorch / PyTorch FSDP	https://github.com/pytorch/pytorch	BSD-3-Clause	2.9.1

Using the Qwen3-4B-Base staleness-128 runs, we report R_m^{local} in the fixed-prefix column and R_m^{seq} in the sequence-level column of Table 6. Both ratios are normalized to the corresponding $m = 1$ estimator within the same old-to-current pair:

$$R_m^{\mathrm{local}} = \frac{\mathbb{E}_{s_t}\left[\mathrm{Var}_{a_{t,1},\ldots,a_{t,m}\sim p_{\mathrm{old}}(\cdot|s_t)}\left(\widehat{L}_m^{\mathrm{MC}}(\theta; s_t) \mid s_t\right)\right]}{\mathbb{E}_{s_t}\left[\mathrm{Var}_{a_t\sim p_{\mathrm{old}}(\cdot|s_t)}\left(\widehat{L}_1^{\mathrm{MC}}(\theta; s_t) \mid s_t\right)\right]}, \tag{16}$$

$$R_m^{\mathrm{seq}} = \frac{\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\widehat{L}_m^{\mathrm{MC}}(\theta; s_t)\right]}{\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\widehat{L}_1^{\mathrm{MC}}(\theta; s_t)\right]}. \tag{17}$$

For R_m^{seq}, the generated prefix path $s_{1:T}$ is fixed when computing the variance; the MC samples are local scorer queries at each fixed prefix, not separate rollout branches.

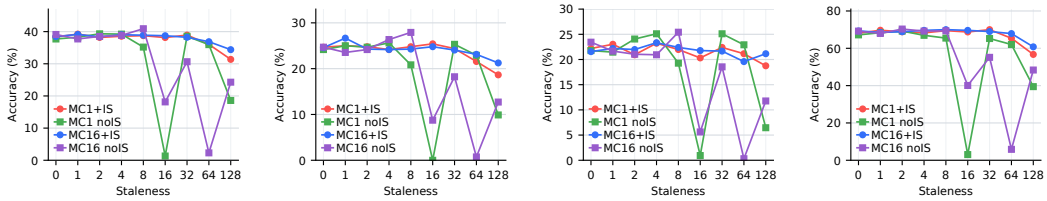
Table 6: MC variance ratios at large staleness. The fixed-prefix column isolates local next-token action-sampling variance; the sequence-level column averages the same estimator over generated timesteps before computing variance.

Diagnostic m	Fixed-prefix ratio R_m^{local}	Sequence-level ratio R_m^{seq}
1	1.000	1.000
4	0.255	0.338
16	0.0588	0.152
64	0.0149	0.112

Table 6 shows that larger m consistently reduces variance. Fixed-prefix ratios closely follow the $1/m$ reference ($m = 64$ leaves 1.49% of the one-sample local variance, versus $1/64 = 1.56\%$). After timestep aggregation, $m = 64$ still leaves only 11.2% of the one-sample sequence-level variance, showing that multi-sample MC reduces variance in practice even though sequence-level averaging makes the reduction less extreme than the local next-token effect.

F Importance-Sampling Ablation

Before treating multi-sampling as an improvement, we separate it from IS. Increasing m changes the Monte Carlo variance of the estimator. It does not define the target distribution. Old-to-current IS is still the mechanism that turns behavior-policy samples into an estimator of the current reverse-KL local gradient. Figure 10 compares MC1 and MC16 with and without IS to test this distinction directly.



(a) Average (b) AIME24 (c) AIME25 (d) AMC

Figure 10: Accuracy comparison under staleness for MC importance-sampling ablations. Increasing the number of samples reduces Monte Carlo variance, but old-to-current IS is still needed to correct stale-policy sampling.

G Theoretical Properties of Multi-Sample MC

We provide the following guarantees for the multi-sample MC estimator at a fixed stale prefix s, separating action-level from prefix-level staleness. All expectations and variances below are over action samples for a fixed old-to-current policy pair. The estimator is

$$\widehat{L}_m^{\mathrm{MC}}(\theta; s) = -\frac{1}{m}\sum_{i=1}^{m} \rho_\theta(a_i, s)\,\mathrm{sg}(A_\theta(a_i, s))\,, \qquad a_i \overset{\text{iid}}{\sim} p_{\mathrm{old}}(\cdot \mid s),$$

where

$$\rho_\theta(a, s) = \frac{p_\theta(a \mid s)}{p_{\mathrm{old}}(a \mid s)}, \qquad A_\theta(a, s) = \log q(a \mid s) - \log p_\theta(a \mid s),$$

and sg denotes stop-gradient.

Unbiasedness. Under full-support rollout sampling and unclipped importance sampling,

$$\begin{aligned}\mathbb{E}\Big[\nabla_\theta \widehat{L}_m^{\mathrm{MC}}(\theta; s) \mid s\Big] &= -\mathbb{E}_{a\sim p_{\mathrm{old}}}\left[\rho_\theta(a, s) A_\theta(a, s) \nabla_\theta \log p_\theta(a \mid s)\right] \\ &= -\mathbb{E}_{a\sim p_\theta}\left[A_\theta(a, s) \nabla_\theta \log p_\theta(a \mid s)\right] \\ &= \nabla_\theta D_R(\theta; s).\end{aligned}$$

Thus, multi-sample MC is unbiased for the gradient of the dense reverse-KL objective at the fixed prefix.

Exact variance reduction. Because the estimator averages m iid behavior-policy samples,

$$\mathrm{Var}\Big(\widehat{L}_m^{\mathrm{MC}} \mid s\Big) = \frac{1}{m}\,\mathrm{Var}\Big(\widehat{L}_1^{\mathrm{MC}} \mid s\Big)\,.$$

The same exact $1/m$ reduction holds for the variance of each gradient component for any fixed old-to-current policy pair.

Staleness-dependent bound. For $a \sim p_{\mathrm{old}}(\cdot \mid s)$,

$$\mathbb{E}[\rho_\theta] = 1, \qquad \mathrm{Var}(\rho_\theta) = \chi^2(p_\theta \| p_{\mathrm{old}}).$$

Here, $\chi^2(p_\theta \| p_{\mathrm{old}})$ denotes the chi-square divergence between the token distributions at s. If $|A_\theta(a, s)| \le M$, the scalar surrogate therefore satisfies

$$\mathrm{Var}\Big(\widehat{L}_m^{\mathrm{MC}} \mid s\Big) \le \frac{M^2\big(1 + \chi^2(p_\theta \| p_{\mathrm{old}})\big) - D_R(\theta; s)^2}{m}.$$

This bound makes the effect of staleness explicit: greater old-to-current policy mismatch can increase one-sample variance, while multi-sampling reduces it exactly by a factor of m.

Table 7: A 200-step head-to-head comparison of forward KL and repaired reverse KL using Qwen3-4B-Base. Avg@32 is final AIME24 accuracy; throughput speedups are relative to the strict-sync baseline for the same objective.

Objective	Scheduler	Train tok/s (× sync)	Overlap	Avg@32
Forward KL	Strict sync	7.2k (1.00×)	0.82	20.73
Forward KL	AsyncOPD	17.9k (2.47×)	2.26	20.52
Repaired reverse KL	Strict sync	8.6k (1.00×)	0.79	24.90
Repaired reverse KL	AsyncOPD	14.0k (1.62×)	2.00	24.69

Table 8: End-to-end AsyncOPD results on code generation using MC64. Throughput speedups are relative to the matched strict-sync baseline. Mean Pass@1 is averaged across HumanEval+, LiveCodeBench v6, and MBPP+.

Student	Scheduler	Train tok/s (× sync)	Mean Pass@1
Qwen3-1.7B-Base	Strict sync	12.2k (1.00×)	28.06
Qwen3-1.7B-Base	AsyncOPD (ours)	30.3k (2.48×)	28.56
Qwen3-4B-Base	Strict sync	8.3k (1.00×)	48.22
Qwen3-4B-Base	AsyncOPD (ours)	18.3k (2.19×)	51.81
Qwen3-8B-Base	Strict sync	8.4k (1.00×)	57.06
Qwen3-8B-Base	AsyncOPD (ours)	13.2k (1.58×)	56.04

H Forward KL versus Repaired Reverse KL

Forward KL is robust to staleness but starts from a lower accuracy at zero staleness than reverse KL. Our repaired reverse KL is designed to retain the higher accuracy of reverse KL under asynchronous training. Table 7 reports the results of a 200-step head-to-head comparison. Under AsyncOPD, repaired reverse KL retains a 4.17-point accuracy advantage over forward KL, achieves $1.62\times$ the strict-synchronous throughput, and remains within 0.21 points of its strict-synchronous accuracy.

I End-to-End Results on Code Generation

We conducted code-generation experiments on HumanEval+, LiveCodeBench v6, and MBPP+ using approximately 25.1K training problems from `sungyub/eurus-2-code-verl`, while keeping all other training settings identical to the mathematical-reasoning experiments. All results report mean Pass@1 across the three benchmarks. We compared AsyncOPD with strict-synchronous training using MC64 under the same settings, with results reported in Table 8. These results show that AsyncOPD improves throughput from $1.58\times$ to $2.48\times$ while preserving comparable final accuracy.

J AsyncOPD Scheduler Details

This appendix gives the implementation details omitted from Section 7. Our scheduler, AsyncOPD, follows the fully asynchronous systems structure of AReaL [5], but the queue contains OPD cache items rather than reward-labeled RL trajectories.

Queue interface. The pipeline has three long-running stages: rollout generation, teacher scoring, and learner training. Rollout workers sample trajectories from their latest synchronized student snapshot. For each visited prefix s, they cache the MC actions, rollout log probabilities under p_{old}, and the rollout student version. The main scheduler comparison uses MC64; the MC1 runs use the same queue interface with one cached action. The teacher scores the cached actions. The learner then recomputes $\log p_\theta(a \mid s)$ and $A_\theta(a, s)$ under the current student, and applies the unclipped old-to-current IS estimator from Section 6.

Weight synchronization and queue capacity. During AsyncOPD weight synchronization, rollout workers pause generation. In the keep-mode used for the scheduler experiments, in-flight requests are not discarded: the already sampled token prefix is kept, the student weights are updated, and the running-request prefix cache is reset so the engine rebuilds the attention state for that prefix under the new weights before generation resumes. Thus, tokens before the synchronization point are reused rather than regenerated, while later tokens are sampled under the new student snapshot. Each completed sample records the token index at which the active weight version changes. The queue-depth parameter τ is enforced as a capacity bound rather than as a learner-side drop rule. The coordinator creates a semaphore with $(\tau + 1)B$ permits, where B is the effective train batch size. The prompt feeder acquires one permit before submitting a prompt to rollout, and the train dispatcher releases permits only after the corresponding samples have been consumed by a learner update. During weight synchronization, a sync gate prevents the feeder from using newly released permits until rollout workers have received the updated weights. Thus, smaller τ limits the amount of unconsumed rollout work in the pipeline, while larger τ permits a deeper backlog and more overlap. The queues themselves remain FIFO; items are not evicted for being stale.

Training throughput metric. The table reports training throughput. Let n_j be the number of response tokens used by learner update j, and let t_j be the train wall-clock time after that update. Discarding the first five warmup updates, we compute

$$\text{throughput} = \frac{\sum_{j=6}^{J} n_j}{\sum_{j=6}^{J}(t_j - t_{j-1})} = \frac{\sum_{j=6}^{J} n_j}{t_J - t_5}.$$

Speedups are normalized to the strict-sync run with the same student and MC setting.

Pipeline overlap metric. Let $\mathcal{S} = \{\text{rollout}, \text{teacher}, \text{train}\}$. For teacher and train, we merge overlapping busy intervals within each stage and compute the merged busy time T_s. Rollout has N_r workers, so we first merge intervals within each worker i and then define the rollout-stage busy time as the worker-normalized average

$$T_{\text{rollout}} = \frac{1}{N_r} \sum_{i=1}^{N_r} T_{\text{rollout},i}.$$

Let T_{wall} be the elapsed train wall-clock interval from the first to the last recorded pipeline-stage interval. We define

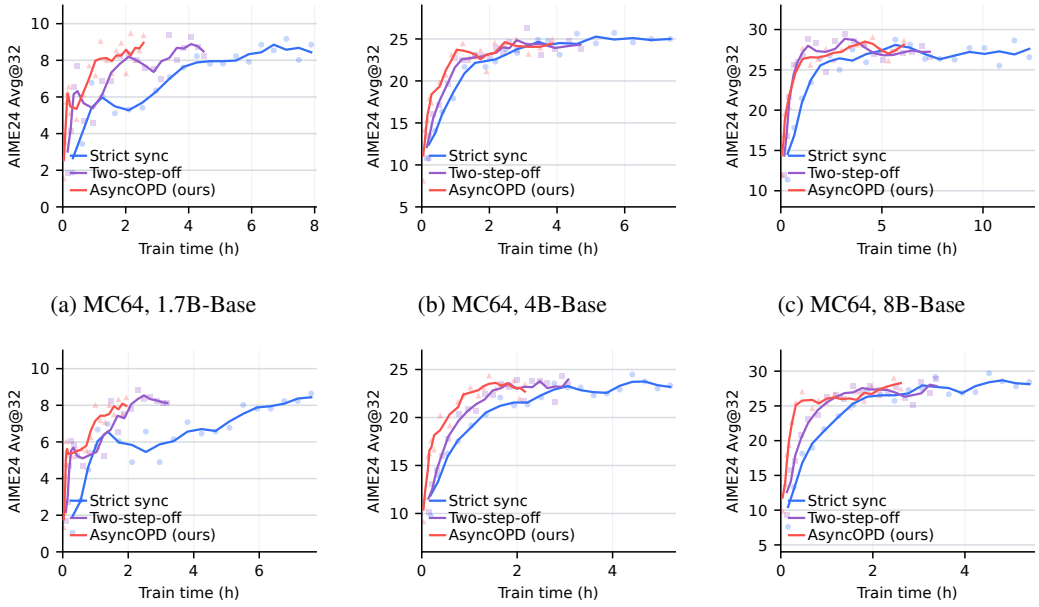
$$\text{overlap} = \frac{\sum_{s \in \mathcal{S}} T_s}{T_{\text{wall}}}.$$

A mostly serial schedule has overlap near 1, and the maximum remains 3: all rollout workers, teacher scoring, and training busy for the full interval.

Hardware and testing protocol. Each scheduler run uses one 8-GPU node. One GPU is reserved for teacher scoring. The remaining seven GPUs are the rollout/training pool. Rollout generation uses data parallelism, and learner training uses PyTorch FSDP. Strict sync runs time-share this pool: all seven GPUs run rollout, then all seven switch to training, and the cycle repeats. The two-step-off and our AsyncOPD runs split the same seven GPUs concurrently: 4 GPUs for rollout workers and 3 GPUs for the FSDP trainer.

For each student size and MC setting, we compare strict sync, two-step-off, and our AsyncOPD scheduler with the same teacher, training data, evaluation metrics, and reverse-KL estimator: current-policy A_θ, no clipping, and old-to-current IS correction. Two-step-off fixes a two-update offset between rollout and the learner update that consumes it, so stale rollout reuse is static and controlled rather than produced by queue timing. We use this offset because it is the fastest static step-off schedule under the 4-rollout/3-trainer split. The OPD pipeline has three serial stages: rollout generation, teacher scoring, and learner training. Therefore, a two-step offset is enough to keep all stages occupied in the gated schedule. Larger offsets only make the consumed data older; they do not create another OPD stage to overlap or remove the step-off batch barrier. We measure final-checkpoint Avg@32 and train wall-clock time over the same training horizon.

Qwen3-Base train-time accuracy. Figure 11 provides the train-time view for Qwen3-Base students. AsyncOPD reaches later checkpoints sooner, so accuracy improves earlier in wall-clock time across student sizes and MC settings.



(a) MC64, 1.7B-Base (b) MC64, 4B-Base (c) MC64, 8B-Base

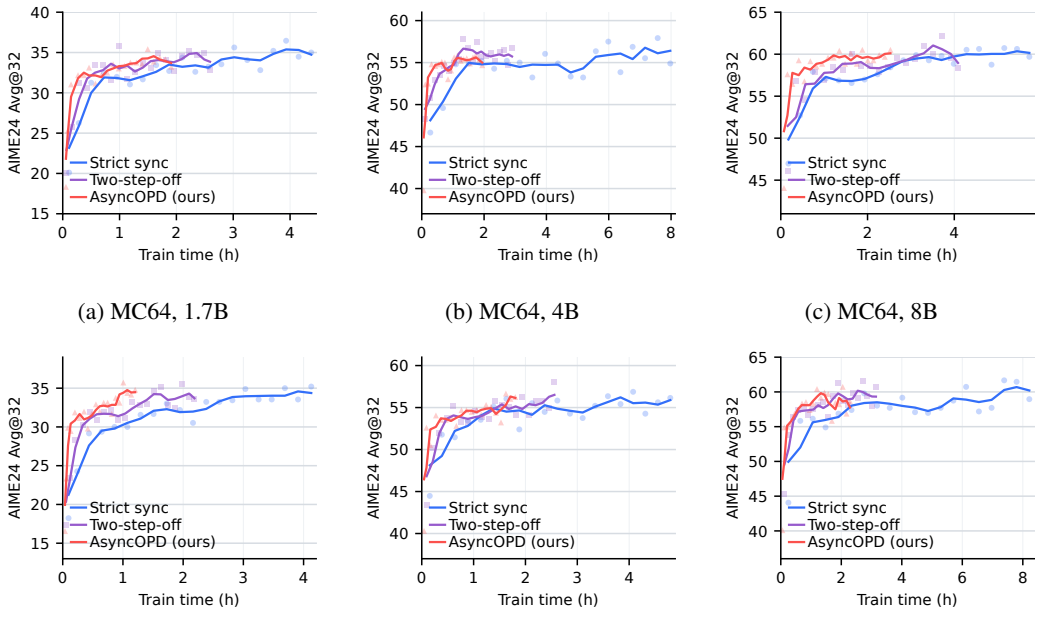
(d) MC1, 1.7B-Base (e) MC1, 4B-Base (f) MC1, 8B-Base

Figure 11: Train-time AIME24 Avg@32 for Qwen3-Base students with MC64 and MC1. Lines are 3-point moving averages; faint markers are raw evaluations; colors denote scheduler. AsyncOPD reaches later checkpoints sooner, so its accuracy improves earlier in wall-clock time.

Table 9: Additional AsyncOPD scheduler results for Qwen3 students with thinking disabled. Train tok/s is training throughput; parentheses show speedup over the matched strict-sync baseline. Overlap is concurrent OPD-stage activity. Avg@32 is final AIME24. AsyncOPD achieves the highest throughput and overlap in all matched settings while maintaining comparable final accuracy.

Student	Scheduler	MC64			MC1		
		Train tok/s (× sync)	Overlap	Avg@32	Train tok/s (× sync)	Overlap	Avg@32
1.7B	Strict sync	11.2k (1.00×)	0.78	**35.00**	11.7k (1.00×)	0.78	**35.21**
	Two-step-off	18.0k (1.61×)	1.58	32.81	20.8k (1.78×)	1.71	33.54
	AsyncOPD (ours)	**24.2k (2.16×)**	**2.07**	33.23	**37.0k (3.17×)**	**2.56**	34.79
4B	Strict sync	4.5k (1.00×)	0.87	54.90	7.4k (1.00×)	0.83	56.15
	Two-step-off	11.5k (2.58×)	1.54	**56.77**	13.0k (1.75×)	1.63	**58.02**
	AsyncOPD (ours)	**17.0k (3.82×)**	**2.15**	54.69	**18.5k (2.49×)**	**2.26**	56.25
8B	Strict sync	5.6k (1.00×)	0.86	59.69	3.9k (1.00×)	0.87	58.96
	Two-step-off	7.5k (1.34×)	1.45	58.33	9.6k (2.45×)	1.61	**60.73**
	AsyncOPD (ours)	**10.1k (1.81×)**	**2.11**	**60.52**	**10.9k (2.78×)**	**2.17**	58.65

Additional Qwen3 AsyncOPD results. For the Qwen3 1.7B, 4B, and 8B student rows, we disable thinking at the tokenizer prompt-formatting level: prompt construction uses the Qwen3 tokenizer's non-thinking chat-template mode before rollout and evaluation. Table 9 and Fig. 12 report this comparison. The systems pattern matches the Qwen3-Base results: AsyncOPD has the highest throughput and overlap, reaching up to 3.8× strict-sync throughput on MC64 and up to 3.2× on MC1. The train-time accuracy plots show the same wall-clock pattern as the main Qwen3-Base results: AsyncOPD reaches later checkpoints sooner, so accuracy improves earlier across student sizes and MC settings.



(a) MC64, 1.7B (b) MC64, 4B (c) MC64, 8B

(d) MC1, 1.7B (e) MC1, 4B (f) MC1, 8B

Figure 12: Train-time AIME24 Avg@32 for Qwen3 1.7B, 4B, and 8B students with thinking disabled, using MC64 and MC1. Lines are 3-point moving averages; faint markers are raw evaluations; colors denote scheduler. AsyncOPD reaches later checkpoints sooner, so its accuracy improves earlier in wall-clock time.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's main contributions and scope: the staleness analysis of asynchronous OPD, the forward-/reverse-KL comparison, the reverse-KL surrogate study, multi-sample MC, and the AsyncOPD system. These claims are supported by the main experiments and analysis in Sections 4–7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and future work in Section 8, including the focus on sparse and Monte Carlo OPD estimators and the absence of full-vocabulary dense-KL evaluation in the asynchronous pipeline.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For the theoretical and estimator-level claims, the paper states the relevant OPD setup and assumptions and provides derivations for the forward-/reverse-KL formulations, the reverse-KL policy-gradient form, and the old-to-current importance-sampling identity. The paper does not rely on separate theorem-style results beyond these analytical derivations.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses the main information needed to reproduce the experimental results, including model choices, datasets, training horizon, optimization settings, generation settings, evaluation protocol, implementation stack, scheduler configuration, and hardware setup. Additional estimator and scheduler details are provided in the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper states that the AsyncOPD implementation is released through an anonymized repository, and the datasets and models used in the experiments are public or cited. The appendix provides the experimental setup needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training and test settings, including the filtered training data, prompt and response lengths, optimizer, rollout generation settings, evaluation sampling protocol, implementation stack, and hardware allocation (see Appendix and code implementation).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The main experiments are computationally expensive, so repeating every experiment across multiple random seeds was not feasible within our available compute budget. We acknowledge this limitation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides the compute setup used for the experiments, including the single $8\times$B200 node, GPU allocation across teacher scoring, rollout, and training, and the throughput and overlap measurements used in the scheduler comparisons (see Appendix).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows standard experimental practice for LLM post-training, uses public or cited datasets and models, preserves submission anonymity, and does not involve human-subject experiments or unsafe data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper discusses positive impact through improved training efficiency and hardware utilization for LLM post-training, and the work is methodological and systems-focused rather than a deployment of a new end-user model or application. No new application-specific societal risks are introduced beyond those of the underlying LLM post-training setting.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped dataset, or other asset requiring special safeguards for responsible release. The released asset is a training-system implementation.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original creators and sources of the existing assets used in this work. Table 5 lists the source URL, license name, and version for each reused asset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces and releases AsyncOPD, a fully asynchronous OPD training pipeline. The implementation, estimator choices, scheduler behavior, and experimental setup are documented in the main text and appendices. Instructions on how to reproduce the results are provided in the code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as an important, original, or non-standard component for developing the research methodology or manuscript. LLMs were used only for writing, or editing (declaration is not required under the checklist policy).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.