NeurIPS 2026

40th Conference on Neural Information Processing Systems

f-OPD: Stabilizing Long-Horizon On-Policy Distillation with Freshness-Aware Control

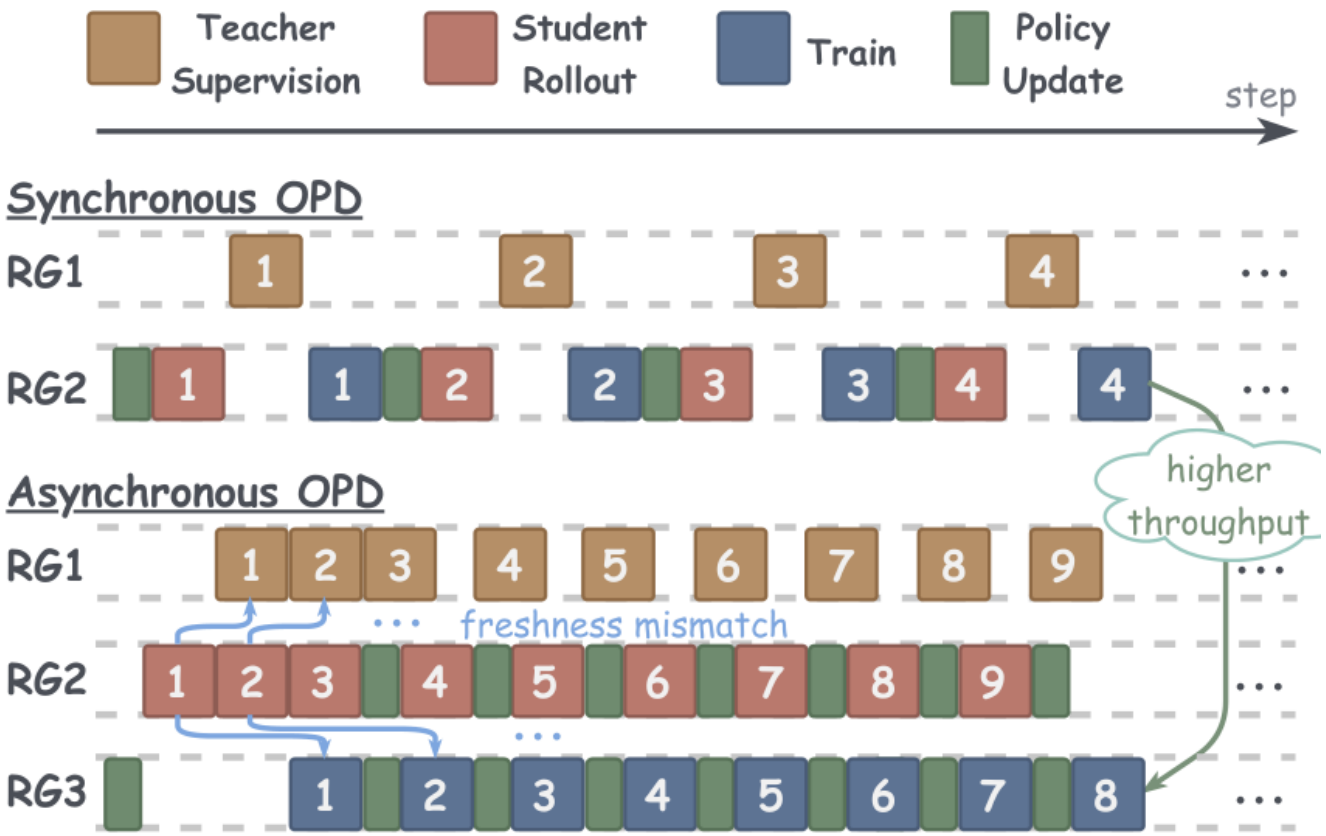
Xianwei Chen, Shimin Zhang, Jibin Wu

The Hong Kong Polytechnic University

jibin.wu@polyu.edu.hk

The Performance–Efficiency Tension

Asynchronous execution overlaps rollout, teacher supervision, and optimization.



Synchronous OPD

Fresh trajectories and targets
Idle time between pipeline stages

Asynchronous OPD

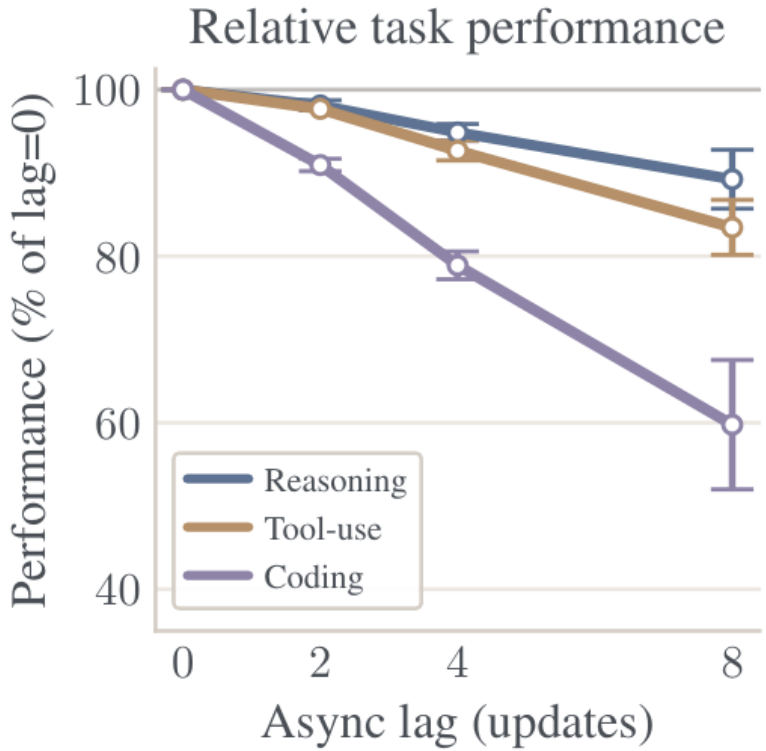
Higher system utilization
Buffered samples become stale

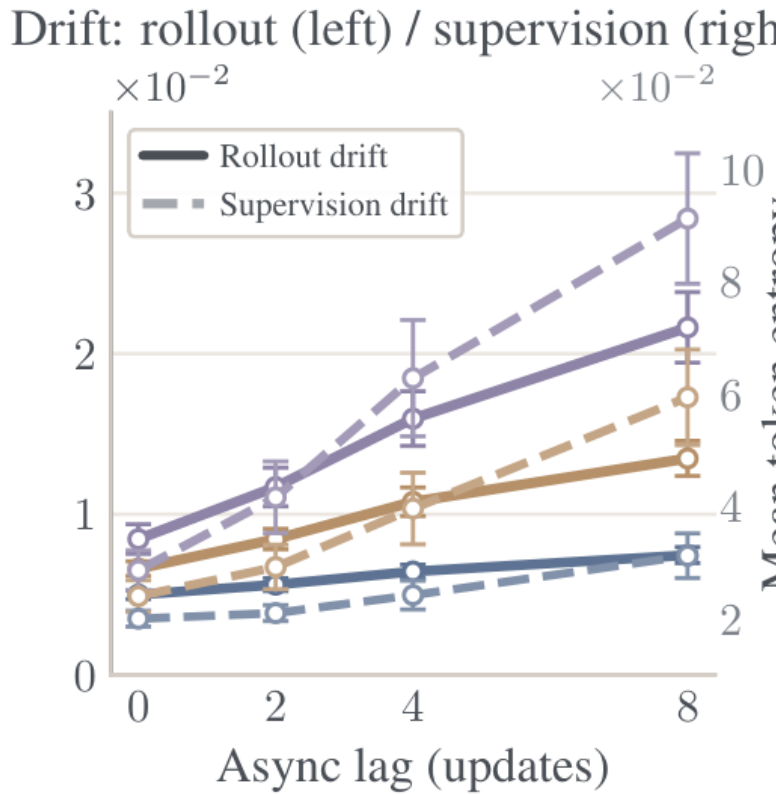
How can OPD retain asynchronous speed while preserving on-policy fidelity?

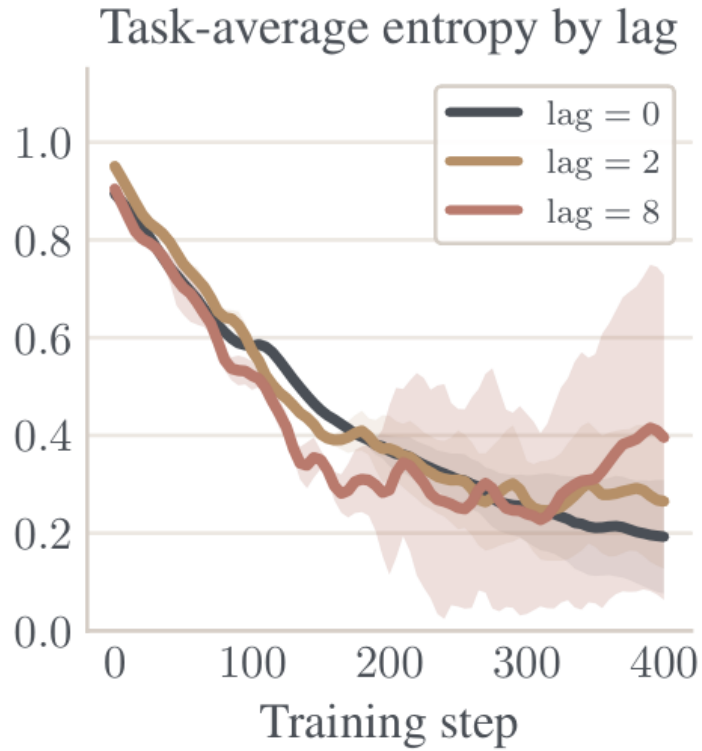
Longer Horizons Amplify Staleness

Controlled update lags of 0, 2, 4, and 8 expose the failure modes of vanilla OPD.

Coding performance falls by roughly 40% at lag 8 relative to lag 0.







Long-horizon agents amplify both objective mismatch and training instability.

Two Sources of Objective Discrepancy

The analysis separates stale student occupancy from stale teacher conditioning.

$$\Delta_t \le C_{\text{roll}}\,\mathrm{TV}(d_t, d_{\text{stale},t}) + C_{\text{sup}}\,\mathbb{E}_{i\sim B^t}\left[\mathrm{TV}\left(q(\cdot \mid c_i^t), q(\cdot \mid c_i^{r(i)})\right)\right]$$

Rollout drift

Older student policies produce a different distribution of trajectory prefixes.
Early deviations propagate downstream.

Supervision drift

Cached teacher labels condition on an outdated context. Tool outputs and repository states can change that context.

Finite support, positive probabilities, and a TV-Lipschitz loss are assumptions of the bound.

A fixed teacher can still produce stale supervision when its context changes.

A Freshness Score for Each Buffered Sample

A continuous score combines update age with both replay-based drift diagnostics.

$$f_i = \frac{1}{\tau_i + 1} \exp\left[-\alpha \sqrt{D_i^{\text{roll}}} - \beta \sqrt{D_i^{\text{sup}}}\right]$$

Age factor

$$\frac{1}{\tau_i + 1}$$

Discounts older trajectories.

Drift penalty

Larger rollout or supervision KL lowers the score smoothly.

Operational scale

$\alpha,\ \beta$

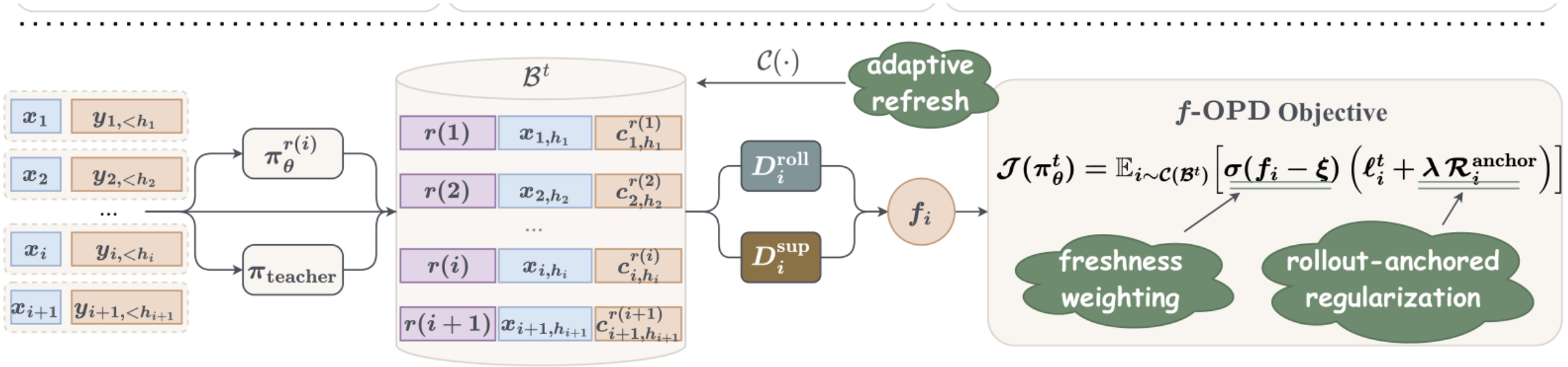
Balance the two drift signals.

The score is a monotone control surrogate, not an exact estimator of objective bias.

Freshness ranks sample reliability and determines how strongly each sample influences training.

f-OPD: Freshness-Aware Control

The score drives complementary interventions at the sample, update, and buffer levels.



Weighting

Reduce the influence of stale samples.

Rollout anchoring

Penalize divergence from the generating policy.

Adaptive refresh

Renew the buffer when freshness or alignment falls.

Weight surviving samples, constrain drift, and refresh the buffer when its support becomes unreliable.

Coding Agents: Quality, Throughput, and Stability

Qwen3-8B on the 250-issue SWE-bench Verified report subset, with five random seeds.

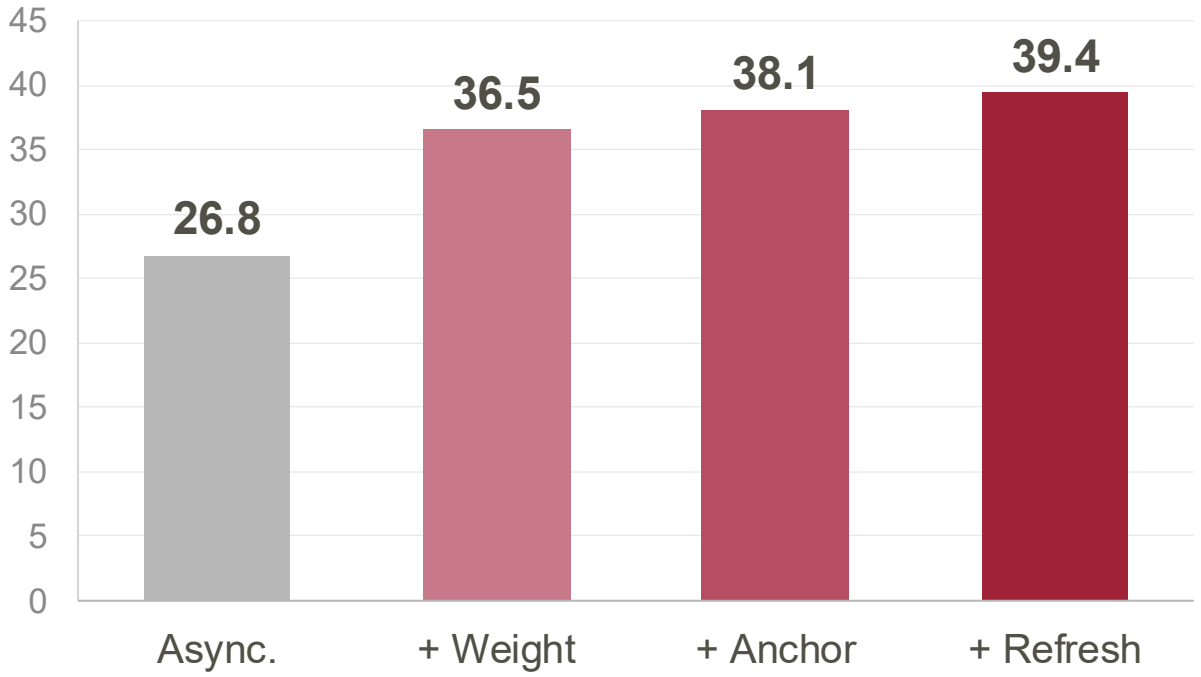
Method	Throughput	Resolve (%)	Post-patch reg. (pp)	Collapse
Sync. OPD	1.00×	41.8	1.6	0/5
Async. OPD	1.61×	26.8	7.4	3/5
+ KL	1.60×	28.2	7.6	3/5
+ clipped IW	1.60×	29.9	7.2	2/5
+ hard refresh	1.15×	35.1	6.6	1/5
+ lag-only weighting	1.54×	34.9	7.6	2/5
f-OPD	**1.46×**	**39.4**	**2.6**	**0/5**

f-OPD gains 12.6 resolve points over Async. OPD, with no observed collapse in 5 runs.

Each Freshness Mechanism Contributes

Incrementally adding controls recovers task quality and reduces collapse.

Coding resolve (%)



Sync. reference: 41.8 coding resolve and 74.8 tool-use accuracy.

Cumulative method	Tool-use Avg@4 (%)	Collapse
Async. OPD	61.6	3/5
+ Weighting	70.8	1/5
+ Anchoring	73.2	1/5
+ Refresh (f-OPD)	**74.4**	**0/5**

Freshness weighting recovers most of the lost quality. Anchoring and refresh add consistent gains.

Summary

f-OPD stabilizes asynchronous distillation by tracking the freshness of individual samples.

Two sources of mismatch

Stale student rollouts and stale teacher conditioning both affect the objective.

Three complementary controls

Freshness weighting, rollout anchoring, and adaptive buffer refresh.

Coding-agent result

39.4% resolve 1.46× throughput 0/5 observed collapses

Freshness-aware control improves the quality–throughput trade-off for long-horizon OPD.

Thank you