

ResOPD: Tail Residualization for Sparse On-Policy Distillation

Penghui Yang^{1,2}, Long Xing³, Xuanlang Dai^{2,4}, Ziyu Liu^{1,2}, Kai Chen² and Yuhang Zang^{†2}

¹Shanghai Jiao Tong University, ²Shanghai Artificial Intelligence Laboratory,

³University of Science and Technology of China, ⁴Fudan University

On-policy distillation (OPD) trains a student model on self-generated trajectories, but transmitting dense teacher distributions across long reasoning traces creates prohibitive communication and memory bottlenecks. Practical systems therefore rely on sparse teacher interfaces, typically transmitting either the sampled-token score or a small Top- k distribution. However, this sparse setting faces a fundamental dilemma: sampled-token estimators are unbiased but suffer from severe gradient variance, whereas directly optimizing Top- k objectives introduces systematic bias. To improve this trade-off, we propose ResOPD (**On-Policy Distillation with Tail Residualization**), which provides unbiased full-vocabulary reverse KL gradient estimation under on-policy sampling, with substantial variance reduction under the same sparse payload in the evaluated settings. ResOPD aggregates the unobserved vocabulary into an observable coarse tail event, computes its exact aggregate gradient, and samples only the fine-grained within-tail residual, which requires no additional teacher queries or forward passes. Extensive experiments demonstrate that ResOPD substantially reduces gradient variance, stabilizes online training dynamics, and improves downstream performance across the evaluated settings. These results establish ResOPD as an efficient, plug-and-play variance reduction primitive for sparse on-policy distillation.

Keywords: Large Language Models, On-Policy Distillation, Variance Reduction, Sparse Teacher Feedback

1. Introduction

On-policy distillation (OPD) [1, 9] transfers reasoning and generation ability from a strong teacher to a smaller student. Unlike offline distillation [17] or supervised fine-tuning, it trains the student on self-generated trajectories with token-level teacher supervision, which reduces exposure bias over long sequences. OPD is now a standard step in Large Language Model (LLM) post-training at frontier labs, including Thinking Machines [20], GLM [29], and DeepSeek [27].

Standard OPD [1] needs the teacher’s full distribution at every generated token, which is prohibitively costly for modern LLMs whose vocabularies can reach 250K tokens, so the payload grows with vocabulary size per token and with rollout length per sequence. Practical training systems therefore expose a sparse teacher interface that, at each step, returns only the teacher’s Top- k tokens with their probabilities, along with the probability of the sampled token [12, 30]. We refer to OPD trained through such an interface as **sparse OPD**. The difficulty is that the omitted tail is never observed: objectives defined directly on the Top- k support, whether by truncation or renormalization [32] or by aggregating the omitted mass into a ghost token [3, 12], are surrogates of the full-vocabulary reverse Kullback-Leibler (rKL) objective and carry a systematic gradient bias (Fig. 1(b)).

Bias can be traded for variance: rather than shrinking the objective to the observed support, a second family of methods retains the full-vocabulary rKL objective and estimates its gradient from the sparse payload, remaining **unbiased but noisy**. However, existing estimators fail to fully exploit this payload. At one extreme, sampled-token OPD [9, 20] queries only the rollout token’s score, suffering from severe gradient variance due to single-sample Monte Carlo estimation (Fig. 1(b)). At the other, methods with Top- k access underutilize it:

[†] Corresponding author: Yuhang Zang (zangyuhang@pjlab.org.cn).
Code is available at <https://github.com/InternLM/ResOPD>.

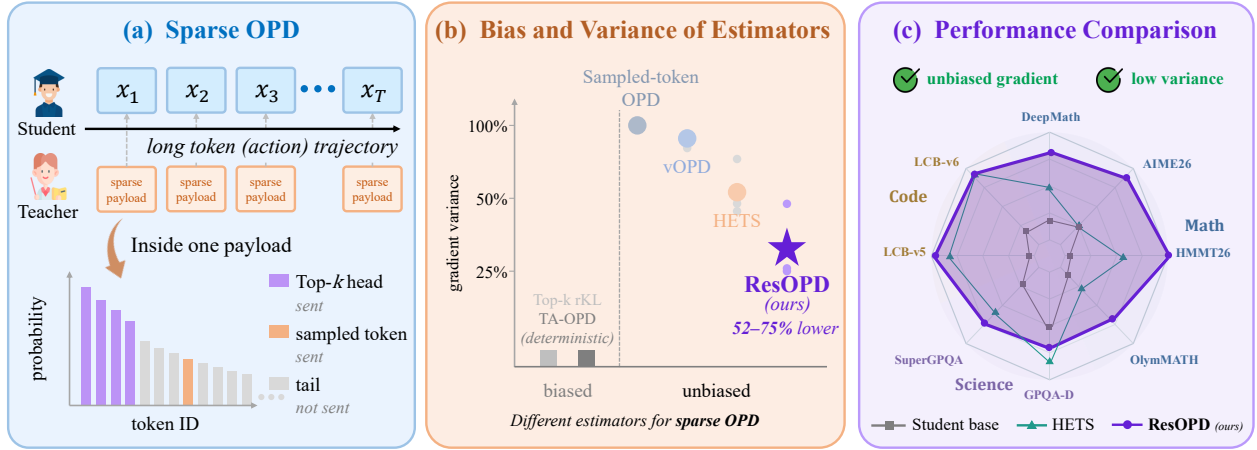


Figure 1: (a) Sparse OPD: the teacher model only sends the Top- k head probabilities and the sampled token score, leaving the tail distribution unobserved. (b) Bias vs. variance trade-off: while deterministic Top- k surrogates suffer from systematic gradient bias, ResOPD provides an unbiased gradient estimator and substantially reduces gradient variance. (c) Downstream evaluation: the radar chart shows benchmark scores across mathematics, coding, and science; while the relative ranking varies by task, ResOPD achieves the best overall performance.

scalar-baseline variants such as vOPD [23] reduce the head to a detached baseline rather than evaluating its token-wise gradient analytically (Tab. 2), while hybrid schemes such as Zhang and Ba [30], which we refer to as Head-Exact Tail-Sampled (HETS), evaluate the head exactly but treat the tail correction as entirely unknown despite knowing its aggregate mass. Consequently, each tail hit must compensate for the entire omitted mass, injecting large gradient spikes; even at 98.4% head coverage, a 30K-token reasoning trace contains 460–470 such spikes, thousands per optimizer step (Sec. 4.3.1). Practitioners are thus left choosing between a biased objective and an unstable gradient, which causes sampled-token OPD and HETS to peak early and degrade later in training (Sec. 4.3).

To resolve this trade-off, we analyze sparse OPD at the level of the gradient estimator. We first show that any estimator evaluating the head exactly, adding a prefix-dependent control vector weighted by the student’s tail mass, and correcting on tail hits is unbiased for the full-vocabulary rKL gradient, with HETS as the special case of a zero control (Sec. 3.2). Our covariance-trace decomposition reveals that variance splits into an invariant within-tail sampling term and a between-event term proportional to the control’s distance from the true tail mean. This explains HETS’s gradient spikes: its zero control pays for the full tail mean on every hit, whereas any control that better predicts this mean strictly reduces variance. Crucially, the residual masses outside the head define a coarse pseudo-token [3, 12] whose analytical gradient serves as a computable tail-mean predictor. Building on this, we propose *On-Policy Distillation with Tail Residualization* (ResOPD): at every step it computes the exact gradient of the $(k + 1)$ -way coarse distribution, adding only a small residual correction on tail hits (Fig. 2(a)). The resulting estimator is provably unbiased while retaining the same sparse teacher payload.

We evaluate ResOPD in two settings. In exact frozen-prefix audits that enumerate every next-token action, ResOPD lowers the gradient covariance trace by 52–75% relative to sampled-token OPD under Top-4 feedback, below both HETS and a payload-matched vOPD baseline (Tab. 2). In end-to-end distillation from Qwen3.5-27B to Qwen3.5-4B, it attains the best average score among the evaluated sparse OPD baselines on four mathematical reasoning benchmarks, 4.40 points above HETS under the same payload (Tab. 3), and its average advantage over HETS consistently holds across every tested support size $k \in \{1, 4, 16, 32\}$, with small margins (Tab. 6).

Our contributions are as follows: (1) **Bias and variance analysis of sparse OPD.** We show that truncated, renormalized, and stopped-coefficient Top- k rKL objectives are all biased surrogates of the full-vocabulary objective, and we derive a covariance-trace decomposition that separates within-tail sampling variance from between-event variance. This yields the exact condition under which a coarse tail control lowers variance relative to HETS, and a two-control family in which HETS, ResOPD, and the concurrent SC-TA [12] appear as special cases. (2) **ResOPD estimator.** We propose an estimator that splits the full-vocabulary rKL gradient into an exact $(k + 1)$ -way coarse term, computed at every step from the Top- k payload and the residual

tail masses, and a within-tail residual correction applied only when the rollout token lands in the tail. The estimator is unbiased under on-policy sampling and requires no additional teacher queries or forward passes. (3) **Empirical validation.** Using exact frozen-prefix gradient audits and end-to-end distillation settings, we show that ResOPD reduces gradient covariance, lowers online residual moments during training, and improves downstream accuracy over the evaluated sparse OPD baselines.

2. Related Work

On-Policy Language-Model Distillation. On-policy distillation (OPD) mitigates exposure bias by training students on self-generated rollouts under dense teacher supervision [1, 9]. Subsequent work explores skew divergences [18], adaptive targets [11, 15, 16], and diagnostic behaviors in complex reasoning regimes [8, 32]. However, full-vocabulary OPD incurs prohibitive communication and memory footprints during distributed rollout generation, severely hindering scalability.

Sparse Divergences and Truncation Biases. To alleviate communication bottlenecks, distributed training frequently adopts sparse Teacher Top- k payloads [32]. Nonetheless, naively renormalizing or truncating to the Top- k support optimizes a biased surrogate that misallocates tail probability mass. While coarse-support approaches like TA-OPD [12] incorporate a ghost token [3] to aggregate omitted mass, they still optimize a surrogate $(k + 1)$ -way divergence that discards the teacher’s fine-grained within-tail geometry.

Stochastic Estimators and Variance Reduction. An alternative paradigm constructs stochastic estimators to recover the true full-vocabulary rKL gradient unbiasedly. Standard sampled-token estimators [9, 20] suffer from extreme Monte Carlo variance, motivating variance-reduction baselines like vOPD [23]. HETS, a hybrid scheme proposed in EMA-PG [30], computes the Top- k head analytically and samples only within the tail, yet still incurs severe gradient bursts upon tail hits, which destabilizes training dynamics.

3. Methodology

In this section, we present the theoretical foundations and algorithmic formulation of *On-Policy Distillation with Tail Residualization* (ResOPD). We begin in Sec. 3.1 by establishing the full-vocabulary reverse-KL gradient and analyzing the trade-off of sparse teacher payloads, exposing how naive Top- k heuristics induce objective bias and why hybrid estimators like HETS suffer from tail-sampling variance. In Sec. 3.2, we formulate ResOPD (illustrated in Fig. 2), demonstrating how aggregate tail masses serve as an analytic control variate to achieve an unbiased, variance-reduced estimator. Sec. 3.3 describes the sampling assumptions and connects theoretical covariance bounds to observable online diagnostics and downstream task performance. Finally, Sec. 3.4 extends the analysis to a two-control family.

3.1. Preliminaries: From sampled gradients to an exact sparse head

At one fixed trajectory prefix, let $p(i) > 0$ and $q(i) > 0$ denote Student and Teacher next-token probabilities over a shared vocabulary $\mathcal{V}$. Define

$$r_i := \log \frac{p_i}{q_i}, \quad \mathbf{s}_i := \nabla_{\theta} \log p_i, \quad \mathbf{G}_i := r_i \mathbf{s}_i. \quad (1)$$

Because $\sum_i p_i \mathbf{s}_i = \mathbf{0}$, the full-vocabulary reverse-KL gradient is

$$\mathbf{g} = \sum_i p_i \mathbf{G}_i = \mathbb{E}_{A \sim p}[\mathbf{G}_A]. \quad (2)$$

Here, $\mathbf{G}_A$ denotes the stopped-coefficient sampled-token OPD gradient. Equation (2) demonstrates that this sampled-token gradient is an unbiased estimator of the full-vocabulary reverse-KL gradient.

Why direct Teacher Top- k rKL is biased. Let S be the fixed Teacher Top- k support, with $0 < k < |\mathcal{V}|$, $P = \sum_{i \in S} p_i$ and $Q = \sum_{i \in S} q_i$. Two common scalar shortcuts have different biases:

$$L_{\text{renormalized}} = D_{\text{KL}}\left(\frac{p_S}{P} \parallel \frac{q_S}{Q}\right), \quad L_{\text{truncated}} = \sum_{i \in S} p_i \log \frac{p_i}{q_i}. \quad (3)$$

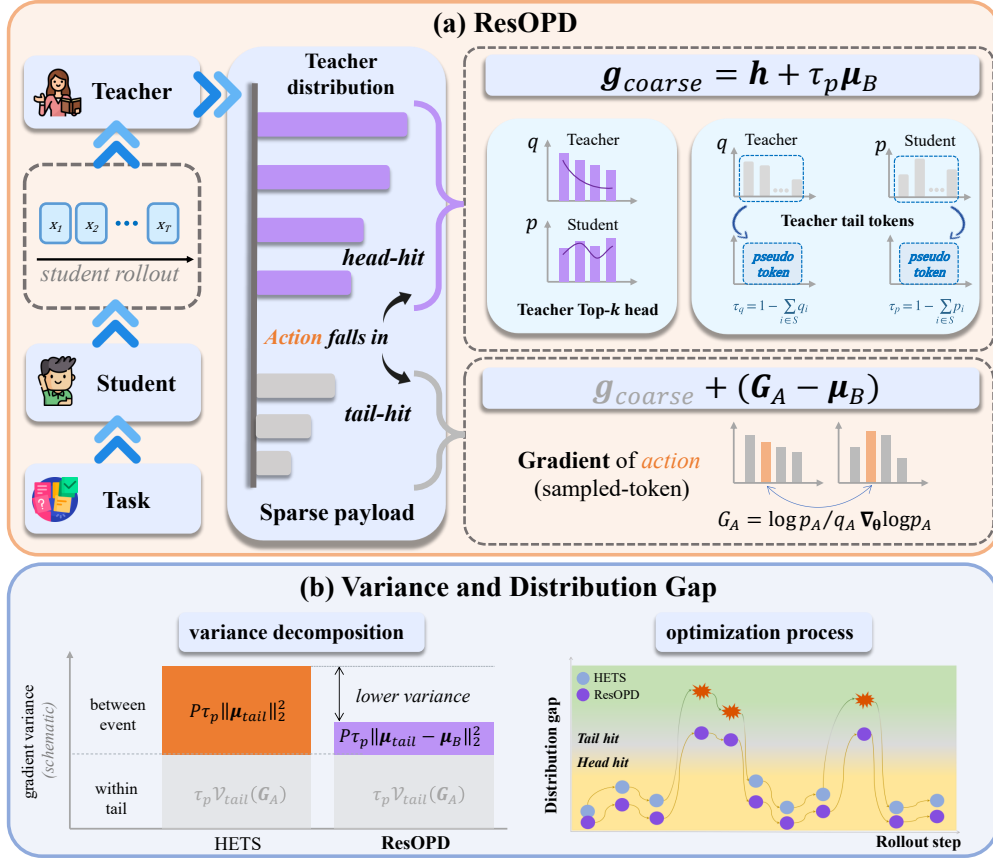


Figure 2: Overview of ResOPD. (a) When the student sampled-token falls in the head, ResOPD uses only the exact coarse gradient; upon a tail hit, it adds a residual correction. (b) ResOPD suppresses between-event variance compared to HETS by centering tail samples around μ_B , achieving superior alignment with the teacher distribution across rollout steps.

Renormalization depends only on the conditional distribution within S . And for every omitted logit z_j , $\partial L_{\text{renormalized}} / \partial z_j = 0$. Truncation retains the unnormalized head but introduces a boundary term:

$$\frac{\partial L_{\text{truncated}}}{\partial z_j} = -p_j \underbrace{\sum_{i \in S} p_i (r_i + 1)}_{C_S}, \quad j \notin S. \quad (4)$$

When $C_S > 0$, gradient descent increases every omitted logit. We adopt the stopped-coefficient alternative [32] $h = \sum_{i \in S} p_i r_i s_i$ in our experiments, which avoids that scalar boundary term but is still biased by the omitted tail. Appendix A.1 proves all three statements.

The interface returns full-normalized q_i for $i \in S$ and q_A for the sampled action. Write

$$\tau_p := 1 - P, \quad \tau_q := 1 - Q, \\ h := P \mathbb{E}[G_A | A \in S] = \sum_{i \in S} p_i G_i, \quad \mu_{\text{tail}} := \mathbb{E}[G_A | A \notin S] = \frac{1}{\tau_p} \sum_{i \notin S} p_i G_i. \quad (5)$$

The head h is exactly known, while the fine-tail mean gradient is unavailable. With $J = \mathbb{I}\{A \notin S\}$, the prior Head-Exact Tail-Sampled estimator is

$$\hat{g}^{\text{HETS}} = h + J G_A, \quad \mathbb{E}[\hat{g}^{\text{HETS}}] = h + \tau_p \mu_{\text{tail}} = g. \quad (6)$$

This HETS baseline [30] equals stopped Teacher Top-k rKL on head draws and adds the entire correction only

on tail draws:

$$\hat{g}^{\text{HETS}} = \begin{cases} \mathbf{h}, & A \in S, \\ \mathbf{h} + \mathbf{G}_A, & A \notin S, \end{cases} \quad \Pr(A \in S) = P. \quad (7)$$

Large P therefore creates a rare-correction regime: most tokens use the biased head-only update, while long minibatches still accumulate many sampled tail corrections.

3.2. Tail residualization

Top- k also reveals aggregate tail masses. Following the tail-bucketing approach in Hübottter et al. [13] and Huang and Wei [12], collapsing the tail into a single pseudo-token yields

$$b := \log \frac{\tau_p}{\tau_q}, \quad \bar{\mathbf{s}}_{\text{tail}} := \mathbb{E}[\mathbf{s}_A \mid J = 1] = \nabla_{\theta} \log \tau_p, \quad \boldsymbol{\mu}_B := b \bar{\mathbf{s}}_{\text{tail}}. \quad (8)$$

The exact gradient of the resulting complete coarse distribution is

$$\mathbf{g}_{\text{coarse}} = \mathbf{h} + \tau_p \boldsymbol{\mu}_B. \quad (9)$$

Here $\boldsymbol{\mu}_B$ predicts, but need not equal, the unavailable fine-tail mean. In fact, this formulation can be generalized to arbitrary baseline vectors: for $A \sim p$ and any prefix-dependent baseline vector $\mathbf{B}$,

$$\hat{g}(\mathbf{B}) = \mathbf{h} + \tau_p \mathbf{B} + J(\mathbf{G}_A - \mathbf{B}), \quad (10)$$

is unbiased, i.e., $\mathbb{E}_{A \sim p}[\hat{g}(\mathbf{B})] = \mathbf{g}$. Fig. 2 summarizes the construction. HETS is the special case $\mathbf{B} = \mathbf{0}$. The expectation is explicit:

$$\mathbb{E}[\hat{g}(\mathbf{B})] = \mathbf{h} + \tau_p \mathbf{B} + \tau_p(\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}) = \mathbf{g}. \quad (11)$$

The control may depend on the fixed prefix and current model probabilities, but must be fixed with respect to the sampled action and stopped before backward. Furthermore, defining the total variance of any random vector $\mathbf{X}$ as $\mathcal{V}(\mathbf{X}) := \text{tr}(\text{Cov}(\mathbf{X}))$, applying the law of total covariance yields the explicit variance of $\hat{g}(\mathbf{B})$:

$$\mathcal{V}(\hat{g}(\mathbf{B})) = \tau_p \mathcal{V}_{\text{tail}}(\mathbf{G}_A) + P \tau_p \|\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}\|_2^2, \quad (12)$$

where $\mathcal{V}_{\text{tail}}(\mathbf{G}_A) := \text{tr}(\text{Cov}(\mathbf{G}_A \mid A \in \text{tail}))$. Here, the first term $\tau_p \mathcal{V}_{\text{tail}}(\mathbf{G}_A)$ captures the baseline-invariant within-tail sampling variance (weighted by the tail probability τ_p), while the second term accounts for the between-event variance induced by the residual approximation error $\|\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}\|_2^2$, scaled by $P \tau_p$, the variance of the binary head-versus-tail indicator. To see this directly, condition on J : the head branch is deterministic, the tail branch retains $\text{Cov}_{\text{tail}}(\mathbf{G}_A)$, and the two conditional means differ by $\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}$. The law of total covariance then gives Equation (12).

Equation (12) also shows that the variance of the estimator $\hat{g}(\mathbf{B})$ is minimized when the control vector is set to the true tail mean, i.e., $\mathbf{B} = \boldsymbol{\mu}_{\text{tail}}$. However, since $\boldsymbol{\mu}_{\text{tail}}$ is unavailable in practice, ResOPD substitutes it with the computable tail-bucket surrogate $\boldsymbol{\mu}_B$:

$$\hat{g}^{\text{ResOPD}} = \mathbf{g}_{\text{coarse}} + J(\mathbf{G}_A - \boldsymbol{\mu}_B). \quad (13)$$

The predictable coarse correction now appears on every draw; only the fine-tail residual remains rare. ResOPD achieves strictly lower variance than HETS whenever $\|\boldsymbol{\mu}_{\text{tail}} - \boldsymbol{\mu}_B\|_2 < \|\boldsymbol{\mu}_{\text{tail}}\|_2$, i.e., whenever the tail-bucket vector $\boldsymbol{\mu}_B$ predicts the true tail mean better than zero, as observed in aggregate in our frozen-prefix audits. The complete derivation is given in Appendix A.2.

3.3. Sampling assumptions and evidence boundaries

The unbiasedness identities above assume $A \sim p_{\theta}$ at the fixed prefix. Rollout filters change this sampling law. Our primary mathematics experiments (Sec. 4.3) use temperature 1.0, rollout top- $p = 0.95$, and rollout top- $k = 20$, while the full-support audits (Sec. 4.2) and cross-domain runs (Sec. 4.4) use the ancestral condition. Sec. 4.1 specifies both, and Sec. 4.6 evaluates the effect of removing the filters.

The coarse-to-fine variance decomposition can also be characterized under an arbitrary realized rollout distribution $\tilde{p}$. Let $\tilde{\tau} = \Pr_{\tilde{p}}(A \notin S)$ and $\tilde{\mu}_{\text{tail}} = \mathbb{E}_{\tilde{p}}[G_A \mid A \notin S]$. For $\hat{g}(B) = h + \tau_p B + \mathbb{I}\{A \notin S\}(G_A - B)$, the estimator still admits the exact decomposition

$$\mathcal{V}_{\tilde{p}}(\hat{g}(B)) = \tilde{\tau} \mathcal{V}_{\tilde{p}, \text{tail}}(G_A) + (1 - \tilde{\tau}) \tilde{\tau} \|\tilde{\mu}_{\text{tail}} - B\|_2^2. \quad (14)$$

Hence, for $0 < \tilde{\tau} < 1$, ResOPD has strictly lower sampling variance than HETS whenever $\|\tilde{\mu}_{\text{tail}} - \mu_B\|_2 < \|\tilde{\mu}_{\text{tail}}\|_2$. The expected update under this realized sampling law is

$$\mathbb{E}_{\tilde{p}}[\hat{g}(B)] = h + \tau_p B + \tilde{\tau}(\tilde{\mu}_{\text{tail}} - B). \quad (15)$$

The variance identity remains exact under $\tilde{p}$, but it does not turn a truncated rollout into an unbiased draw from p_θ . Crucially, truncation is confined strictly to sequence generation. All student quantities (p_i, r_i, s_i, τ_p , and the coarse-tail terms) are faithfully recomputed via a fresh, untruncated forward pass during training.

Equation (12) characterizes estimator quality at fixed prefixes. We test it by exact action enumeration (Sec. 4.2). During online OPD, both policies and visited prefixes change, while sparse Teacher feedback hides the fine tail. We instead monitor residual second moments M_2 as cheap mechanism-aligned online diagnostics. Their decline is consistent with reduced head-versus-tail event noise, but does not by itself estimate full gradient covariance. Appendix D.1 gives the exact definitions.

Nevertheless, variance reduction does not imply monotonic downstream accuracy gains. Lower covariance trace reduces stochastic optimization noise, but final performance is still shaped by training dynamics and decoding randomness, both of which are discretized by benchmark scoring. Thus, ResOPD can reliably suppress gradient noise without strictly outperforming HETS at every checkpoint. We therefore view theoretical covariance, online residual moments, and downstream accuracy as complementary signals across different stages of the learning pipeline.

3.4. A Unified View of Sparse Gradient Estimators

ResOPD changes the between-event component of the HETS variance while preserving its conditional within-tail randomness. This distinction provides a useful way to compare sparse unbiased estimators: a second, conditionally zero-mean control can act inside the tail. We present the family, its oracle optimum, and the exact conditions governing the ResOPD–SC-TA comparison here. The corresponding audits and downstream results appear in Sec. 4.8.

3.4.1. Separating between-event and within-tail controls

The single-control family acts only on the head-versus-tail event component. A second control can reduce the conditional variance that remains inside the tail without changing the sparse Teacher payload. Define the conditionally zero-mean score fluctuation $C_A = b(s_A - \bar{s}_{\text{tail}})$, so that $\mathbb{E}[C_A \mid J = 1] = 0$. For any finite prefix-measurable B and scalar γ , both stopped before drawing A , define

$$\hat{g}(B, \gamma) = h + \tau_p B + J(G_A - B - \gamma C_A). \quad (16)$$

Equation (16) satisfies

$$\begin{aligned} \mathbb{E}[\hat{g}(B, \gamma)] &= h + \tau_p B + \tau_p(\mu_{\text{tail}} - B) - \gamma \tau_p \mathbb{E}[C_A \mid J = 1] \\ &= h + \tau_p \mu_{\text{tail}} = g. \end{aligned} \quad (17)$$

Conditioned on the head event its value is $h + \tau_p B$, while conditioned on the tail event its mean is $h + \tau_p B + \mu_{\text{tail}} - B$. Their difference is $\mu_{\text{tail}} - B$, while only the tail branch retains conditional randomness. The law of total covariance therefore gives

$$\mathcal{V}(\hat{g}(B, \gamma)) = \tau_p \mathcal{V}_{\text{tail}}(G_A - \gamma C_A) + P \tau_p \|\mu_{\text{tail}} - B\|_2^2. \quad (18)$$

This factorization proves that the two controls are orthogonal in trace covariance: B changes only the between-event term and γ only the within-tail term. We call Equation (16) the *Unified* family: the first coordinate

Table 1: Coordinates in the two-control family.

Unbiased estimator	$\mathbf{B}$	γ	Role
HETS	$\mathbf{0}$	0	structural baseline
ResOPD	$\boldsymbol{\mu}_B$	0	our estimator
SC-TA	$\boldsymbol{\mu}_B$	1	boundary estimator
Joint oracle	$\boldsymbol{\mu}_{\text{tail}}$	γ^*	unavailable optimum

generalizes the single control in Equation (10), while the second subtracts a known-zero-mean within-tail fluctuation. If $\mathcal{V}_{\text{tail}}(\mathbf{C}_A) > 0$, their joint oracle is

$$\mathbf{B}^* = \boldsymbol{\mu}_{\text{tail}}, \quad \gamma^* = \frac{\text{tr}(\text{Cov}_{\text{tail}}(\mathbf{G}_A, \mathbf{C}_A))}{\mathcal{V}_{\text{tail}}(\mathbf{C}_A)}, \quad (19)$$

and completing the square in both coordinates yields

$$\begin{aligned} & \mathcal{V}(\hat{\mathbf{g}}(\mathbf{B}, \gamma)) - \mathcal{V}(\hat{\mathbf{g}}(\mathbf{B}^*, \gamma^*)) \\ &= P\tau_p \|\mathbf{B} - \boldsymbol{\mu}_{\text{tail}}\|_2^2 + \tau_p \mathcal{V}_{\text{tail}}(\mathbf{C}_A)(\gamma - \gamma^*)^2. \end{aligned} \quad (20)$$

If $\mathcal{V}_{\text{tail}}(\mathbf{C}_A) = 0$, γ is immaterial and the first term remains exact. Importantly, neither oracle coordinate is available online: $\mathbf{B}^*$ requires the full Teacher tail, while the per-prefix γ^* requires its conditional covariance. ResOPD instead makes the simple but effective choice $(\boldsymbol{\mu}_B, 0)$, and the global fitted coefficient is a separate offline diagnostic.

3.4.2. ResOPD and SC-TA as endpoints

We now specialize the two-level family to $\mathbf{B} = \boldsymbol{\mu}_B$ and connect it to sample-corrected TA-OPD (SC-TA) from the concurrent work of Huang and Wei [12]. Their primary method, TA-OPD, augments the Teacher Top- k support with a tail pseudo-token and optimizes the resulting coarse $(k+1)$ -way reverse-KL objective. SC-TA is subsequently derived as a sample-corrected variant that restores full-vocabulary unbiasedness. We complement this construction with a gradient-level characterization within our two-control family. In particular, we derive the exact pointwise relation between ResOPD and SC-TA, characterize their covariance ordering as a function of the within-tail control coefficient γ , identify the trace-optimal γ^* , and evaluate these predictions across natural prompt distributions.

Exact fine-tail decomposition. On a tail draw,

$$r_A = \log \frac{p_A}{q_A} = \underbrace{\log \frac{\tau_p}{\tau_q}}_b + \log \frac{p_A/\tau_p}{q_A/\tau_q}. \quad (21)$$

Using $\boldsymbol{\mu}_B = b\bar{\mathbf{s}}_{\text{tail}}$ therefore gives

$$\mathbf{G}_A = \underbrace{\boldsymbol{\mu}_B + b(\mathbf{s}_A - \bar{\mathbf{s}}_{\text{tail}})}_{\mathbf{C}_A} + \underbrace{(r_A - b)\mathbf{s}_A}_{\mathbf{R}_A}, \quad \mathbb{E}[\mathbf{C}_A \mid J=1] = \mathbf{0}. \quad (22)$$

The terms have different roles. $\boldsymbol{\mu}_B$ is the coarse tail-bucket prediction; $\mathbf{C}_A$ is a known-zero-mean fluctuation around the conditional Student score; and $\mathbf{R}_A$ records the unresolved Student–Teacher conditional tail-shape mismatch.

With $\mathbf{g}_{\text{coarse}} = \mathbf{h} + \tau_p \boldsymbol{\mu}_B$, the specialization of Equation (16) is

$$\hat{\mathbf{g}}_\gamma = \mathbf{g}_{\text{coarse}} + J[\mathbf{R}_A + (1 - \gamma)\mathbf{C}_A]. \quad (23)$$

Equation (17) already proves that every fixed γ targets $\mathbf{g}$. Its two named endpoints are

$$\begin{aligned} \hat{\mathbf{g}}^{\text{ResOPD}} &= \hat{\mathbf{g}}_{\gamma=0} = \mathbf{g}_{\text{coarse}} + J(\mathbf{R}_A + \mathbf{C}_A), \\ \hat{\mathbf{g}}^{\text{SC}} &= \hat{\mathbf{g}}_{\gamma=1} = \mathbf{g}_{\text{coarse}} + J\mathbf{R}_A. \end{aligned} \quad (24)$$

Hence their signed pointwise relationship is

$$\widehat{\mathbf{g}}^{\text{SC}} = \widehat{\mathbf{g}}^{\text{ResOPD}} - J\mathbf{C}_A \quad (25)$$

The difference has zero expectation but is generally nonzero on tail actions.

Variance optimum and endpoint certificate. Let $\mathbf{Z}_0 = \mathbf{R}_A + \mathbf{C}_A = \mathbf{G}_A - \boldsymbol{\mu}_B$. Decomposing the total variance into the common between-event component and the γ -dependent within-tail component yields

$$\begin{aligned} \mathcal{V}(\widehat{\mathbf{g}}_\gamma) &= \tau_p \mathcal{V}_{\text{tail}}(\mathbf{Z}_0 - \gamma \mathbf{C}_A) + P_{\tau_p} \|\boldsymbol{\mu}_{\text{tail}} - \boldsymbol{\mu}_B\|_2^2, \\ \mathcal{V}_{\text{tail}}(\mathbf{Z}_0 - \gamma \mathbf{C}_A) &= \mathcal{V}_{\text{tail}}(\mathbf{Z}_0) - 2\gamma \text{tr}(\text{Cov}_{\text{tail}}(\mathbf{Z}_0, \mathbf{C}_A)) + \gamma^2 \mathcal{V}_{\text{tail}}(\mathbf{C}_A). \end{aligned} \quad (26)$$

When $\mathcal{V}_{\text{tail}}(\mathbf{C}_A) > 0$, completing the square with respect to γ identifies the trace optimum:

$$\begin{aligned} \gamma^* &= \frac{\text{tr}(\text{Cov}_{\text{tail}}(\mathbf{Z}_0, \mathbf{C}_A))}{\mathcal{V}_{\text{tail}}(\mathbf{C}_A)}, \\ \mathcal{V}(\widehat{\mathbf{g}}_\gamma) - \mathcal{V}(\widehat{\mathbf{g}}_{\gamma^*}) &= \tau_p \mathcal{V}_{\text{tail}}(\mathbf{C}_A) (\gamma - \gamma^*)^2. \end{aligned} \quad (27)$$

Evaluating the quadratic regret at the boundary points $\gamma = 1$ (SC-TA) and $\gamma = 0$ (ResOPD) reveals that SC-TA achieves strictly lower trace covariance than ResOPD if and only if $\gamma^* > 1/2$. Substituting $\mathbf{Z}_0 = \mathbf{R}_A + \mathbf{C}_A$ expands the optimal coefficient as

$$\gamma^* = 1 + \frac{\text{tr}(\text{Cov}_{\text{tail}}(\mathbf{R}_A, \mathbf{C}_A))}{\mathcal{V}_{\text{tail}}(\mathbf{C}_A)}. \quad (28)$$

Consequently, SC-TA outperforms ResOPD exactly when $\text{tr}(\text{Cov}_{\text{tail}}(\mathbf{R}_A, \mathbf{C}_A)) > -\mathcal{V}_{\text{tail}}(\mathbf{C}_A)/2$. By the Cauchy–Schwarz inequality, a sufficient condition is

$$\mathcal{V}_{\text{tail}}(\mathbf{R}_A) < \mathcal{V}_{\text{tail}}(\mathbf{C}_A)/4. \quad (29)$$

In particular, when the conditional tail shapes match, the residual vanishes ($\mathbf{R}_A = \mathbf{0}$), yielding $\gamma^* = 1$. These conditions serve as distribution-dependent certificates rather than guaranteeing a universal variance ordering.

4. Experiments

We evaluate three complementary questions: whether residualization reduces exact conditional gradient covariance, whether its observable effects persist during online training, and whether the resulting policies improve downstream performance. We first specify the protocols, then present frozen-prefix audits, mathematics results and diagnostics, cross-domain transfer, and controlled ablations of support size, rollout sampling, learning rate, and the within-tail control.

4.1. Experimental Setup

Offline mechanism audit. We freeze Qwen3.5 [24] models and sample one 64-token trajectory per prompt at selected frozen Math prefixes, measuring covariance trace and Noise-to-Signal Ratio: $\text{NSR} = \frac{\mathcal{V}(\widehat{\mathbf{g}})}{\|\mathbf{g}_{\text{full-rkl}}\|_2^2}$. Experiments cover five settings: GSM8K [4] (32 prompts) across 0.8B/4B, 4B/27B and 9B/27B pairs; all 33 HMMT26 prompts [5] (4B/27B); and 64 prompts sampled from DeepMath-103K [10] (4B/27B).

Online OPD training for mathematical reasoning. We distill the thinking-enabled Qwen3.5-4B student from the Qwen3.5-27B teacher. Starting from DeepMath-103K, we filter for challenging competition problems after deduplication and benchmark contamination removal. To ensure an appropriate teacher-student capability gap, we retain prompts where the teacher succeeds twice while the student fails at least once across sampled rollouts. This yields 1,276 unique problems, partitioned into 1,024 for training and 252 for conditional held-out evaluation. Training runs for 3 epochs (192 steps), saving intermediate checkpoints at steps 6, 32, 64, 128, and 192.

Table 2: Exact frozen-prefix covariance trace and NSR. Covariance traces of student logit-gradients are computed via full action enumeration and averaged across rollouts; NSR normalizes each trace by the squared L_2 norm of the full-rKL gradient. Sparse estimators (vOPD_{T-Top- k} , HETS, ResOPD) share an identical teacher budget of $k = 4$, with vOPD_{T-Top- k} adopting a detached conditional-KL baseline over this support. Best results per row are in **bold**, and second-best are underlined.

Dataset	Models	Covariance Trace ↓				NSR ↓			
		Sampled	vOPD _{T-Top-k}	HETS	ResOPD	Sampled	vOPD _{T-Top-k}	HETS	ResOPD
GSM8K	0.8B / 4B	2.146	1.730	<u>1.559</u>	1.018	5.422	4.370	<u>3.940</u>	2.573
GSM8K	9B / 27B	0.215	0.192	<u>0.107</u>	0.068	6.202	5.560	<u>3.101</u>	1.955
GSM8K	4B / 27B	0.333	0.297	<u>0.183</u>	0.108	6.058	5.400	<u>3.335</u>	1.964
HMMT26	4B / 27B	0.208	0.189	<u>0.092</u>	0.052	6.589	5.977	<u>2.914</u>	1.630
DeepMath	4B / 27B	0.183	0.169	<u>0.087</u>	0.047	7.352	6.785	<u>3.500</u>	1.907

Evaluation. Our evaluation spans four benchmarks: the DeepMath held-out set, AIME26 (30 problems) [31], HMMT26 (33 problems) [5], and the English split of OlymMATH (200 problems) [26]. Given the limited problem counts in AIME26 and HMMT26, we report the average over 16 sampled responses per problem (avg@16) to mitigate evaluation variance, while DeepMath held-out and OlymMATH use avg@2. In addition to terminal performance, we evaluate all intermediate saved checkpoints to compute the Area Under the Curve (AUC), quantifying learning dynamics and training efficiency. We compute normalized trapezoidal AUC over the saved checkpoint trajectory, and average the benchmark-wise scores and AUCs for the aggregate columns. Full configurations for both training and evaluation are collected in Appendix B.

Realized rollout distributions. Following the Qwen3.5 official recipe, the primary mathematics rollouts and evaluations use temperature 1.0, top- $p = 0.95$, and top- $k = 20$. Truncating low-probability tokens helps avoid degenerate prefixes that can destabilize OPD and Teacher feedback [8]. The Student probabilities used in training are recomputed by a fresh, untruncated forward pass. As explained in Sec. 3.3, this setup retains a well-defined sampling-variance comparison but does not satisfy the idealized full-rKL unbiasedness assumption. We therefore also report a matched full-support mathematics ablation (Sec. 4.6) and cross-domain training with top- $p = 1.0$, top- $k = -1$ (Sec. 4.4).

4.2. Exact frozen-prefix covariance audits

Tab. 2 reports the exact covariance trace and NSR across five frozen-prefix panels. By analytically enumerating every next-token action at each Student-visited prefix, we isolate the conditional sampling variance of each estimator without decoding noise or optimizer feedback. Under the panel-level aggregation used in the table, both metrics exhibit a consistent ordering across every panel: ResOPD < HETS < vOPD < sampled-token OPD (lower is better). Relative to sampled-token OPD, the aggregate covariance trace is reduced by 7.7%–19.4% for payload-matched Teacher-Top- k vOPD, 27.3%–55.8% for HETS, and 52.5%–75.0% for ResOPD.

While Teacher-Top- k vOPD serves as a competitive baseline in our end-to-end experiments (Tab. 3), its detached scalar control fails to exploit the structural head–tail decomposition, leaving the entire action vector sampled at every step. Consequently, its variance reduction remains modest compared to HETS and ResOPD. Because HETS and ResOPD share the identical payload budget, exact Top-4 head integration, and conditional target structure, their contrast directly isolates the gain from coarse-tail residualization without baseline confounds. We therefore focus our primary ablations on the HETS-to-ResOPD progression.

As the support size k increases, the exact Top- k head absorbs more probability mass, causing the residual tail opportunity P_{τ_p} to contract. Accordingly, the marginal covariance reduction of ResOPD over HETS gradually narrows. The full support-size sweep in Appendix D.3 tests this communication–variance trade-off directly.

4.3. End-to-end mathematical reasoning

Tab. 3 reports the primary distillation results: ResOPD achieves the strongest aggregate performance among all evaluated OPD baselines, attaining both the highest terminal average score (50.12%) and learning-trajectory

Table 3: Qwen3.5-27B-to-4B OPD results. DeepMath held-out and OlymMATH are evaluated using avg@2, while AIME26 and HMMT26 adopt avg@16 due to their smaller problem sets.

Method	Benchmark Accuracy (%)				Aggregate (%)	
	DeepMath held-out (avg@2)	AIME26 (avg@16)	HMMT26 (avg@16)	OlymMATH (avg@2)	Avg. Score	Avg. AUC
▼ Original Models						
Qwen3.5-4B Student	48.41	56.25	35.23	31.50	42.85	–
Qwen3.5-27B Teacher	89.29	80.62	62.88	63.25	74.01	–
▼ Alternative Sparse OPD (Biased)						
Teacher Top- k fKL	50.20	54.58	35.04	34.50	43.58	42.80
Teacher Top- k rKL	51.79	55.83	34.66	36.75	44.76	44.63
TA-OPD	58.13	58.96	<u>39.58</u>	36.00	48.17	<u>47.94</u>
▼ Unbiased rKL Estimators						
Sampled-token OPD	48.21	57.50	36.17	34.25	44.03	46.69
vOPD _{T-Top-k}	<u>58.93</u>	60.00	39.20	<u>37.25</u>	<u>48.85</u>	47.49
HETS	54.76	56.25	37.88	34.00	45.72	46.18
ResOPD (ours)	60.91	<u>59.38</u>	39.96	40.25	50.12	47.96

Table 4: Aggregated online diagnostics across training (lower is better). Values represent the average over all 192 optimization steps. Δ denotes the relative percentage reduction of ResOPD compared to HETS.

Diagnostic Metric	Top- k rKL	vOPD _{T-Top-k}	HETS	ResOPD (ours)	Δ (%)↓
Support-mass gap ($\times 10^{-3}$)	11.17	3.53	5.17	3.50	−32.3%
Tail log-mass mismatch	0.715	0.858	0.723	0.322	−55.5%
Token residual M_2 ($\times 10^{-3}$)	–	10.55	11.73	10.67	−9.0%
Trajectory residual M_2 ($\times 10^{-5}$)	–	2.76	3.50	2.65	−24.3%

AUC (47.96%). This favorable training trajectory also extends to external mathematical competitions. Across AIME26, HMMT26, and OlymMATH, competing baselines exhibit pronounced post-peak degradation during later stages of training, whereas ResOPD avoids severe late-stage degradation and recovers to its best aggregate performance at the terminal checkpoint. Step-wise evaluations over intermediate checkpoints are provided and analyzed in Appendix C.1.

4.3.1. Training dynamics

As illustrated in Tab. 4 and Fig. 3, ResOPD consistently suppresses both student–teacher mass discrepancies and stochastic residual fluctuations relative to HETS on average over 192 training steps. Across the matched HETS–ResOPD runs, ResOPD reduces the support-mass gap by 32.3% and the tail log-mass mismatch by 55.5%. Although Teacher-Top- k vOPD exhibits a comparable support-event M_2 , this scalar residual on self-generated trajectories omits centered log-ratio and score geometry, and exact gradient covariance comparisons on shared frozen prefixes (Tab. 2) confirm ResOPD’s structural advantage.

This provides a mechanistic explanation for the optimization behavior of ResOPD. Under HETS, the full fine-tail gradient correction is concentrated on rare tail-hit events. Despite approximately 98.4% head coverage, a 30K-token reasoning trace still contains 460–470 such corrections, accumulating to thousands per optimizer step. By subtracting the payload-available tail-bucket component μ_B , ResOPD redistributes the predictable between-event component while preserving the full-rKL expectation under the on-policy condition in Sec. 3.3. Downstream accuracy remains noisier because it depends on discrete answer correctness, but the completion-only Teacher NLL audit below shows consistently higher teacher likelihood of student-generated trajectories for ResOPD. Together with the matched endpoint and curve-AUC results, these diagnostics connect the estimator-level covariance reduction to a smoother optimization path and improved downstream performance.

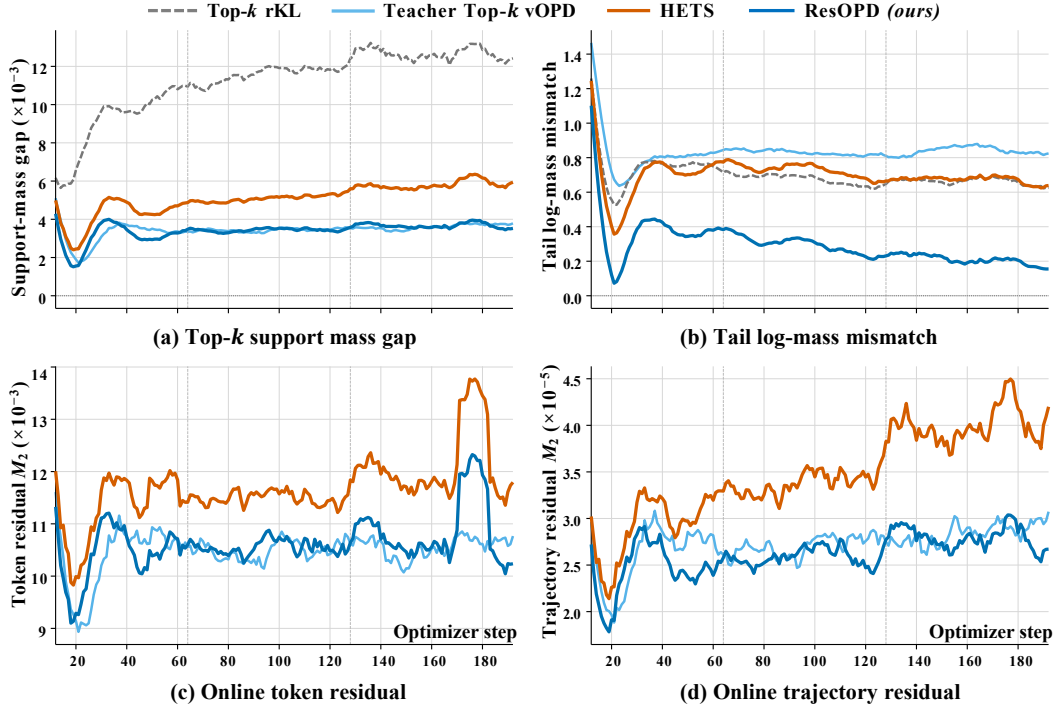


Figure 3: Training-time diagnostics (64 steps/epoch; lower is better). **Panels (a) and (b)** illustrate the student–teacher mismatch in Top- k support mass and tail log-mass across vOPD, HETS, ResOPD, and direct Teacher Top- k rKL. **Panels (c) and (d)** compare vOPD, HETS and ResOPD via the squared residual between the binary Top- k indicator and its student-predicted probability, evaluated at the token level and trajectory level, respectively. Detailed formulations are provided in Appendix D.1.

4.3.2. Behavioral fidelity

Beyond task-specific accuracy metrics (e.g., avg@2 and avg@16 on downstream math benchmarks), monitoring the student’s negative log-likelihood under the teacher model provides an intrinsic, task-agnostic readout of distillation fidelity. Specifically, a lower Teacher NLL indicates that the student samples receive higher likelihood under the teacher.

Fig. 4 tracks this metric across the matched Math run. As illustrated in Panel A, while both estimators begin around 0.46 nats/token, ResOPD achieves a steeper drop by step 6 and sustains a monotonic decline down to 0.36 nats/token by step 192. In contrast, HETS plateaus around 0.385 nats/token and displays slight drift in later stages. The difference $\Delta\text{Teacher-NLL}$ (ResOPD – HETS) in Panel B remains strictly negative across all optimizer steps, widening to over -0.027 nats/token. This demonstrates that ResOPD consistently preserves higher behavioral fidelity to the teacher than HETS, offering evidence of higher teacher-likeness, complementary to accuracy.

4.4. Cross-domain transfer to science and coding

To further evaluate the generalization and cross-domain effectiveness of ResOPD, we follow Heo et al. [11] and extend our investigation to science and coding tasks. Our experimental protocol largely mirrors that of Heo et al. [11]: we distill from a Qwen3-30B-A3B-Instruct [28] teacher into a Qwen3-4B student, training on a 30k-sample mixture compiled from OpenMathReasoning [21], OpenScienceReasoning-2 [22], and OpenCodeReasoning [2]. The run uses full-support ancestral rollouts with temperature 1.0, top- $p = 1.0$, and top- $k = -1$, without thinking mode. For evaluation, we benchmark the code domain on LiveCodeBench [LCB; 14] v5 (167 problems) and v6 (175 problems). For the science domain, we evaluate on GPQA-Diamond [GPQA-D, 198 problems; 25] and SuperGPQA [6], adopting the exact 650-problem subset used in Heo et al. [11].

Tab. 5 reports the evaluation results. Overall, ResOPD achieves the strongest aggregate results, leading in both terminal average score (44.15%) and trajectory AUC (42.65%). Although ResOPD and HETS nominally surpass the teacher on code tasks (e.g., ResOPD’s 48.90% vs. 45.51% on LCB-v5), paired bootstrap 95% CIs

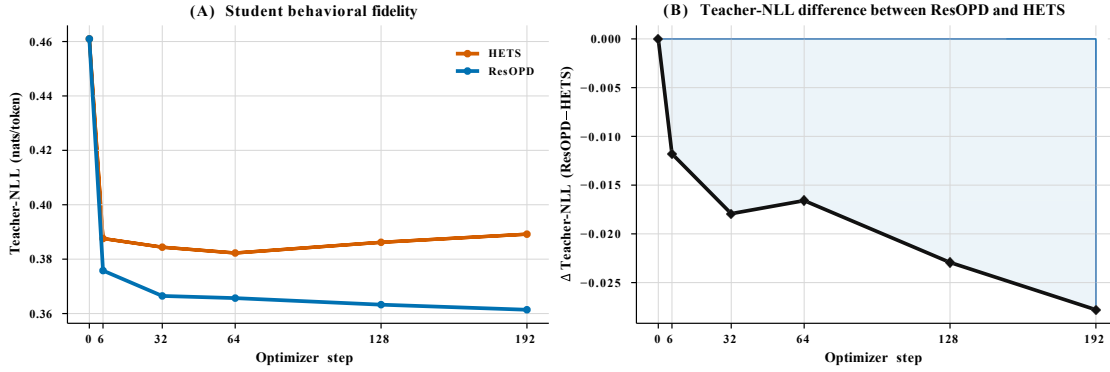


Figure 4: Teacher negative log-likelihood (NLL) across training in the Math OPD setup. Evaluations are computed over the four core benchmark families with identical weights to the main suite: AIME26, HMMT26, DeepMath held-out, and OlymMATH. Lower values indicate higher fidelity to the Teacher distribution.

Table 5: Distillation results from Qwen3-30B-A3B-Instruct to Qwen3-4B. Evaluations use avg@10 for GPQA-D and avg@3 for SuperGPQA, LCB-v5, and LCB-v6. AUC integrates performance across the fixed training trajectory (Base/10/25/50/100).

Method	Science Benchmarks (%)		Code Benchmarks (%)		Aggregate (%)	
	GPQA-D (avg@10)	SuperGPQA (avg@3)	LCB-v5 (avg@3)	LCB-v6 (avg@3)	Avg. Score	Avg. AUC
▼ Original Models						
Student	45.25	32.46	21.76	24.38	30.96	–
Teacher	56.06	48.87	45.51	41.71	48.04	–
▼ Unbiased rKL Estimators						
Sampled-token OPD	50.30	<u>36.87</u>	43.31	39.43	42.48	41.21
HETS	48.89	36.10	<u>46.31</u>	43.24	<u>43.63</u>	<u>42.13</u>
ResOPD (ours)	47.12	37.54	48.90	<u>43.05</u>	44.15	42.65

cross zero (e.g., ResOPD–Teacher spans $[-2.48, +5.33]$ on LCB-v6), indicating on-par performance rather than statistically significant super-teacher generalization. Across training scales, ResOPD maintains consistent optimization advantages (e.g., reducing trajectory residual M_2 by 9.5%–12.6% and gradient norms by 14.6%–23.7%) and exhibits directional AUC gains over HETS on code (+0.48 and +0.73 points), even though Sampled-token OPD yields the strongest performance on science tasks (detailed training checkpoint curves are provided in Appendix C.2).

4.5. Online support-size ablation

Appendix D.3 examines the impact of the support size k with frozen prefixes, reporting the exact covariance trace and NSR. In contrast, here we provide an end-to-end online ablation of OPD. Specifically, Tab. 6 and Fig. 5 show the results across teacher support sizes $k \in \{1, 4, 16, 32\}$. As shown in Tab. 6, ResOPD consistently outperforms HETS across all tested support sizes in average downstream accuracy, with improvements ranging from +0.11 percentage points at $k = 1$ up to +1.56 points at $k = 32$. However, while downstream task scores consistently favor ResOPD, the intrinsic variance reduction benefit offered by ResOPD is expected to diminish as k increases, since a larger teacher support naturally captures more probability mass and renders tail corrections less frequent.

Fig. 5 validates this mechanism on the runs from Tab. 6. In Fig. 5(a), HETS and ResOPD exhibit nearly identical variance factors P_{τ_p} , both dropping by over three orders of magnitude from $k = 1$ to 32. This directly scales the theoretical ResOPD–HETS variance contrast alongside tail-mean approximation quality. Meanwhile, Fig. 5(b) shows that ResOPD consistently lowers tail log-mass mismatch across all supports, with the largest margin in tight-support regimes.

Table 6: Performance under different Top- k support sizes (matched one-epoch budget). Evaluation setups follow Tab. 3 (avg@16 for AIME26/HMMT26 and avg@2 for OlymMATH). Δ denotes the average improvement of ResOPD over HETS.

k	Method	AIME26	HMMT26	OlymMATH	Average	Δ
1	HETS	60.62	37.69	34.75	44.35	+0.11
	ResOPD	56.88	39.02	37.50	44.47	
4	HETS	59.17	36.93	34.50	43.53	+1.01
	ResOPD	60.00	37.88	35.75	44.54	
16	HETS	57.92	39.39	35.75	44.35	+0.70
	ResOPD	59.38	39.02	36.75	45.05	
32	HETS	59.79	39.02	36.75	45.19	+1.56
	ResOPD	61.67	38.07	40.50	46.75	

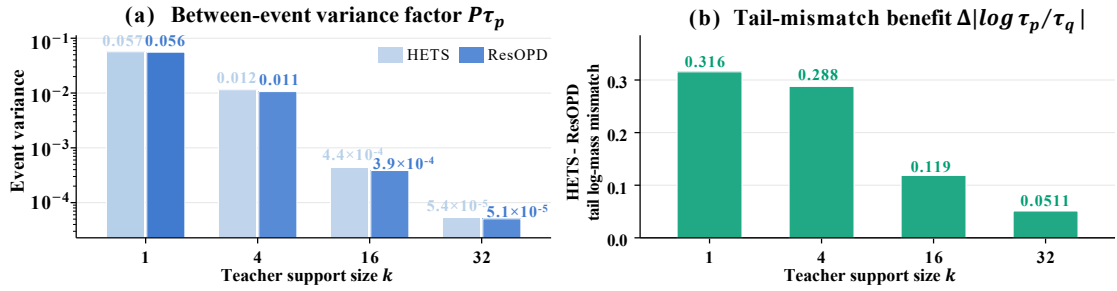


Figure 5: Online diagnostic metrics across teacher Top- k support sizes. (a) Between-event variance factor P_{τ_p} for HETS and ResOPD, characterizing the scaling coefficient of the head-versus-tail partition. (b) Tail log-mass mismatch reduction (HETS – ResOPD, measured via $\Delta |\log(\tau_p/\tau_q)|$).

4.6. Rollout sampling: default filters versus full support

While the idealized ancestral sampling condition $A \sim p_\theta$ is strictly enforced in both our exact enumeration audits (Sec. 4.2) and the cross-domain transfer experiments (Sec. 4.4), we also verify for completeness that the primary math results do not depend on rollout truncation. Specifically, we evaluate the first training epoch under full-support sampling (top- $p = 1.0$, top- $k = -1$) while keeping all other training hyperparameters identical. As shown in Tab. 7, removing the rollout filters lowers AUC for all three evaluated methods. Nevertheless, ResOPD consistently maintains its advantage over HETS and the naive baseline across both benchmark accuracy and trajectory AUC. In addition, we observe that removing rollout filters yields longer reasoning trajectories and a higher frequency of hitting the response token limit during training.

4.7. Learning-rate sensitivity

Experimental setup and evaluation protocol. To evaluate optimization robustness under aggressive step sizes, we conduct a controlled learning-rate sweep ($k = 4$) identical to the primary Math experimental setup, except that training is compressed to 1 epoch (64 optimizer steps). All runs share a batch size of 16, a 32K response budget, and identical sampling configurations with the primary experiments. Checkpoints at steps 0, 6, 32, and 64 are evaluated under the Macro-4 protocol (averaging 16 responses per problem for AIME26 and HMMT26, and 2 responses for DeepMath held-out and OlymMATH).

Performance resilience across learning rates. As shown in Tab. 8, ResOPD demonstrates substantially stronger resilience at elevated learning rates. When scaling the learning rate to 5×10^{-6} , HETS suffers severe degradation, with its final score dropping from 45.94% to 42.54% (−3.40%) and its trajectory AUC declining by 2.24%. In contrast, ResOPD retains consistent performance (46.35% final score, 47.77% AUC), widening its lead over HETS to +3.81% in final score and +3.89% in AUC. Examining the sub-benchmark breakdown at this aggressive rate, ResOPD’s advantage is consistently positive across all four constituent benchmarks at step 64: +3.96% on AIME26, +2.08% on HMMT26, +6.94% on DeepMath held-out, and +2.25% on OlymMATH. We provide detailed training dynamics across different learning rates in Appendix D.4.

Table 7: First-epoch Math performance under default vs. full-support rollouts. Benchmark scores and Avg. denote step-64 accuracy (%); AUC (%) measures the normalized learning curve area over steps $\{0, 6, 32, 64\}$. *Default*: $\text{top-}p = 0.95$, $\text{top-}k = 20$; *Full*: $\text{top-}p = 1.0$, $\text{top-}k = -1$.

Rollout	Method	DeepMath	AIME26	HMMT26	OlymMATH	Avg.	AUC
Default	Naive Top- k rKL	53.17	56.04	35.42	35.00	44.91	45.45
	HETS	53.17	59.17	36.93	34.50	45.94	46.12
	ResOPD	54.76	60.00	37.88	35.75	47.10	47.57
Full	Naive Top- k rKL	45.44	55.42	35.80	34.25	42.72	44.96
	HETS	49.80	57.50	37.88	35.50	45.17	45.68
	ResOPD	57.14	60.42	37.88	36.00	47.86	46.79

Table 8: Learning-rate sensitivity of HETS vs. ResOPD. Δ denotes the improvement of ResOPD over HETS.

Learning Rate	Final Score (%)			Trajectory AUC (%)		
	HETS	ResOPD	Δ	HETS	ResOPD	Δ
1×10^{-6}	45.94	47.10	+1.16	46.12	47.57	+1.45
2×10^{-6}	45.94	46.61	+0.67	45.69	46.17	+0.48
5×10^{-6}	42.54	46.35	+3.81	43.88	47.77	+3.89

Stability boundaries and failure modes. When the learning rate is scaled further to 1×10^{-5} , both methods diverge rapidly, triggering heavy gradient clipping (4/6 steps for HETS and 5/6 steps for ResOPD) and severe response-length collapse within six optimizer steps. While ResOPD does not shift the ultimate divergence boundary (1×10^{-5}) under our default training recipe, its coarse-tail residualization provides a substantially wider stability margin in moderately aggressive regimes. Scaling beyond this threshold may require coordinating higher learning rates with complementary stabilization techniques (such as larger batch sizes or tighter clipping thresholds), which we leave to future work.

4.8. Comparing the two-control endpoints

The analysis in Sec. 3.4 identifies a distinct opportunity to reduce within-tail variance through γ . We evaluate that prediction using SC-TA ($\gamma = 1$) and ResOPD ($\gamma = 0$), keeping the observable coarse control $\mathbf{B} = \mu_B$ fixed.

4.8.1. Cross-distribution fixed-prefix audit

While the two-control framework indicates that SC-TA ($\gamma = 1$) can mitigate within-tail fluctuations beyond ResOPD ($\gamma = 0$), Equation (29) implies this ordering is distribution-dependent rather than universal. To evaluate stability across domains, we freeze the Qwen3.5-4B/27B pair (Top-4 support, 64-token horizon) across three natural prompt distributions (DeepMath [10], MegaScience [7] and SynLogic [19]), evaluating 128 trajectories (8,192 prefixes) with 20,000 bootstrap resamples. We track the normalized trace variance reduction:

$$\Delta := \frac{\mathcal{V}_{\text{SC}} - \mathcal{V}_{\text{ResOPD}}}{\mathcal{V}_{\text{sampled}}}, \quad (30)$$

where negative values favor SC-TA.

As reported in Tab. 9, the aggregate trace variance consistently favors SC-TA over ResOPD across all three evaluated domains ($\Delta \in [-0.207, -0.126]$). Simultaneously, prefix-level crossover remains non-trivial (up to 18.0% of prefixes favor ResOPD in SynLogic). This empirically corroborates the qualified prediction of our two-control formulation: while setting $\gamma = 1$ may yield lower variance in expectation, the local variance ordering remains strictly prefix-dependent rather than globally uniform.

In principle, one could pre-select a subset of prompts from the training distribution to empirically calibrate an optimal coefficient γ^* , potentially unlocking additional variance reduction during optimization (even though γ^* inherently evolves alongside model weights). In practice, however, our experiments indicate that the incremental gains from such calibration over the fixed heuristic $\gamma = 1$ are marginal. We therefore omit further discussion in favor of simplicity.

Table 9: SC-TA vs. ResOPD covariance trace ordering across natural prompt distributions.

Dataset	$\mathcal{V}_{SC}/\mathcal{V}_{ResOPD}$	Δ	ResOPD-Lower Prefixes
DeepMath	0.512	-0.126	6.8%
MegaScience	0.532	-0.207	11.3%
SynLogic	0.558	-0.161	18.0%

Table 10: Matched evaluation curves across training updates (%). Evaluated on Qwen3.5-27B-to-4B distillation across AIME26/HMMT26 (avg@16) and held-out DeepMath/OlymMATH (avg@2).

Method	Base (Step 0)	Step 6	Step 32	Step 64	Step 128	Step 192
HETS	42.85	48.86	45.04	45.94	46.60	45.72
ResOPD	42.85	48.23	47.95	47.10	47.71	50.12
SC-TA	42.85	47.66	46.69	47.26	48.27	46.33

4.8.2. Downstream evaluation comparison

While SC-TA halves the fixed-prefix residual covariance trace of ResOPD, this advantage does not necessarily translate into superior downstream accuracy. As shown in Tab. 10, SC-TA does not exhibit a consistent downstream advantage over ResOPD: it is slightly stronger at several intermediate checkpoints, whereas ResOPD achieves the higher terminal score, with trajectory-level AUC remaining similar.

This decoupling between gradient covariance trace and downstream performance is informative rather than contradictory. While lower estimator variance validates the two-control residualization mechanism locally, downstream benchmark accuracy reflects the global trajectory of policy updates. Consequently, while SC-TA exhibits lower overall variance across the evaluated audit sets, ResOPD retains strong aggregate performance in this evaluated setting without explicitly subtracting the additional within-tail score fluctuation.

5. Conclusion

We introduced ResOPD, a principled estimator for full-vocabulary reverse-KL distillation under sparse teacher feedback. By integrating the observable coarse distribution analytically and sampling only the unresolved fine-tail residual, ResOPD reduces variance in the evaluated settings while preserving exact unbiasedness under ancestral on-policy sampling, without expanding the teacher payload or adding forward passes. Supported by theoretical analysis, exact covariance audits, and empirical evaluations across math, science, and coding tasks, ResOPD demonstrates higher aggregate downstream scores in the evaluated settings.

References

- [1] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263, 2024. 1, 2
- [2] Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. Opencodereasoning: Advancing data distillation for competitive coding. *arXiv preprint arXiv:2504.01943*, 2025. 4.4
- [3] Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18085–18108, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.885. URL <https://aclanthology.org/2025.acl-long.885/>. 1, 1, 2
- [4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 4.1
- [5] Jasper Dekoninck, Nikola Jovanović, Tim Gehringer, Kári Rognvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. 2026. URL <https://arxiv.org/abs/2605.00674>. 4.1

- [6] Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. SuperGPQA: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38, 2026. 4.4
- [7] Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *arXiv preprint arXiv:2507.16812*, 2025. 4.8.1
- [8] Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiakai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026. 2, 4.1
- [9] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717, 2024. 1, 2
- [10] Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. 2025. URL <https://arxiv.org/abs/2504.11456>. 4.1, 4.8.1
- [11] Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026. 2, 4.4
- [12] Huipeng Huang and Hongxin Wei. Tail-aware top- k on-policy distillation. *arXiv preprint arXiv:2608.14728*, 2026. 1, 1, 2, 3.2, 3.4.2
- [13] Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. 3.2
- [14] Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pages 58791–58831, 2025. 4.4
- [15] Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 42217–42227, 2026. 2
- [16] Woogyol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. 2
- [17] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL <https://aclanthology.org/D16-1139/>. 1
- [18] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024. 2
- [19] Junteng Liu, Yuanxiang Fan, Jiang Zhuo, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, et al. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. *Advances in Neural Information Processing Systems*, 38:100976–100997, 2026. 4.8.1
- [20] Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. <https://thinkingmachines.ai/blog/on-policy-distillation>. 1, 2
- [21] Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025. 4.4
- [22] NVIDIA Corporation. Opensciencereasoning-2. Hugging Face dataset, 2025. URL <https://huggingface.co/datasets/nvidia/OpenScienceReasoning-2>. 4.4
- [23] Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026. 1, 2
- [24] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>. 4.1
- [25] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023. 4.4
- [26] Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models, 2025. URL <https://arxiv.org/abs/2503.21380>. 4.1
- [27] Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. 1

- [28] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4.4
- [29] Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. 1
- [30] Lunjun Zhang and Jimmy Ba. Ema policy gradient: Taming reinforcement learning for llms with ema anchor and top-k kl. *arXiv preprint arXiv:2602.04417*, 2026. 1, 2, 3.1
- [31] Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026. 4.1
- [32] Siqu Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026. 1, 2, 3.1

A. Detailed estimator derivations

The main text gives the estimator, its unbiasedness identity, and the unified variance analysis. This appendix supplies the detailed Top- k bias derivations and the single-control oracle calculation, including the boundary-term cancellation for the complete coarse distribution.

A.1. Three biased forms of Teacher Top- k reverse KL

This subsection distinguishes three operations that are often conflated: renormalizing both distributions on the head, truncating the scalar reverse-KL objective, and truncating only the stopped-coefficient gradient. Each is computable from Teacher Top- k , but none generally estimates correctly the gradient of full-vocabulary reverse KL. Here “bias” is relative to that target gradient, not Monte Carlo sampling bias. Fix the support S and Teacher probabilities, let z be unrestricted Student logits, and define

$$\begin{aligned} R &:= \sum_i p_i r_i = D_{\text{KL}}(p||q), & R_S &:= \sum_{i \in S} p_i r_i, \\ P &:= \sum_{i \in S} p_i, & Q &:= \sum_{i \in S} q_i. \end{aligned} \quad (31)$$

The exact full-rKL logit gradient, which provides the reference below, is

$$[\nabla_z R]_j = p_j(r_j - R). \quad (32)$$

Renormalized Top- k reverse KL. Define $\tilde{p}_i = p_i/P$, $\tilde{q}_i = q_i/Q$ and $\tilde{r}_i = \log(\tilde{p}_i/\tilde{q}_i)$ for $i \in S$, with $\tilde{R} = \sum_{i \in S} \tilde{p}_i \tilde{r}_i$. Since $\tilde{p}_i = \exp(z_i) / \sum_{\ell \in S} \exp(z_\ell)$, the objective and its logit gradient are

$$\begin{aligned} L_{\text{renormalized}} &= \sum_{i \in S} \tilde{p}_i \log \frac{\tilde{p}_i}{\tilde{q}_i}, \\ [\nabla_z L_{\text{renormalized}}]_j &= \mathbb{I}\{j \in S\} \tilde{p}_j (\tilde{r}_j - \tilde{R}). \end{aligned} \quad (33)$$

Every tail-logit component is exactly zero, whereas Equation (32) is generally nonzero there. The global minimum is the set

$$\tilde{p}_i = \tilde{q}_i \quad (i \in S) \iff p_i = \frac{P}{Q} q_i \quad (i \in S), \quad (34)$$

with arbitrary P and arbitrary tail shape. Renormalization therefore learns only the conditional head distribution; it identifies neither head mass nor the omitted conditional distribution.

Scalar-truncated Top- k reverse KL. The unnormalized scalar shortcut is $L_{\text{truncated}} = R_S = \sum_{i \in S} p_i r_i$. Writing $C_S := \sum_{i \in S} p_i(r_i + 1) = R_S + P$, direct differentiation gives

$$\begin{aligned} \nabla_\theta L_{\text{truncated}} &= \mathbf{h} + \nabla_\theta P, \\ [\nabla_z L_{\text{truncated}}]_j &= \mathbb{I}\{j \in S\} p_j(r_j + 1) - p_j C_S. \end{aligned} \quad (35)$$

Thus its bias relative to full rKL is $\nabla_\theta P - \tau_p \boldsymbol{\mu}_{\text{tail}}$. In particular, for $j \notin S$, $[\nabla_z L_{\text{tr}}]_j = -p_j C_S$; when $C_S > 0$, gradient descent increases every omitted logit.

The corresponding global optimum makes the failure more explicit. For each head coordinate,

$$x \log \frac{x}{q_i} \geq -\frac{q_i}{e}, \quad \text{with equality iff } x = \frac{q_i}{e}. \quad (36)$$

Because $\sum_{i \in S} q_i/e = Q/e < 1$, all coordinatewise minima are jointly feasible, and therefore

$$p_i^* = \frac{q_i}{e} \quad (i \in S), \quad P^* = \frac{Q}{e}, \quad \sum_{j \notin S} p_j^* = 1 - \frac{Q}{e}, \quad (37)$$

with arbitrary allocation inside the tail. Even if Q is near one, this objective prefers roughly $1 - 1/e$ Student mass outside the Teacher head; the tail-push is thus an objective-level property, not merely a transient gradient sign.

Stopped-coefficient Top- k reverse-KL gradient. Stopping the coefficients removes the ∇P boundary term, yet leaves the tail expectation uncorrected:

$$\begin{aligned} \mathbf{g}_{\text{Top-}k} &:= \sum_{i \in S} \text{stopgrad}(p_i r_i) \mathbf{s}_i = \mathbf{h}, \\ [\mathbf{g}_{\text{Top-}k}]_j &= \mathbb{I}\{j \in S\} p_j r_j - p_j R_S, \quad \mathbf{g}_{\text{Top-}k} - \mathbf{g} = -\tau_p \boldsymbol{\mu}_{\text{tail}}. \end{aligned} \quad (38)$$

In unrestricted categorical-logit space with a non-empty tail, vanishing gradients ($\mathbf{g}_{\text{Top-}k} = \mathbf{0}$) force $R_S = 0$ via the tail components ($j \notin S$), which subsequently implies $r_i = 0$ for all head components ($i \in S$). Consequently, $p_i = q_i$ on the head, correctly matching the aggregate tail mass ($1 - P = 1 - Q$); however, the conditional distribution over the tail remains unconstrained and arbitrary. While this stopped-gradient formulation alters the underlying target the least among the three heuristics, it remains inherently biased whenever the tail vector is non-zero. HETS and ResOPD resolve this residual bias by introducing a sampled tail correction to restore the missing expectation.

A.2. Single-control residualization and its variance optimum

Use the head/tail notation of Equation (5), let $J = \mathbb{I}\{A \notin S\}$, and let $\mathbf{B}$ be any finite prefix-measurable vector stopped before sampling $A \sim p$. Consider the one-control estimator

$$\hat{\mathbf{g}}(\mathbf{B}) := \mathbf{h} + \tau_p \mathbf{B} + J(\mathbf{G}_A - \mathbf{B}). \quad (39)$$

It is unbiased for every such $\mathbf{B}$:

$$\mathbb{E}[\hat{\mathbf{g}}(\mathbf{B})] = \mathbf{h} + \tau_p \mathbf{B} + \tau_p (\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}) = \mathbf{g}. \quad (40)$$

Conditioned on $J = 0$, its value is the deterministic vector $\mathbf{h} + \tau_p \mathbf{B}$. Conditioned on $J = 1$, its mean is $\mathbf{h} + \tau_p \mathbf{B} + \boldsymbol{\mu}_{\text{tail}} - \mathbf{B}$ and its covariance is $\text{Cov}_{\text{tail}}(\mathbf{G}_A)$. The two conditional means differ by $\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}$. Hence the law of total covariance gives

$$\mathcal{V}(\hat{\mathbf{g}}(\mathbf{B})) = \tau_p \mathcal{V}_{\text{tail}}(\mathbf{G}_A) + P \tau_p \|\boldsymbol{\mu}_{\text{tail}} - \mathbf{B}\|_2^2. \quad (41)$$

The first term is independent of $\mathbf{B}$. Therefore, whenever $0 < P < 1$, the unique trace-covariance minimizer and its exact regret are

$$\begin{aligned} \mathbf{B}^* &= \boldsymbol{\mu}_{\text{tail}}, \\ \mathcal{V}(\hat{\mathbf{g}}(\mathbf{B})) - \mathcal{V}(\hat{\mathbf{g}}(\boldsymbol{\mu}_{\text{tail}})) &= P \tau_p \|\mathbf{B} - \boldsymbol{\mu}_{\text{tail}}\|_2^2. \end{aligned} \quad (42)$$

The oracle removes all head-versus-tail event variance, but it is unavailable under a sparse Teacher interface because it requires the full conditional fine-tail mean.

HETS and ResOPD are the two payload-compatible choices

$$\hat{\mathbf{g}}^{\text{HETS}} = \hat{\mathbf{g}}(\mathbf{0}), \quad \hat{\mathbf{g}}^{\text{ResOPD}} = \hat{\mathbf{g}}(\boldsymbol{\mu}_B). \quad (43)$$

Their regrets relative to the single-control oracle are respectively $P \tau_p \|\boldsymbol{\mu}_{\text{tail}}\|_2^2$ and $P \tau_p \|\boldsymbol{\mu}_{\text{tail}} - \boldsymbol{\mu}_B\|_2^2$. Subtraction yields

$$\mathcal{V}(\hat{\mathbf{g}}^{\text{HETS}}) - \mathcal{V}(\hat{\mathbf{g}}^{\text{ResOPD}}) = P \tau_p (2\langle \boldsymbol{\mu}_{\text{tail}}, \boldsymbol{\mu}_B \rangle - \|\boldsymbol{\mu}_B\|_2^2). \quad (44)$$

Thus ResOPD has lower trace covariance exactly when $\boldsymbol{\mu}_B$ is closer to $\boldsymbol{\mu}_{\text{tail}}$ than the HETS control $\mathbf{0}$ is; no unconditional ordering is asserted.

It remains to verify why $\boldsymbol{\mu}_B$ is available without the fine Teacher tail. The reverse KL of the complete coarse distribution over the k head tokens and one tail bucket is

$$L_{\text{coarse}} := \sum_{i \in S} p_i \log \frac{p_i}{q_i} + \tau_p \log \frac{\tau_p}{\tau_q}. \quad (45)$$

Since $\nabla \tau_p = \tau_p \bar{\mathbf{s}}_{\text{tail}} = -\nabla P$, direct differentiation gives

$$\begin{aligned} \nabla L_{\text{coarse}} &= \mathbf{h} + \nabla P + (b+1) \nabla \tau_p \\ &= \mathbf{h} + b \nabla \tau_p = \mathbf{h} + \tau_p \boldsymbol{\mu}_B = \mathbf{g}_{\text{coarse}}. \end{aligned} \quad (46)$$

The boundary terms cancel because the tail pseudo-token completes the coarse support. This is why $\mathbf{g}_{\text{coarse}}$ is an exact gradient (but biased), even though $\boldsymbol{\mu}_B$ need not equal the true conditional fine-tail mean.

Table 11: Primary Math OPD experimental configurations and hyperparameters.

Category	Hyperparameter	Configuration / Value
Models	Student model	Qwen3.5-4B
	Teacher model	Qwen3.5-27B
Data	Training dataset	1,024 prompts sampled from DeepMath-103K
	Max prompt length	2,048
Sampling	Enable thinking	True
	Max response length	32,768 (training and evaluation)
	temperature	1.0
	top- p	0.95
	top- k	20
Training	Training epochs	3
	Global batch size	16
	Optimizer	AdamW
	Gradient clipping	1.0
	Learning rate	1×10^{-6}
System & Hardware	Compute infrastructure	16 $\times$ NVIDIA H200 GPUs
	Resource allocation	8 GPUs for Student rollout, 4 for Student training, 4 for hosting the Teacher inference service
	Software frameworks	Training: verl Inference: vLLM

B. Detailed training and evaluation configurations

Tabs. 11 and 12 collect the complete settings for the primary mathematics and cross-domain runs. Their sampling assumptions are discussed in Secs. 3.3 and 4.1; the matched full-support mathematics ablation is reported in Sec. 4.6.

C. Step-wise evaluation results

C.1. Math OPD experiments

Fig. 6 illustrates the training trajectories across evaluation suites. On the DeepMath held-out set, ResOPD displays consistent late-stage gains, achieving the highest terminal accuracy among all estimators. While competitive baselines such as TA-OPD and Teacher Top- k vOPD show steady improvements, they fall short of ResOPD at the terminal checkpoint. Meanwhile, Sampled-token OPD deteriorates severely after an early peak, and HETS suffers a late-stage drop. A similar trend holds on the external math competitions: although TA-OPD maintains strong intermediate trajectory performance and vOPD remains competitive, ResOPD exhibits superior late-stage resilience and decisively surpasses all alternatives at the final step, avoiding the post-peak degradation observed in estimators like Sampled-token OPD and HETS.

Together with the lower fixed-prefix covariance trace and improved online training diagnostics reported elsewhere in the paper, these trajectories are consistent with the proposed coarse-tail residualization reducing stochastic gradient noise and improving late-stage optimization stability.

C.2. Cross-Domain OPD Experiments

Fig. 7 illustrates the checkpoint evaluation trajectories along the fixed training progression (steps 0, 10, 25, 50, and 100) across all four benchmarks. On code benchmarks (LiveCodeBench v5 and v6), ResOPD and HETS rapidly converge and consistently outperform Sampled-token OPD by a substantial margin throughout training, with ResOPD achieving the highest overall trajectory AUC. In contrast, science benchmarks display mixed

Table 12: Cross-domain Qwen3 OPD experimental configurations and hyperparameters.

Category	Hyperparameter	Configuration / Value
Models	Student model	Qwen3-4B
	Teacher model	Qwen3-30B-A3B-Instruct-2507
Data	Training dataset	30,000 deduplicated prompts; Math/Code/Science: 10,000 each
	Max prompt length	1,024
Sampling	Enable thinking	False
	Max response length	8192 (training) and 32768 (evaluation)
	temperature	1.0
	top- p	1.0
	top- k	−1
Training	Training steps	100
	Global batch size	288
	Optimizer	AdamW
	Gradient clipping	1.0
	Peak learning rate	5×10^{-6}
	LR schedule	10-step warmup; cosine decay to $0.1 \times \text{peak}$
System & Hardware	Compute infrastructure	16× NVIDIA H200 GPUs
	Resource allocation	8 GPUs for student rollout, 4 for student training, 4 for hosting the teacher inference service
	Software frameworks	Training: verl Inference: vLLM

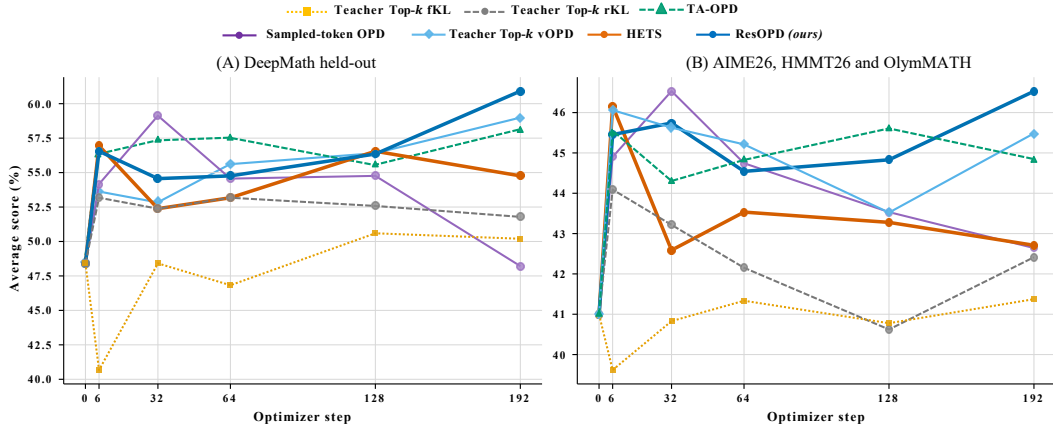


Figure 6: Step-wise learning curves corresponding to Tab. 3 on DeepMath held-out (left) and external competitions (right) across 192 optimizer steps.

domain dynamics: Sampled-token OPD demonstrates superior retention and final accuracy on GPQA-D, whereas ResOPD remains competitive across training steps and achieves the best terminal performance on SuperGPQA. These trajectories corroborate that while Sampled-token OPD remains strong in specialized scientific reasoning, ResOPD provides transfer performance on code tasks without sacrificing overall multi-domain balance.

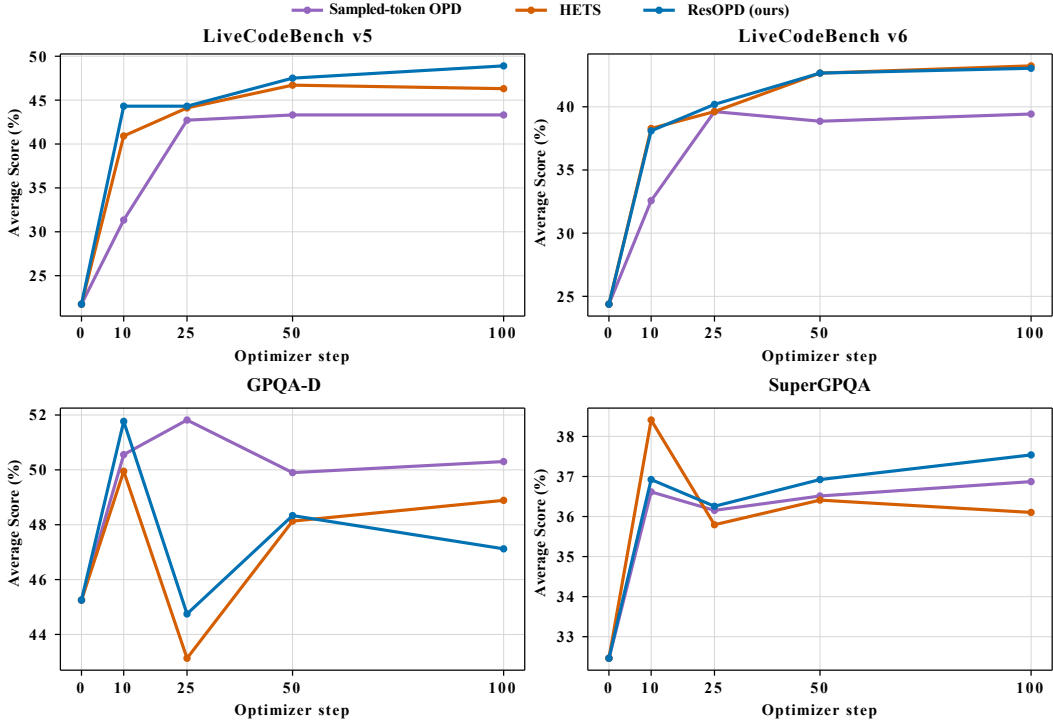


Figure 7: Evaluation trajectories during cross-domain transfer distillation from Qwen3-30B-A3B-Instruct to Qwen3-4B across optimizer steps (0, 10, 25, 50, and 100).

D. Details of diagnostics

D.1. What the online residual moments measure

At optimizer update u , let $i \in \{1, \dots, B_u\}$ index the non-empty sampled responses and $t \in \{1, \dots, T_{ui}\}$ index their valid tokens. We define the scalar support residual and its components as

$$P_{uit} := p_\theta(S_{uit} | x_{uit}), \quad H_{uit} := \mathbb{I}\{A_{uit} \in S_{uit}\}, \quad Z_{uit} := H_{uit} - P_{uit}. \quad (47)$$

Here, P_{uit} denotes the Student probability mass allocated to the transmitted Teacher Top- k support S_{uit} , while H_{uit} indicates whether the sampled Student token lands within this set. Under an exact on-policy draw, the residual is conditionally zero-mean, $\mathbb{E}[Z_{uit} | x_{uit}] = 0$, with conditional variance $\mathbb{E}[Z_{uit}^2 | x_{uit}] = P_{uit}(1 - P_{uit})$. More generally, its uncentered second moment captures both intrinsic sampling variability and any realized calibration drift.

Fig. 3 reports two diagnostic aggregations of this quantity. Panel C displays the masked token-level second moment:

$$M_{2,\text{tok}}^{(u)} = \frac{1}{\sum_{i=1}^{B_u} T_{ui}} \sum_{i=1}^{B_u} \sum_{t=1}^{T_{ui}} Z_{uit}^2, \quad (48)$$

whereas Panel D evaluates the trajectory-level variance by first averaging within each response and then weighting responses uniformly:

$$\bar{Z}_{ui} = \frac{1}{T_{ui}} \sum_{t=1}^{T_{ui}} Z_{uit}, \quad M_{2,\text{traj}}^{(u)} = \frac{1}{B_u} \sum_{i=1}^{B_u} \bar{Z}_{ui}^2. \quad (49)$$

Intuitively, $M_{2,\text{tok}}$ captures the instantaneous magnitude of the unresolved head-versus-tail events (which is leveraged in Fig. 5 to approximate the between-event variance factor P_{τ_p}), whereas $M_{2,\text{traj}}$ measures the persistent, signed event imbalance surviving along sequence rollouts. Defining the batch mean $\bar{Z}_u =$

Table 13: Support-size (k) sweep under fixed Teacher Top- k payloads. “Var. Removed” measures the empirical covariance-trace reduction relative to sampled-token OPD. Payload accounts for k Top- k entries plus the sampled token (INT32 ID and FP32 log-probability each). Student mass and Head Hit denote the predicted and empirical probabilities of landing in the Top- k support, respectively. Best variance reduction and NSR per setting are in **bold**.

Dataset Student/Teacher	k	Payload (B/token)	Student mass	Head Hit	Var. Removed (%) $\uparrow$		NSR $\downarrow$	
					HETS	ResOPD	HETS	ResOPD
GSM8K 0.8B/4B	1	16	75.0%	74.8%	−0.5	20.5	5.446	4.310
	4	40	93.1%	92.9%	27.3	52.5	3.940	2.573
	16	136	97.7%	97.9%	62.6	75.3	2.027	1.337
	64	520	99.3%	99.7%	85.1	89.5	0.807	0.570
HMMT26 4B/27B	1	16	89.1%	89.3%	−1.0	32.0	6.656	4.479
	4	40	99.1%	98.8%	55.8	75.3	2.914	1.630
	16	136	99.9%	100.0%	89.8	94.2	0.675	0.382

$B_u^{-1} \sum_{i=1}^{B_u} \bar{Z}_{ui}$, the empirical standard error of the trajectory mean is given by

$$\widehat{\text{SE}}(\bar{Z}_u) = \sqrt{\frac{\max\{M_{2,\text{traj}}^{(u)} - \bar{Z}_u^2, 0\}}{B_u - 1}}. \quad (50)$$

The connection to ResOPD is exact at the scalar-coefficient level. Noting that tail indicators and tail masses satisfy $J_{uit} = 1 - H_{uit}$ and $\tau_{p,uit} = 1 - P_{uit}$, we obtain $Z_{uit} = \tau_{p,uit} - J_{uit}$. Substituting this into Equation (10) yields

$$\hat{\mathbf{g}}_{uit}^{\text{ResOPD}} - \hat{\mathbf{g}}_{uit}^{\text{HETS}} = Z_{uit} \boldsymbol{\mu}_{B,uit}. \quad (51)$$

Hence, the two M_2 metrics directly track the magnitude and sequence-level persistence of the random scalar modulating ResOPD’s bucket control. Because they intentionally abstract away $\|\boldsymbol{\mu}_{B,uit}\|$, its alignment with the fine-tail mean, and the orthogonal within-tail fluctuations, they serve as computationally lightweight, mechanism-aligned online diagnostics rather than full covariance estimators or replacements for fixed-prefix NSR. Finally, while Z can trivially be recorded for a sparse Top- k objective, direct Top- k rKL lacks a sampled tail correction and remains inherently biased with respect to the true rKL target. Its zero action-sampling variance is therefore not comparable to the unbiased vOPD, HETS and ResOPD estimators in Panels C and D.

D.2. Finite gradient-enumeration check

We also verify the unbiasedness identities without Monte Carlo error on a four-action softmax. The Student and Teacher probabilities are respectively (0.35, 0.30, 0.20, 0.15) and (0.60, 0.25, 0.10, 0.05), and the first two actions form the Teacher Top-2 support. We differentiate the full reverse KL directly, enumerate each possible Student action, and weight its estimator gradient by the exact Student action probability. The resulting HETS expectation matches the dense full-rKL gradient to maximum absolute error 5.6×10^{-17} .

ResOPD changes the per-action vector by $(\tau_p - J)\boldsymbol{\mu}_B$ relative to HETS. Because the same enumeration has $\mathbb{E}[J] = \tau_p$, this correction sums exactly to zero and ResOPD recovers the same dense gradient. This finite check complements the symbolic identity in Equation (10): it validates the stopped-coefficient and support-mask implementation semantics, but it is not empirical evidence about covariance or downstream accuracy.

D.3. Complete frozen-prefix support diagnostics

Across both suites in Tab. 13, ResOPD consistently outperforms HETS at every matched support size, achieving higher covariance-trace elimination and lower NSR. Expanding k from 4 to 16 and 64 scales transmission costs by $3.4\times$ and $13\times$ ($40 \rightarrow 136 \rightarrow 520$ B/token) for diminishing marginal variance gains. Crucially, the head hit rates demonstrate why high support coverage alone is insufficient: even when head mass reaches 98.8% ($k = 4$ on HMMT26), the rare 1.2% tail draws still induce a substantial HETS–ResOPD variance gap (55.8% vs. 75.3%

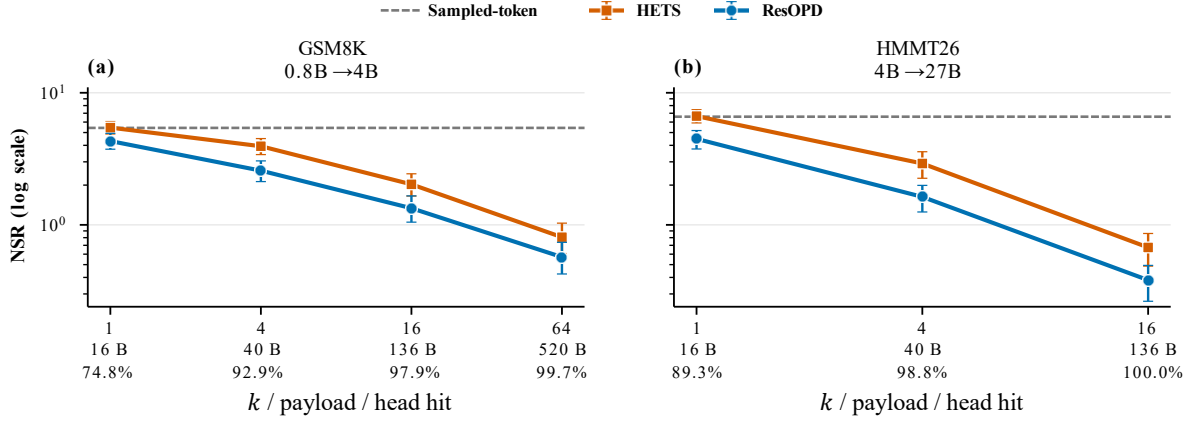


Figure 8: Comprehensive frozen-prefix support-size (k) diagnostics across benchmarks. ResOPD consistently reduces the dimensionless NSR across all evaluated support sizes on GSM8K and HMMT (log scale). Axis annotations detail the transmitted Teacher support size k , the corresponding communication payload (bytes/-token), and the empirical head-hit probability. Error bars denote 95% trajectory-level bootstrap confidence intervals.

removed). Moreover, at $k = 1$ where HETS exhibits negative reduction (-1.0%), ResOPD reliably eliminates 32.0% of the gradient variance. This firmly corroborates the theoretical condition in Equation (44), confirming that uncorrected tail draws dominate residual gradient noise despite their low nominal probability mass.

D.4. Training dynamics across learning rates

Underlying training dynamics. The training trajectories in Fig. 9 reveal two clear mechanisms behind this divergence. First, in terms of gradient magnitude, ResOPD maintains substantially lower pre-clipping gradient norms at 5×10^{-6} (mean 0.43 vs. 0.48; p99 0.97 vs. 1.84) and completely avoids gradient clipping (0/64 steps vs. 2/64 steps for HETS). Second, regarding residual stabilization, Panels (B)–(D) show that ResOPD consistently curbs the support-mass gap ($Q - P$) and suppresses both token- and trajectory-level residual second moments (M_2), consistent with the more stable learning trajectory observed at this step size.

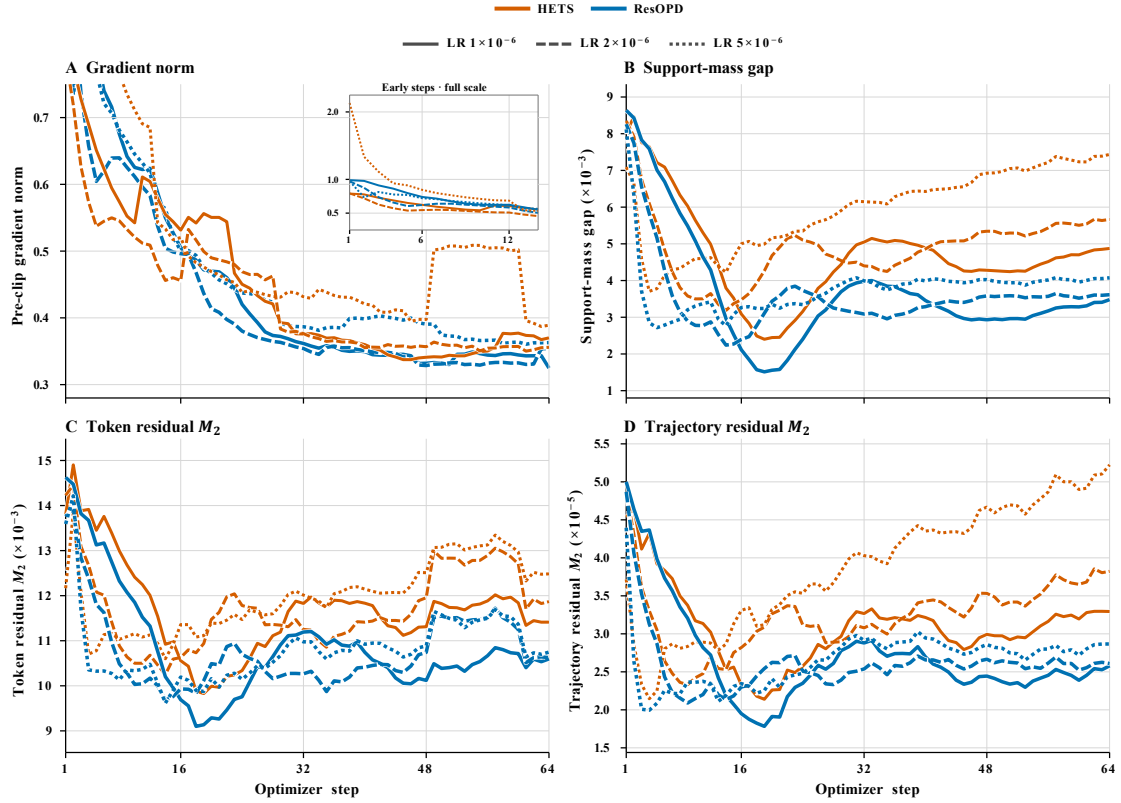


Figure 9: Training dynamics and learning-rate sensitivity of HETS (orange) vs. ResOPD (blue) on DeepMath across three learning rates (1×10^{-6} solid, 2×10^{-6} dashed, 5×10^{-6} dotted). **(A)** Pre-clipping global gradient norm (inset displays full-scale unclipped peaks over early steps). **(B)** Support-mass gap ($Q - P$) on the transmitted Teacher Top-4 tokens. **(C, D)** Token-level and trajectory-level second residual moments (M_2). While HETS exhibits large gradient-norm excursions and growing mass/residual discrepancies as the learning rate scales up to 5×10^{-6} , ResOPD maintains stable dynamics and stays close to its low-rate trajectories throughout optimization.