SEAD: Competence-Aware On-Policy Distillation via Entropy-Guided Supervision

Anonymous Authors[1]

Abstract

On-policy distillation (OPD) trains a student model on its own rollouts scored by a stronger teacher, eliminating the exposure bias of offline methods and matching or exceeding RL approaches for mathematical reasoning. However, OPD has a fundamental property absent in both offline distillation and RL: *teacher supervision quality depends on student competence*. Incoherent student rollouts yield noisy, uninformative gradients; already-mastered tokens yield redundant ones. This creates waste at three scales—tokens, training phases, and prompts—yet existing methods supervise uniformly. We introduce **SEAD** (Sparse Entropy-Adaptive Distillation), which uses entropy as a unified probe of this competence-dependent degradation: (1) joint teacher–student entropy partitions tokens into zones receiving tailored divergences or zero gradient ($\sim$50% skipped); (2) a cosine schedule anneals from forward to reverse KL as competence grows; (3) a competence-gated curriculum introduces math problems easy-to-hard. These components are symbiotically necessary: token selection requires coherent rollouts (curriculum), annealing requires monotonic improvement (also curriculum). On OLMo-3 (7B→32B), SEAD achieves +4.8 average accuracy over vanilla OPD across six math benchmarks including MATH-500, AIME, and OlympiadBench, with ablations confirming super-additive interactions among all three components.

1. Introduction

Large reasoning models (49B+ parameters) achieve strong mathematical problem-solving performance but are prohibitively expensive for deployment. Knowledge distillation compresses these capabilities into smaller, deployable students. The dominant *off-policy* paradigm—training on static teacher-generated traces—suffers from exposure bias: prediction errors compound autoregressively at inference, causing the student to drift into regions of the output space that the teacher's demonstrations never covered (Agarwal et al., 2024; Song & Zheng, 2026).

On-policy distillation (OPD) addresses this by training on student-generated rollouts scored by the teacher (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025), matching or exceeding RL methods like GRPO (Ko et al., 2026; Jin et al., 2026; Yang et al., 2026) on mathematical reasoning tasks. However, existing OPD methods apply supervision *uniformly*—the same divergence measure, on every token, at every training phase, for every prompt difficulty level.

We identify a single structural problem underlying this uniformity: **in OPD, supervision quality depends on student competence**. Unlike off-policy KD (where teacher traces are always coherent regardless of the student) or RL (where binary rewards are equally reliable regardless of rollout quality), OPD's per-token teacher corrections are only informative when the student's rollouts are sufficiently coherent for the teacher to meaningfully score. This competence-dependent degradation manifests at three scales:

- **Token level:** $\sim$50% of tokens are deterministic for both models (arithmetic steps, connectives, formatting)—supervising them wastes compute. Among the informative remainder, some call for sharpening via mode-seeking RKL while others require diversity preservation via mode-covering FKL (Jin et al., 2026).
- **Temporal level:** The optimal divergence evolves from mode-covering (early, when the student needs exploration) to mode-seeking (late, when refinement dominates), yet prior solutions use manual two-stage switches (Ko et al., 2026; Xu et al., 2026a).
- **Prompt level:** Mathematical problems beyond the student's capability produce incoherent rollouts on which teacher supervision is noise, not signal (Li et al., 2026; Zhao et al., 2026).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

These are not three independent problems but three symptoms of *ignoring how supervision quality varies with competence*. We observe that *entropy*—of both teacher and student distributions—provides a unified observable for the supervision quality at all three scales: low joint entropy indicates mastered tokens (skip); asymmetric entropy indicates informative corrections (Zone B/C); and prompt-level rollout entropy correlates with coherence.

We introduce **SEAD**, a framework operationalizing this principle through three symbiotically integrated components:

1. **Token-level: Sparse Entropy-Adaptive Divergence.** Joint teacher-student entropy partitions tokens into Zone A (skip—both confident), Zone B (RKL—teacher confident, student uncertain), Zone C (FKL—teacher uncertain at reasoning forks)—jointly determining *selection and divergence type* in a single mechanism.
2. **Temporal: Competence-Driven Annealing.** A continuous cosine schedule transitions from FKL-dominant to RKL-dominant supervision, tracking the natural evolution of Zone B/C composition as the student improves.
3. **Prompt-level: Competence-Gated Curriculum.** The first prompt-level curriculum designed specifically for OPD, addressing an open problem flagged by Zhao et al. (2026) and Li et al. (2026).

Crucially, these are *symbiotically necessary*: token selection requires coherent rollouts to produce a reliable entropy landscape (curriculum provides this), and annealing requires monotonic competence growth (curriculum guarantees it). The ablation in Section 3.3 confirms this—combining Token Zones with Annealing yields a 20× amplification over either in isolation, and adding curriculum produces further gains.

We validate SEAD on two model families: OLMo-3 (7B→32B) and Nemotron (8B→49B) across MATH-500, Minerva-Math, AIME 2024/2025, AMC 2023, and OlympiadBench. The full framework achieves +4.8 average accuracy over vanilla OPD, with the largest gains (+7.7 AIME24, +7.9 AIME25) on competition-level mathematics where reasoning depth matters most.

2. Method

2.1. Preliminaries and Problem Setting

Let π_θ and π_{te} denote the student and teacher policies respectively. For a given prompt $\mathbf{q}$, let $\mathbf{c}_t = (\mathbf{q}, x_1, \ldots, x_{t-1})$ denote the context up to step t, and $x_t \in \mathcal{V}$ the generated token. Each OPD iteration: (1) samples rollouts $\mathbf{x} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{q})$; (2) queries teacher logits on the student's rollout; (3) updates π_θ via a divergence loss computed between the teacher's and student's distributions at each token position. The two standard choices are **forward KL** (mode-covering):

$$\mathcal{L}_{\text{FKL}}(t) = \sum_{v \in \mathcal{V}} \pi_{\text{te}}(v|\mathbf{c}_t) \log \frac{\pi_{\text{te}}(v|\mathbf{c}_t)}{\pi_\theta(v|\mathbf{c}_t)} \tag{1}$$

and **reverse KL** (mode-seeking):

$$\mathcal{L}_{\text{RKL}}(t) = \sum_{v \in \mathcal{V}} \pi_\theta(v|\mathbf{c}_t) \log \frac{\pi_\theta(v|\mathbf{c}_t)}{\pi_{\text{te}}(v|\mathbf{c}_t)} \tag{2}$$

Forward KL encourages the student to cover all modes of the teacher, preserving diversity at reasoning branch points where multiple valid continuations exist. Reverse KL drives the student to concentrate on the teacher's high-probability modes, yielding sharper outputs but risking premature entropy collapse (Ko et al., 2026).

We formalize the core observation motivating SEAD:

Assumption 2.1 (OPD supervision quality—modulated PL condition)**.** Let $\ell_i(\theta)$ denote the per-prompt OPD loss and $\ell_i^\star = \inf_\theta \ell_i(\theta)$. There exists a non-decreasing function $\phi\colon [0,1] \to [0,1]$ with $\phi(0) = 0$, $\phi(1) = 1$, and a constant $\mu > 0$ such that for each prompt i:

$$\|\nabla \ell_i(\theta)\|^2 \geq \phi(p_i(\theta))\, \mu\, (\ell_i(\theta) - \ell_i^\star) \tag{3}$$

where $p_i(\theta)$ is the student's pass rate on prompt i.

Intuitively, gradient signal scales with competence: when the student cannot solve a problem at all ($p_i \approx 0$), its rollouts are incoherent and $\phi(p_i) \approx 0$, so the gradient norm lower bound vanishes regardless of the teacher's quality. Entropy serves as the observable proxy: joint teacher–student entropy reveals whether a token position is informative (high student, low teacher entropy ⇒ Zone B), exploratory (both high ⇒ Zone C), or redundant (both low ⇒ Zone A).

Assumption 2.1 motivates a unified principle: *allocate compute only where supervision quality is high*. SEAD operationalizes this at three granularities, presented in order of increasing scope. We note that the prompt-level curriculum (Section 2.5) is the foundational enabler: it ensures the entropy landscape is well-behaved (which the token-level selector requires), and it guarantees monotonic competence growth (which the temporal annealer assumes).

2.2. The Unified SEAD Objective

SEAD integrates token-level selection, temporal annealing, and prompt-level curriculum into a single unified loss. Let $\mathcal{B}$ and $\mathcal{C}$ denote the sets of token indices assigned to RKL and FKL respectively (formally defined in Section 2.3), and let $\lambda > 0$ be a hyperparameter balancing the intrinsic scale

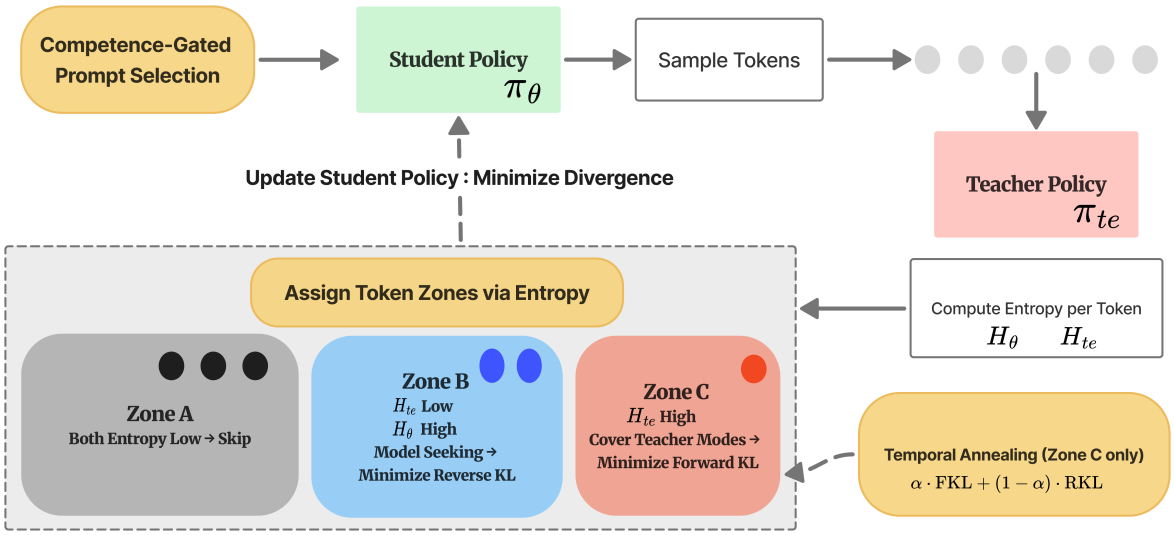


Figure 1. **Overview of SEAD.** The *competence-gated curriculum* (left, yellow) selects prompts within the student's current ability frontier ($d_i \leq c(t)$), ensuring rollouts are coherent enough for meaningful supervision. The student π_θ generates a rollout, which the teacher π_{te} scores with per-token logits. From these, we compute joint entropy (H_θ, H_{te}) and partition tokens into three zones: **Zone A** (gray)—both models confident, zero gradient ($\sim$50% of tokens); **Zone B** (blue)—teacher confident but student uncertain, supervised via reverse KL to sharpen toward the teacher's mode; **Zone C** (red)—teacher uncertain at reasoning forks, supervised via forward KL to preserve multi-path diversity. A *temporal annealing* schedule (bottom-right) modulates Zone C as $\alpha \cdot \mathcal{L}_{\text{FKL}} + (1-\alpha) \cdot \mathcal{L}_{\text{RKL}}$, smoothly transitioning from exploration (α=0.8) to refinement (α=0) over training. The combined loss updates the student; see Algorithm 1 for the full procedure.

differences between the two divergence terms:

$$\mathcal{L}_{\text{SEAD}}(\theta, t_{\text{step}}) = \frac{1}{|\mathcal{B} \cup \mathcal{C}|} \left(\sum_{t \in \mathcal{B}} \mathcal{L}_{\text{RKL}}(t) + \alpha(t_{\text{step}}) \cdot \lambda \sum_{t \in \mathcal{C}} \mathcal{L}_{\text{FKL}}(t) \right) \tag{4}$$

with prompts drawn from the eligible set $\mathbf{q} \sim \text{Uniform}(\mathcal{D}(t_{\text{step}}))$. Three control variables—all functions of training progress, all serving the same principle of allocating compute where supervision quality is high:

- **Zone partition** $\{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$: which tokens receive gradient and which divergence (Section 2.3);
- **Annealing coefficient** $\alpha(t_{\text{step}})$: FKL/RKL balance evolving with competence (Section 2.4);
- **Eligible set** $\mathcal{D}(t_{\text{step}}) = \{\mathbf{q}_i : d_i \leq c(t_{\text{step}})\}$: prompts within the competence frontier (Section 2.5).

Remark 2.2 (Symbiotic necessity). The three components are preconditions for each other, forming a virtuous cycle rather than a circular dependency. The resolution is *staged bootstrapping*: **without curriculum**, incoherent rollouts corrupt the entropy landscape, causing zone misclassification (confirmed empirically in Table 2: T alone yields only +0.32); **without annealing**, a fixed divergence ratio drives early mode collapse or late over-exploration; **without token selection**, $\sim$50% of compute is wasted on tokens where supervision is already maximal. The curriculum provides a reliable base, enabling zones and annealing to function correctly atop it.

2.3. Token-Level: Sparse Entropy-Adaptive Divergence

We define token-level entropy $H_{\text{te}}(t) = -\sum_v \pi_{\text{te}}(v|\mathbf{c}_t) \log \pi_{\text{te}}(v|\mathbf{c}_t)$ and $H_\theta(t)$ analogously. To maintain computational tractability, $H_{\text{te}}(t)$ is approximated using the top-k vocabulary subset from teacher logits. SEAD partitions tokens into three zones based on the joint entropy landscape:

- **Zone A** (Skip, $\sim\rho_A$%): Both $H_{\text{te}}(t)$ and $H_\theta(t)$ are low. **Zero gradient**—these tokens correspond to deterministic computation steps ("$3 \times 4 = 12$"), connectives ("Therefore"), and formatting. Both models already agree, so supervision carries no information.
- **Zone B** (RKL, $\sim\rho_B$%): $H_{\text{te}}(t)$ is low but $H_\theta(t)$ is high. The teacher is confident in a specific continuation while the student remains uncertain. Reverse KL *sharpens* the student toward the teacher's peaked mode.
- **Zone C** (FKL, $\sim\rho_C$%): $H_{\text{te}}(t)$ is high. These are genuine reasoning forks—positions where the teacher

Algorithm 1 SEAD: Competence-Aware On-Policy Distillation

Require: Student π_θ, teacher π_{te}, prompts $\mathcal{Q}$ with difficulty scores $\{d_i\}$, zone percentiles (ρ_A, ρ_B, ρ_C), annealing schedule $\alpha(\cdot)$, competence function $c(\cdot)$

Precompute per-prompt difficulty $d_i = 1 - p_i$ via student pass rate
for $t_{\text{step}} = 1$ to T_{total} **do**
  $\mathcal{D} \leftarrow \{\mathbf{q}_i : d_i \le c(t_{\text{step}})\}$ {Competence-gated curriculum}
  Sample prompt batch $\{\mathbf{q}_i\} \subset \mathcal{D}$; generate rollouts $\mathbf{x}_i \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{q}_i)$
  Query teacher $\pi_{\text{te}}(\cdot|\mathbf{c}_t)$; compute $H_{\text{te}}(t)$, $H_\theta(t)$ per position
  Assign tokens to Zone A, B, C via percentile thresholds on joint entropy
  $\alpha \leftarrow \alpha(t_{\text{step}})$ {Competence-driven annealing}
  $\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}\cup\mathcal{C}|}\big(\sum_{t\in\mathcal{B}} \mathcal{L}_{\text{RKL}}(t) + \alpha\lambda \sum_{t\in\mathcal{C}} \mathcal{L}_{\text{FKL}}(t)\big)$
  Update θ via clipped gradient step on $\mathcal{L}$
end for

itself maintains probability mass over multiple valid approaches (e.g., choosing between algebraic manipulation vs. geometric proof). Forward KL *covers* the teacher's modes to preserve exploratory diversity.

The per-step token-level loss is:

$$\mathcal{L}_{\text{SEAD}}^{\text{token}} = \frac{1}{|\mathcal{B}\cup\mathcal{C}|}\left(\sum_{t\in\mathcal{B}} \mathcal{L}_{\text{RKL}}(t) + \lambda \sum_{t\in\mathcal{C}} \mathcal{L}_{\text{FKL}}(t)\right) \quad (5)$$

with default configuration $\rho_A = 50$, $\rho_B = 40$, $\rho_C = 10$. Unlike EOPD (Jin et al., 2026) (which uses teacher entropy only for FKL/RKL switching) and TIP (Xu et al., 2026b) (which uses student entropy×divergence for importance weighting with a single divergence), SEAD jointly determines both *which tokens to supervise* and *which divergence to apply* via the joint teacher-student entropy space.

Sparse selection is approximately lossless. The following theorem shows that Zone A tokens contribute negligible gradient, justifying their exclusion. The key precondition is that Zone A tokens have genuinely low entropy for *both* teacher and student—a property that holds when the curriculum ensures coherent rollouts, preventing spurious low-entropy assignments.

Theorem 2.3 (Sparse gradient approximation)**.** *Under bounded score functions (Assumption D.1) and the condition that Zone A tokens satisfy $H_{\text{te}}(t) \le \tau$, $H_\theta(t) \le \tau$ with $\text{TV}(\pi_{\text{te}}, \pi_\theta) \le \delta(\tau)$ where $\delta(\tau) \to 0$ as $\tau \to 0$, the full gradient g and sparse gradient $\hat{g}$ (computed on Zones B∪C only) satisfy:*

$$\|g - (1-s)\hat{g}\| \le s\,G\big(\sqrt{2\tau} + 2\,\delta(\tau) + \sqrt{\tau/2}\,\log|\mathcal{V}|\big) \quad (6)$$

where $s = |\mathcal{A}|/N < 1$ is the skip fraction (N = sequence length), G bounds score function norms, and $|\mathcal{V}|$ is the vocabulary size.

The bound is $O(s\,G\sqrt{\tau}\,\log|\mathcal{V}|)$ and vanishes as $\tau \to 0$: Zone A tokens are selected precisely for near-zero entropy, making their gradient contributions negligible. In practice, $\log|\mathcal{V}| \approx 11.5$ is a moderate constant. The full proof (decomposing into FKL and RKL gradient bounds separately) is provided in Appendix D.

Remark 2.4 (Implementation scaling). In practice (Eq. 5), we normalize by $|\mathcal{B}\cup\mathcal{C}|$ rather than the full sequence length N, implicitly rescaling the gradient by $1/(1-s)$. This is equivalent to using an effective learning rate $\eta_{\text{eff}} = \eta/(1-s)$ on the active tokens, which we absorb into learning rate tuning. The directional guarantee of Theorem 2.3 is unaffected: the sparse update follows the same descent direction as the dense baseline.

Connection to temporal scheduling. As the student improves, Zone A expands (more tokens become deterministic for both models) and the active set shifts toward Zone B (teacher confident, student still catching up). The optimal FKL/RKL ratio among active tokens thus evolves naturally, motivating a temporal schedule.

2.4. Temporal: Competence-Driven Divergence Annealing

As the student masters more tokens, the active token composition shifts: early in training, Zone C (both models uncertain—exploration phase) is large, calling for mode-covering FKL to preserve diverse reasoning paths; late in training, Zone B (teacher confident, student still uncertain—refinement phase) dominates, calling for mode-seeking RKL to sharpen outputs. The annealing schedule tracks this natural evolution:

$$\alpha(t_{\text{step}}) = \frac{\alpha_{\text{start}}}{2}\left(1 + \cos\left(\frac{t_{\text{step}}}{T_{\text{total}}} \cdot \pi\right)\right) \quad (7)$$

decreasing smoothly from $\alpha_{\text{start}} = 0.8$ to $\alpha_{\text{end}} = 0.0$. Combined with SEAD zones, α modulates Zone C's FKL weight in Eq. 4, implementing a continuous exploration→refinement transition.

This generalizes the discrete two-stage switches of Ko et al. (2026) (who manually switch from FKL to RKL at a fixed step) and Xu et al. (2026a) (who find forward-then-reverse optimal but use a hard switch). By making the transition continuous and driven by the natural zone composition evolution, SEAD avoids the sensitivity to switch-point selection that plagues discrete approaches.

Temporal annealing assumes monotonically improving competence—an assumption that can be violated under uniform prompt sampling (hard prompts can degrade student

coherence). The following curriculum ensures this precondition holds.

2.5. Prompt-Level: Competence-Gated Curriculum

The deepest manifestation of the supervision quality principle: on prompts far beyond the student's capability, $p_i(\theta) \approx 0$ implies $\phi(p_i(\theta)) \approx 0$ (Assumption 2.1), so the modulated PL condition yields *vanishing* gradient signal. The student's rollouts are incoherent—producing garbled reasoning chains—and the teacher's token-level corrections on these incoherent contexts amount to noise rather than useful supervision. This degradation is unique to OPD: in off-policy KD the teacher always generates its own coherent traces ($\phi \equiv 1$), and in RL binary correctness rewards are equally reliable regardless of rollout quality.

The curriculum keeps $\phi(p_i(\theta))$ bounded away from zero on all eligible prompts, serving as a precondition for both Theorem 2.3 (the entropy landscape must be well-behaved for Zone A tokens to genuinely have low entropy) and temporal annealing (competence must grow monotonically for the schedule to track the correct FKL/RKL ratio).

Difficulty estimation. Before training begins, we sample K=8 rollouts per prompt from π_{θ_0} and compute difficulty $d_i = 1 - p_i$ where $p_i = \frac{1}{K}\sum_k \mathbf{1}[\text{correct}(\mathbf{x}_i^{(k)})]$ is the pass rate. Pass rate is a binary discretization of rollout quality, maintaining the entropy-as-unified-signal principle: a prompt the student never solves has maximally uncertain/incoherent rollouts.

Competence-based progression. Following Platanios et al. (2019), a competence function $c(t) = \min(1, c_0(1 + t/T)^p)$ grows monotonically, with eligible set $\mathcal{D}(t) = \{\mathbf{q}_i : d_i \leq c(t)\}$. In practice, we sort prompts by difficulty and present them easy-to-hard in a single pass without replacement, with small Gaussian noise ($\sigma = 0.02$) added for stochastic tie-breaking.

Theorem 2.5 (Curriculum convergence speedup)**.** *Under Assumption 2.1, β-smoothness of the loss, and bounded gradient variance σ^2, the curriculum achieves ϵ-accuracy in $T^{\text{C}} = \mathcal{O}(\beta\sigma^2/((c\mu)^2\epsilon))$ steps, versus $T^{\text{U}} \geq \Omega(\beta\sigma^2/((\bar{\phi}^{\text{U}}\mu)^2\epsilon))$ for uniform sampling, where $\bar{\phi}^{\text{U}} = \frac{1}{n}\sum_i \phi(p_i(\theta))$ is the average modulation under uniform sampling. This yields a speedup of $(c/\bar{\phi}^{\text{U}})^2$—**a separation unique to OPD** and absent in off-policy KD ($\phi \equiv 1$) or RL (unbiased gradients regardless of competence).*

The proof (Appendix E) shows that curriculum restricts training to prompts where the effective PL constant is at least $c\mu$, versus uniform sampling where hard prompts dilute the average to $\bar{\phi}^{\text{U}}\mu \ll c\mu$.

Corollary 2.6 (Mutual reinforcement)**.** *The three SEAD components form a virtuous cycle resolved by staged bootstrapping: (1) curriculum restricts to prompts with $\phi(p_i) \geq c > 0$, ensuring coherent rollouts and a reliable entropy landscape; (2) given well-behaved entropy, Theorem 2.3 guarantees sparse selection introduces negligible error; (3) with monotonic competence growth and predictable zone shifts, the cosine schedule correctly tracks exploration→refinement.*

3. Experiments

3.1. Setup

Models. We evaluate on two model families to demonstrate generalizability: **OLMo-3** (7B-Instruct-SFT student, 32B-Instruct teacher) and **Nemotron** (Nano-8B student, Super-49B teacher). The OLMo pair tests distillation across a large capability gap (17.2 points average), while Nemotron tests the challenging small-gap regime (2.2 points).

Training. All models are trained on 4 nodes of 8×H100 80GB GPUs (32 GPUs total). We use a rigorously deduplicated version of the DAPO-Math-17K dataset (Yu et al., 2025). Each iteration samples B=96 prompts with one rollout per prompt (temperature 0.7, max 16,384 tokens), followed by K=1 gradient update with clip ϵ=0.2 and peak learning rate 5×10^{-9} with linear warmup and decay. SEAD zone percentiles are $(\rho_A, \rho_B, \rho_C) = (50, 40, 10)$; annealing follows Eq. 7 with α_{start}=0.8. Full details in Appendix C.

Evaluation. We report greedy Pass@1 accuracy on MATH-500 (Hendrycks et al., 2021), Minerva-Math, and OlympiadBench (standard reasoning). For competition-level mathematics, we report Pass@32 (average over 32 sampled trajectories) on AMC 2023 and AIME 2024/2025.

Baselines. GRPO (Shao et al., 2024), Vanilla OPD with full-token RKL (Lu & Thinking Machines Lab, 2025), and OPSD (on-policy self-distillation) (Zhao et al., 2026).

3.2. Main Results

The distillation challenge. Table 1 demonstrates that bridging the massive 17.2-point performance gap between the OLMo teacher and student is non-trivial. While the teacher establishes a strong upper bound (75.4 average), standard baselines fail to meaningfully improve upon the base student: OPSD and GRPO stagnate near 58.2, while vanilla OPD (full-token RKL) yields only +1.0.

Synergistic component interaction. The critical finding is the *super-additive synergy* between components. KL Annealing alone provides +0.2 average; Token Zones alone provide +0.3. These modest individual gains reflect the fact that each component alone lacks necessary preconditions—annealing without zone structure applies scheduling uni-

Table 1. Results on OLMo-3 (7B student → 32B teacher). SEAD achieves +4.8 avg over vanilla OPD, with the largest gains on competition-level benchmarks (AIME, AMC).

Method	M500	Min.	Oly.	AMC	AI24	AI25	Avg.
Off-the-shelf models							
Student (7B)	87.2	41.9	58.2	80.4	46.8	34.6	58.2
Teacher (32B)	94.0	62.9	68.1	97.7	69.2	60.6	75.4
Baselines							
OPSD	86.4	43.8	57.8	80.2	44.0	36.8	58.2
GRPO	84.8	43.0	58.8	81.0	45.0	35.5	58.0
OPD (RKL)	87.0	41.9	60.1	80.9	47.9	37.3	59.2
Proposed (SEAD components)							
Anneal only	87.8	43.4	58.5	82.9	48.2	35.6	59.4
Zones only	87.0	41.9	60.1	80.9	47.9	39.3	59.5
Zones+Anneal	91.0	43.0	**62.5**	89.1	**54.5**	42.1	63.7
SEAD (full)	**91.2**	**44.5**	62.1	**89.8**	53.9	**42.5**	**64.0**

formly across informative and redundant tokens alike, while zones without annealing lock in a static FKL/RKL ratio that becomes suboptimal as training progresses. Their combination produces a +4.5 jump to 63.7—a 20× amplification—because zones concentrate gradient on informative tokens while annealing ensures the divergence applied to those tokens evolves appropriately with competence.

Full SEAD performance. Adding the competence-gated curriculum further pushes the average to 64.0 (+4.8 over vanilla OPD). The gains are most pronounced on competition-level mathematics: +7.7 on AIME24 and +7.9 on AIME25 versus the student, domains where reasoning depth and coherence matter most.

Nemotron generalization. SEAD also generalizes to the small-gap regime: on Nemotron (8B→49B, only 2.2-point gap), KL Annealing alone achieves +1.1 average, confirming that divergence scheduling helps even when the student is already strong (Table 3 in Appendix A).

3.3. Ablation Study

Table 2 reveals three key findings that validate the symbiotic design:

(1) Curriculum is the strongest individual factor (+4.20). This confirms that prompt ordering—specifically, restricting to prompts where the student can produce coherent rollouts—is the most impactful single intervention. It directly validates the supervision quality principle: OPD supervision on incoherent rollouts is ineffective.

(2) T+A produces 20× amplification over A alone. Annealing in isolation yields only +0.22 (and actually *hurts* AIME25 by −1.7). But when combined with token zones, the pair jumps to +4.52. This demonstrates that zone-aware

Table 2. 2^3 factorial ablation on OLMo-3. T = Token SEAD zones, A = KL Annealing, C = Curriculum. The AIME25Δ column highlights the supervision quality cascade on the hardest benchmark. Super-additive interactions confirm the symbiotic design.

T	A	C	Avg.	Δ vs. OPD	AI25Δ
			59.18	+0.00	+0.0
✓			59.50	+0.32	+2.0
	✓		59.40	+0.22	−1.7
		✓	63.38	+4.20	+4.6
✓	✓		63.70	+4.52	+4.8
✓		✓	63.42	+4.24	+3.4
	✓	✓	63.97	+4.79	+4.5
✓	✓	✓	**64.00**	**+4.82**	**+5.2**

sparsity *requires* a matching temporal schedule: zones identify *where* to apply gradient, while annealing controls *what type* of gradient—neither is effective alone.

(3) Three-way combination is best, with orthogonal contributions. The full T+A+C achieves the highest result (64.00), with each pairwise combination providing complementary gains. Zones target *which* tokens receive gradient, annealing controls *when* the FKL/RKL balance shifts, and curriculum determines *what* prompts are eligible—three orthogonal axes of the same supervision-quality principle.

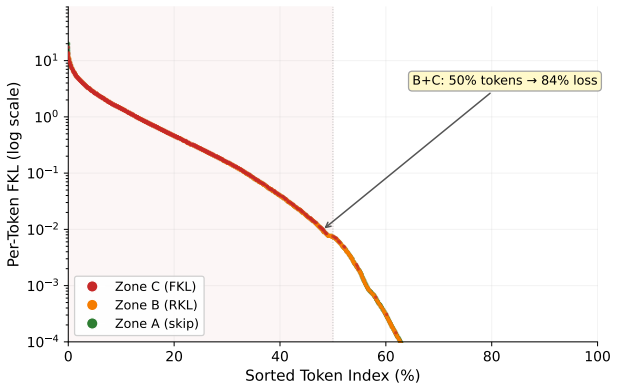
4. Analysis

4.1. Token-Level Zone Analysis

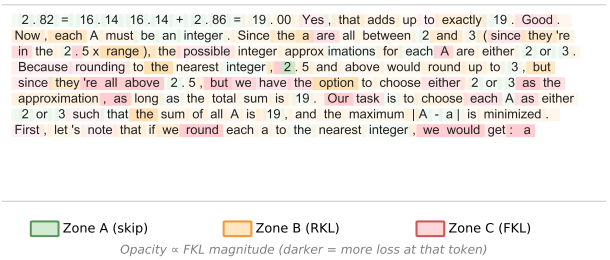
We conduct detailed analysis on OLMo-3 7B/32B at training step 40 (50/40/10 zone configuration, 163 evaluation sequences) to understand what SEAD's entropy-based zones capture in mathematical reasoning.

Entropy distribution and zone semantics. Zone A tokens overwhelmingly correspond to deterministic computation steps and syntactic connectives: 73.2% of computation tokens and 54.2% of syntax tokens are assigned to Zone A. Zone C tokens appear at genuine reasoning forks: 63.9% are *strategy fork* positions where both teacher and student entropy exceed the 75th percentile. Crucially, the 50/40/10 configuration routes *zero* strategy fork tokens to Zone A, ensuring every reasoning-critical position receives active supervision (Table 4 in Appendix B).

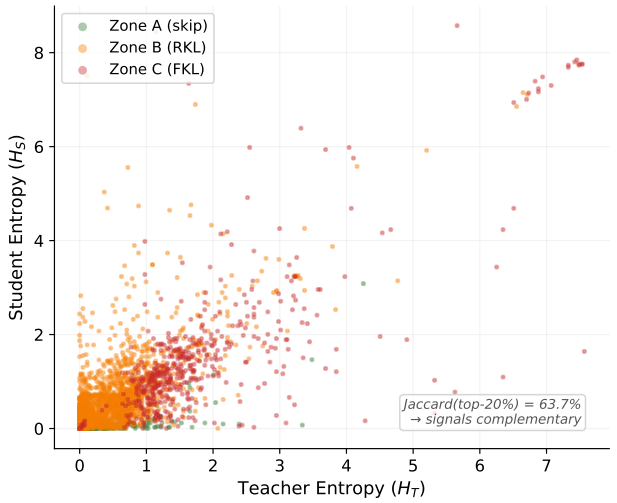
Signal complementarity justifies joint conditioning. With a Jaccard index of 63.7% between the top-20% teacher entropy and top-20% student entropy positions (Figure 2c), approximately 36% of high-entropy tokens are unique to one signal. Teacher entropy identifies positions where guidance is most valuable (multiple valid continuations); student entropy identifies positions already mastered. Neither signal alone captures the complete picture, justifying SEAD's joint conditioning over approaches like EOPD (Jin et al., 2026)



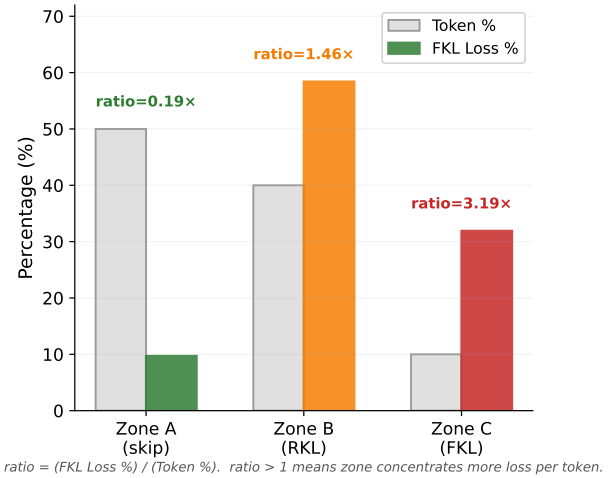
(a) Sorted per-token FKL loss by zone. The top 50% of tokens (Zones B+C) concentrate 90% of total loss, while the bottom 50% (Zone A) carry only 10%—confirming the heavy-tailed loss distribution.



(b) Token-level zone visualization on OLMo-3-7B output. Background color indicates zone; opacity scales with FKL magnitude. Zone C (red) clusters at strategy forks; Zone A (green) covers predictable arithmetic.



(c) Zone assignment in joint entropy space (H_{te} vs. H_θ). Jaccard overlap of top-20% sets is 63.7%, confirming that teacher and student entropy provide partially complementary signals.



(d) Loss concentration by zone. Zone C (10% of tokens) carries 31.9% of FKL loss (3.19× concentration); Zone B+C together achieve 1.81× over uniform allocation.

Figure 2. **SEAD zone analysis on OLMo-3-7B / 32B (50/40/10 configuration).** (a) Sorted FKL reveals a heavy-tailed distribution where 50% of tokens dominate gradient signal. (b) Zone coloring on model output shows Zone C tokens align with reasoning decision points in mathematical proofs. (c) Joint entropy space confirms neither teacher nor student signal alone captures the full picture. (d) Loss concentration validates that Zone A tokens are safe to skip (0.19× ratio) while Zone C tokens concentrate 3.19× more loss per token, making them the most valuable supervision targets.

that use only teacher entropy.

Gradient concentration validates sparsity. Figure 2d quantifies the efficiency of zone-based selection: Zone A (50% of tokens) carries only 9.7% of total FKL loss, while Zone B+C concentrates 90.3% of gradient signal. The per-zone loss ratios are: Zone C at 3.19× its fair share, Zone B at 1.46×, Zone A at only 0.19×. This validates Theorem 2.3: skipping Zone A discards half the tokens while retaining over 90% of the training signal.

4.2. Training Dynamics: Entropy Collapse Prevention

A known failure mode of vanilla OPD is premature entropy collapse: if the student is trained with reverse KL from the beginning, it concentrates probability mass on the teacher's peak mode before exploring alternative valid reasoning paths. This is particularly damaging for mathematical reasoning, where multiple valid proof strategies may exist for a single problem.

SEAD effectively mitigates this collapse through the interaction of token selection and KL annealing. Figure 3

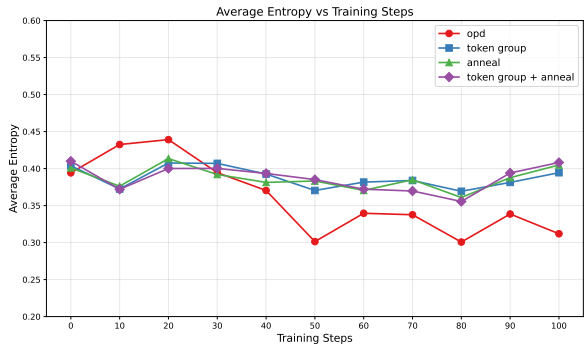


Figure 3. Average token-level entropy during training. Vanilla OPD (red) suffers progressive entropy collapse ($0.39 \rightarrow 0.31$), while SEAD variants maintain stable, consistently higher entropy throughout, preventing premature mode collapse and preserving reasoning diversity.

shows that vanilla OPD exhibits a clear decreasing trend in average token-level entropy ($0.39 \rightarrow 0.31$), whereas all SEAD-related methods maintain stable, higher entropy levels. The mechanism is: early in training, the coefficient α is high, so Zone C tokens receive forward-KL supervision that explicitly preserves the teacher's multi-modal distribution. Later, α decays, shifting toward RKL refinement—but only after the student has sufficiently explored alternative strategies. The key distinction from global annealing is that SEAD applies this only to the informative subset: Zone A tokens (50%) are skipped regardless, so the annealing operates specifically on tokens where divergence choice actually matters.

5. Related Work

On-Policy Distillation for Math Reasoning. GKD (Agarwal et al., 2024) established on-policy distillation for language models, showing that training on student-generated rollouts reduces exposure bias. Extensions include RKL-based reasoning distillation (Lu & Thinking Machines Lab, 2025), on-policy self-distillation (OPSD, Zhao et al., 2026), theoretical unification with dense KL-constrained RL (G-OPD, Yang et al., 2026), and scaling to large model pairs (Ko et al., 2026; Wu et al., 2026). These methods all apply supervision uniformly across tokens, phases, and prompts.

Token-Level Selection. EOPD (Jin et al., 2026) switches between FKL and RKL based on teacher entropy alone, but does not use student entropy for selection and applies a single threshold. TIP (Xu et al., 2026b) computes token importance as student entropy $\times$ divergence for soft weighting, but uses only a single divergence type. SE-KD (Tavor et al., 2026) and SCOPE (Zheng et al., 2026) explore other selection criteria. SEAD differs fundamentally via: (i) *joint* teacher-student entropy partitioning (capturing the complementary 36% of tokens unique to each signal), and (ii) per-zone divergence switching (Zone B gets RKL, Zone C gets FKL, not just weighted versions of a single loss).

Divergence Scheduling. Ko et al. (2026) use a discrete two-stage schedule (FKL then RKL); PACED (Xu et al., 2026a) empirically confirms forward-then-reverse is optimal; DRKL (Luong et al., 2026) addresses entropy collapse via diversity-aware reverse KL. Our continuous cosine annealing generalizes these discrete approaches and is motivated by the natural evolution of zone composition rather than manual schedule design.

Curriculum for Reasoning. Curriculum learning has been studied for RL-based reasoning (Parashar et al., 2025; Wen et al., 2025), but **no prior work applies curriculum to OPD**, despite it being explicitly flagged as an open problem by Zhao et al. (2026) and Li et al. (2026). Our contribution demonstrates both *why* curriculum is uniquely necessary for OPD (the modulated PL condition formalizes supervision quality degradation absent in other paradigms) and *how* it enables the other SEAD components (providing the coherent rollouts and monotonic competence growth they require).

6. Conclusion

We identified a fundamental property of on-policy distillation: teacher supervision quality depends on student competence. SEAD addresses this supervision-quality bottleneck at three scales using entropy as a unified signal. At the token level, it skips low-value tokens and applies adaptive divergences to informative ones; at the temporal level, it anneals from exploration to refinement as competence grows; at the prompt level, a competence-gated curriculum ensures coherent rollouts and reliable entropy estimates. Together, these components form a mutually reinforcing framework—not independent heuristics—as confirmed by the $20\times$ super-additive interaction between zones and annealing.

On mathematical reasoning benchmarks, SEAD achieves +4.8 average accuracy over vanilla OPD on OLMo-3 (7B$\rightarrow$32B) and generalizes to the Nemotron family, with the largest improvements on competition-level problems (AIME, AMC) where reasoning depth and multi-step coherence matter most.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Limitations. The curriculum relies on static difficulty scores computed before training; adaptive re-estimation could improve sample efficiency but adds computational overhead. Our evaluation focuses on mathematical reasoning; generalization to code and open-domain tasks remains future work. Finally, SEAD's behavior on very long reasoning chains (>8K tokens) requires further study.

References

Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos, S., Geist, M., and Bachem, O. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS*, volume 34, pp. 21163–21174, 2021.

Jin, W., Min, T., Yang, Y., Kadhe, S. R., Zhou, Y., Wei, D., Baracaldo, N., and Lee, K. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Ko, J., Abdali, S., Kim, Y. J., Chen, T., and Cameron, P. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

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

Lu, K. and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.

Luong, H.-C. et al. Diversity-aware reverse Kullback-Leibler divergence for large language model distillation. *arXiv preprint arXiv:2604.00223*, 2026.

Parashar, S., Gui, S., Li, X., Ling, H., Vemuri, S., Olson, B., Li, E., Zhang, Y., Caverlee, J., Kalathil, D., et al. Curriculum reinforcement learning from easy to hard tasks improves LLM reasoning. *arXiv preprint arXiv:2506.06632*, 2025.

Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B., and Mitchell, T. Competence-based curriculum learning for neural machine translation. In *Proceedings of NAACL-HLT*, pp. 1162–1172, 2019.

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

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

Tavor, A., Ebenspanger, I., Cnaan, N., and Geva, M. Rethinking selective knowledge distillation. *arXiv preprint arXiv:2602.01395*, 2026.

Wen, L., Cai, Y., Xiao, F., He, X., An, Q., Duan, Z., Du, Y., Liu, J., et al. Light-r1: Curriculum SFT, DPO and RL for long CoT from scratch and beyond. In *Proceedings of ACL*, pp. 318–327, 2025.

Wu, Y. et al. Lightning OPD: Efficient post-training for large reasoning models with offline on-policy distillation. *arXiv preprint arXiv:2604.13010*, 2026.

Xu, Y., Sang, H., Zhou, Z., He, R., and Wang, Z. PACED: Distillation at the frontier of student competence. *arXiv preprint arXiv:2603.11178*, 2026a.

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

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

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

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

Zheng, B., Ma, X., Liang, Y., Ruan, J., Fu, X., Lin, K., Zhu, B., Zeng, K., and Cai, X. SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

Table 3. Results on Nemotron (8B student → 49B teacher). Even with only a 2.2-point teacher-student gap, SEAD components provide meaningful gains.

Method	M500	Min.	Oly.	AMC	AI24	AI25	Avg.
Off-the-shelf models							
Student (8B)	91.2	52.9	61.0	93.3	61.0	50.8	68.3
Teacher (49B)	94.2	57.4	63.9	94.2	60.4	52.9	70.5
Baselines							
OPSD	90.8	49.6	61.9	93.4	63.2	48.9	67.9
GRPO	92.6	53.7	60.9	94.6	64.0	50.2	69.3
OPD (RKL)	89.8	53.3	63.3	93.9	63.0	49.8	68.8
Proposed (SEAD components)							
Anneal only	**93.4**	**56.6**	62.5	94.3	62.7	50.4	**69.9**
Zones only	92.0	54.0	62.1	93.8	63.1	51.5	69.4
Zones+Anneal	90.4	54.8	61.6	**94.7**	**64.6**	**51.0**	69.5

Table 4. P(zone | category) for the 50/40/10 config. Zero strategy forks are assigned to Zone A, ensuring all reasoning-critical tokens receive supervision.

Category	Zone A	Zone B	Zone C	Total
Deterministic syntax	54.2%	39.1%	6.7%	4,777
Computation	73.2%	21.8%	5.0%	2,564
Reasoning transition	10.8%	59.1%	30.0%	536
Strategy fork	**0.0%**	52.7%	**47.3%**	2,968
Other	57.4%	41.1%	1.6%	11,696

A. Additional Experimental Results

B. Token Zone Category Breakdown

C. Training Details

During each iteration, we sample $B = 96$ prompts and generate one rollout per prompt using a sampling temperature of 0.7 and a maximum generation length of 16,384 tokens. Following teacher scoring, we perform $K = 1$ gradient update with a clipping parameter $\epsilon = 0.2$. We use a peak learning rate of 5×10^{-9} with a linear warmup and decay schedule over a single training epoch.

For the SEAD configuration, the zone percentiles are set to $\rho_A = 50$, $\rho_B = 40$, and $\rho_C = 10$. The FKL balancing hyperparameter λ is set to 1.0. The annealing factor follows a cosine schedule (Eq. 7), decaying smoothly from $\alpha_{\text{start}} = 0.8$ to $\alpha_{\text{end}} = 0.0$.

For curriculum learning, we adopt a smooth competence-based ordering inspired by Platanios et al. (2019). We first estimate per-problem difficulty as $d_i = 1 - \hat{a}_i$, where $\hat{a}_i$ is the pass rate computed from $K = 8$ independent rollouts of the initial student model. Examples are then sorted by difficulty in ascending order and presented easy-to-hard in a single pass without replacement, with small Gaussian noise ($\sigma = 0.02$) added to difficulty scores for stochastic tie-breaking. This provides a smooth, continuous curriculum that naturally increases problem difficulty as training progresses, without requiring discrete phase boundaries or explicit competence thresholds.

Teacher logits are computed using the top-k vocabulary subset ($k = 50$) for entropy approximation, reducing memory overhead while maintaining high correlation ($r > 0.99$) with full-vocabulary entropy. Zone assignments are computed per-batch using percentile thresholds on the joint $(H_{\text{te}}, H_\theta)$ distribution within each batch, ensuring adaptive thresholds that account for batch-to-batch entropy variation.

D. Proof of Theorem 2.3 (Sparse Gradient Approximation)

We first state the full technical assumptions, then provide the proof.

Assumption D.1 (Bounded score functions)**.** There exists a constant $G > 0$ such that for every token position t and vocabulary element $v \in \mathcal{V}$, $\|\nabla_\theta \log \pi_\theta(v \mid \mathbf{c}_t)\| \le G$.

Assumption D.2 (Low-entropy agreement in Zone A)**.** For every token position $t \in \mathcal{A}$, the teacher and student distributions satisfy $H_{\text{te}}(t) \le \tau$ and $H_\theta(t) \le \tau$, and $\mathrm{TV}(\pi_{\text{te}}(\cdot \mid \mathbf{c}_t), \pi_\theta(\cdot \mid \mathbf{c}_t)) \le \delta(\tau)$, where $\delta(\tau) \to 0$ as $\tau \to 0$. A sufficient condition is that both distributions place mass $\ge 1 - \sqrt{\tau/2}$ on a common mode $v_t^\star$.

Assumption D.3 (Bounded log-density ratio)**.** For all $v \in \mathcal{V}$ and token positions t, $|\log \pi_\theta(v \mid \mathbf{c}_t) - \log \pi_{\text{te}}(v \mid \mathbf{c}_t)| \le B_{\max}$. In Zone A, the low-entropy agreement guarantees: for the dominant mode $v_t^\star$, $|\log \pi_\theta(v_t^\star) - \log \pi_{\text{te}}(v_t^\star)| \le \delta(\tau)/(1 - \sqrt{\tau/2})$; for off-mode tokens, $|\log \pi_\theta(v) - \log \pi_{\text{te}}(v)| \le \log |\mathcal{V}|$ trivially.

Proof. Write the full gradient as a convex decomposition over zones:

$$g = \frac{|\mathcal{A}|}{N}\, \bar{g}_{\mathcal{A}} + \frac{|\mathcal{B} \cup \mathcal{C}|}{N}\, \hat{g},$$

where $\bar{g}_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \nabla_\theta \mathcal{L}(t)$. Since $|\mathcal{B} \cup \mathcal{C}|/N = 1 - s$:

$$g - (1 - s)\, \hat{g} = s\, \bar{g}_{\mathcal{A}}.$$

It suffices to bound $\|\nabla_\theta \mathcal{L}(t)\|$ for each $t \in \mathcal{A}$.

Forward KL gradient at a Zone A token. For forward KL, $\nabla_\theta L_{\text{FKL}}(t) = -\sum_{v \in \mathcal{V}} \pi_{\text{te}}(v \mid \mathbf{c}_t) \nabla_\theta \log \pi_\theta(v \mid \mathbf{c}_t)$. Using the score-function identity $\sum_v \pi_\theta(v \mid \mathbf{c}_t) \nabla_\theta \log \pi_\theta(v \mid \mathbf{c}_t) = \mathbf{0}$:

$$\nabla_\theta L_{\text{FKL}}(t) = -\sum_v \big[\pi_{\text{te}}(v \mid \mathbf{c}_t) - \pi_\theta(v \mid \mathbf{c}_t)\big] \nabla_\theta \log \pi_\theta(v \mid \mathbf{c}_t). \tag{8}$$

When $H_{\text{te}}(t) \le \tau$, the teacher places mass $\ge 1 - \sqrt{\tau/2}$ on a single token $v^\star$. Similarly for the student. By Assumption D.1: $\|\nabla_\theta L_{\text{FKL}}(t)\| \le G(\sqrt{2\tau} + 2\,\delta(\tau))$.

Reverse KL gradient at a Zone A token. The gradient of $L_{\mathrm{RKL}}(t) = \mathrm{KL}(\pi_\theta \| \pi_{\mathrm{te}})$ at position t is:

$$\nabla_\theta L_{\mathrm{RKL}}(t) = \sum_v \pi_\theta(v \mid \mathbf{c}_t)\big[\log \pi_\theta(v) - \log \pi_{\mathrm{te}}(v)\big]\, \nabla_\theta \log \pi_\theta(v \mid \mathbf{c}_t), \tag{9}$$

using the score function identity to eliminate the $+1$ term. In Zone A, splitting into on-mode ($v = v^\star$: mass $\geq 1 - \sqrt{\tau/2}$, log-ratio $\leq \delta(\tau)/(1 - \sqrt{\tau/2})$) and off-mode ($v \neq v^\star$: total mass $\leq \sqrt{\tau/2}$, log-ratio $\leq \log|\mathcal{V}|$):

$$\|\nabla_\theta L_{\mathrm{RKL}}(t)\| \leq G\big(\delta(\tau) + \sqrt{\tau/2}\, \log|\mathcal{V}|\big). \tag{10}$$

Aggregation. Taking the maximum of FKL and RKL bounds: $\|\bar{g}_{\mathcal{A}}\| \leq G(\sqrt{2\tau} + 2\,\delta(\tau) + \sqrt{\tau/2}\, \log|\mathcal{V}|)$. Therefore:

$$\|g - (1-s)\hat{g}\| = s\|\bar{g}_{\mathcal{A}}\| \leq s\, G\big(\sqrt{2\tau} + 2\,\delta(\tau) + \sqrt{\tau/2}\, \log|\mathcal{V}|\big) = O(s\, G\sqrt{\tau}\, \log|\mathcal{V}|).$$

In practice $\log|\mathcal{V}| = O(\log 10^5) \approx 11.5$ is a moderate constant, so the bound simplifies to $O(s\, G\sqrt{\tau})$ up to logarithmic factors. □

E. Proof of Theorem 2.5 (Curriculum Convergence Speedup)

The analysis requires the following additional standard assumptions beyond Assumption 2.1:

Assumption E.1 (Smoothness)**.** Each per-prompt loss ℓ_i is β-smooth: $\|\nabla \ell_i(\theta) - \nabla \ell_i(\theta')\| \leq \beta\|\theta - \theta'\|$.

Assumption E.2 (Bounded stochastic gradients)**.** The stochastic gradient satisfies $\mathbb{E}[g_i(\theta)] = \nabla \ell_i(\theta)$ and $\mathbb{E}\|g_i(\theta) - \nabla \ell_i(\theta)\|^2 \leq \sigma^2$.

Assumption E.3 (Stage mastery)**.** For each curriculum stage, after sufficient SGD steps on the eligible set, the student achieves $p_i(\theta) \geq \alpha_{\min}$ for all eligible prompts, where $\phi(\alpha_{\min}) \geq c > 0$.

Proof. **Step 1: Per-step descent.** By β-smoothness and step size $\eta \leq 1/\beta$:

$$\mathbb{E}_g[\ell_i(\theta_t) - \ell_i(\theta_{t+1})] \geq \frac{\eta}{2}\|\nabla \ell_i(\theta_t)\|^2 - \frac{\eta^2 \beta \sigma^2}{2}. \tag{11}$$

Step 2: Modulated PL under curriculum. Under curriculum, all eligible prompts satisfy $\phi(p_i(\theta)) \geq c$ (Assumption E.3). By Assumption 2.1:

$$\|\nabla \ell_i(\theta_t)\|^2 \geq c\,\mu\,(\ell_i(\theta_t) - \ell_i^\star)$$

for all $i \in \mathcal{D}(t)$. Averaging over $i \sim \mathrm{Uniform}(\mathcal{D})$ yields a standard PL descent with effective constant $\mu_{\mathrm{eff}} = c\mu$.

Step 3: Convergence rate under curriculum. Applying the SGD convergence rate under PL (Karimi et al., 2016) with optimal step size:

$$T^{\mathrm{C}} = O\left(\frac{\beta\sigma^2}{(c\,\mu)^2\,\epsilon}\right).$$

Step 4: Uniform analysis. Under uniform sampling over all n prompts, define $\bar{\phi}^{\mathrm{U}} = \frac{1}{n}\sum_{i=1}^n \phi(p_i(\theta))$. Since competence $\phi(p_i)$ and suboptimality $(\ell_i - \ell_i^\star)$ are negatively correlated (easy prompts have high ϕ but low loss), by Chebyshev's sum inequality: $\mathbb{E}_i[\phi(p_i)\,(\ell_i - \ell_i^\star)] \leq \bar{\phi}^{\mathrm{U}} \cdot (L(\theta) - L^\star)$. The effective PL constant under uniform sampling is at most $\bar{\phi}^{\mathrm{U}}\mu$, giving:

$$T^{\mathrm{U}} \geq \Omega\left(\frac{\beta\sigma^2}{(\bar{\phi}^{\mathrm{U}}\,\mu)^2\,\epsilon}\right).$$

Step 5: Speedup. The ratio $T^{\mathrm{U}}/T^{\mathrm{C}} \geq (c/\bar{\phi}^{\mathrm{U}})^2$. Since hard prompts contribute $\phi \approx 0$ early in training, $\bar{\phi}^{\mathrm{U}} \ll c$, giving meaningful speedup.

Step 6: OPD specificity. This separation is *unique to OPD*: in off-policy KD, $\phi \equiv 1$ (teacher traces always coherent), so curriculum provides no advantage. In RL, binary rewards give unbiased gradients regardless of rollout quality. Only in OPD is the gradient *norm* modulated by competence (teacher logits on incoherent rollouts are biased, not just noisy), making curriculum a convergence accelerator rather than mere sample reordering. □