# DOSD: DIFFICULTY-PRIVILEGED ON-POLICY SELF-DISTILLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy self-distillation trains an unassisted student using a reference-conditioned teacher evaluated on the student's own trajectories. Reference access can supply useful supervision, but can also make that supervision depend on information unavailable to the student. We introduce Difficulty-privileged On-policy Self-Distillation (DOSD), which allocates teacher-side reference context using a per-problem rolling record of verified outcomes. An offline-initialized, fixed-length queue is updated with each current student response before the teacher scores that same response. Its success fraction determines the token fraction of a solution prefix exposed to the teacher, while the final answer remains separately available to the teacher and all privileged information remains hidden from the student. We formalize this training procedure, establish bounded changes in its exposure budget, and characterize the additional outcome-conditioned target variability introduced by immediate feedback. These properties motivate controlled tests of adaptive exposure; they do not establish a performance improvement or the elimination of privileged-information dependence. Across Qwen3-1.7B, 4B, and 8B, DOSD changes three-benchmark average accuracy relative to OPSD by $+0.44$, $+0.36$, and $-0.78$ percentage points, respectively. In a Qwen3-8B-thinking controller comparison, DOSD exceeds a reported budget-calibrated fixed-prefix controller by 2.32 points and is 0.18 points below full-prefix OPSD. Because the available experiment record contains aggregate point estimates but not repeated seeds, complete run metadata, or realized context budgets, these results are descriptive rather than evidence of statistical significance or reduced end-to-end cost.

## 1 INTRODUCTION

Reasoning models must ultimately solve problems without access to their solutions. During training, however, verified answers and reference derivations can provide substantially richer supervision than the model will receive at deployment. Reinforcement learning with verifiable rewards (RLVR) uses this asymmetry conservatively: a verifier evaluates a completed response, and its outcome drives policy improvement. Group Relative Policy Optimization (GRPO; Shao et al., 2024) obtains relative advantages from multiple responses, but a group of uniformly incorrect responses has no correctness-based contrast. On-policy distillation offers a complementary source of supervision: a teacher evaluates the student's own reasoning states and supplies token-level distributions, including on trajectories that end in failure (Agarwal et al., 2024). *The central question is how to obtain useful dense feedback without teaching the student to depend on information it will not observe.*

On-policy self-distillation (OPSD; Zhao et al., 2026) makes this question particularly concrete. A model acts as an unassisted student and as a reference-conditioned teacher. The teacher scores the student's trajectory rather than generating an additional target response, creating an informational advantage without requiring a separately trained, larger teacher. Yet access to a reference is not equivalent to a universally better teaching signal. A teacher may favor a particular derivation, use answer-conditioned shortcuts, or describe a reference that is absent from the student's prompt. Matching such preferences need not improve unassisted reasoning.

This distinction is supported by recent analyses. Yang et al. (2026a) report reference-dependent behavior and late-training degradation in their OPSD settings, and propose RLSD, which uses privileged policy differences to modulate verifier-derived advantages instead of directly matching



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the privileged distribution. Shen et al. (2026) separately isolate reference-induced shortcuts in self-distillation targets. These findings motivate examining *what enters the target*, not merely whether the teacher predicts tokens confidently. In particular, lower teacher–student divergence is not by itself evidence of better reasoning: it can reflect either improved learning or a less informative teacher.

The resulting design problem is a trade-off, rather than a mandate to remove reference context. Too much context may make a target specific to the supplied solution; too little may remove the very guidance that makes self-distillation useful. Adaptive teacher exposure already addresses this tension (Tan & Hong, 2026; Han et al., 2026; Yang et al., 2026b), so adaptivity alone is not our claimed distinction. Moreover, Harne et al. (2026) find that reducing privileged context can weaken useful supervision along with reference-induced bias. We therefore ask a narrower question: *can a lightweight record of each problem's verified outcomes allocate teacher context to where it is needed, without changing the student's rollout distribution?*

We propose **Difficulty-privileged On-policy Self-Distillation (DOSD)**. Each problem maintains a fixed-length queue of binary outcomes, initialized from offline unassisted rollouts. During training, the student generates one unassisted response; its verified outcome immediately replaces the oldest queue entry. The resulting success rate $k$ determines a teacher-side reference-prefix ratio $\max(0.1, 1 - k)$, measured in tokens. The teacher then scores *that same response* with the updated prefix hint. The hint excludes the final-answer segment, while the correct answer remains separately visible to the teacher. Neither channel is exposed to the student. Thus, outcome feedback controls the *information supplied to the teacher*, rather than acting as a group-relative advantage, a rollout hint, or a learned stopping reward.

The working hypothesis is that a problem with a stronger success history can be taught with less specific reference, whereas repeated failure warrants more. This is a testable allocation hypothesis, not an assumption that hard problems always benefit from longer hints. Outcomes are policy-dependent and historical; failures may also reflect verifier errors, insufficient exploration, or an unsuitable reference. Immediate feedback introduces an additional subtlety: although the student remains unassisted, its completed response influences its own teacher target through the hint length, and our analysis accounts for this retrospective dependence explicitly.

This work makes three methodological contributions. First, it specifies a per-problem, offline-initialized outcome memory with same-response teacher-context adaptation, requiring no extra online competence rollouts. Second, it separates reference-prefix exposure from the retained answer channel and formalizes the resulting on-policy distillation update. Third, it establishes bounded changes in the exposure budget and an exact forward-KL decomposition for outcome-conditioned targets. These results delimit what DOSD controls—teacher context allocation—and what remains an empirical question: whether that allocation improves the balance between helpful supervision, reference dependence, and total training cost.

## 2 Related Work

**On-policy and context-based distillation.** Context distillation transfers behavior elicited by richer training-time context into a model that receives less context at inference (Snell et al., 2022). Generalized knowledge distillation trains on student-generated sequences to address the mismatch between training states and the student's own generations (Agarwal et al., 2024). OPSD combines student rollouts with the same model under reference condition, and studies divergence choices and clipping (Zhao et al., 2026). DOSD retains this trajectory-scoring setup; its distinguishing component is the outcome-memory rule for teacher context, not self-distillation or student-generated training data themselves.

**Privileged-information dependence and corrective supervision.** RLSD retains the sign of verifier-derived scalar advantages while using stopped-gradient privileged evidence to change token-level magnitudes (Yang et al., 2026a). DOSD instead changes teacher inputs and retains distribution matching, so it does not inherit this advantage-sign anchoring. Purified OPSD uses a reference-only teacher to isolate a shortcut component and constructs a residual, PMI-based target (Shen et al., 2026). OGLS-SD uses outcome feedback to steer logits (Yang et al., 2026c); GATES uses tutor consensus when verified labels are unavailable (Stein et al., 2026). These target- or reliability-side interventions

Table 1: Where privileged or outcome-derived information enters training. "Student rollout" refers to generation, not subsequent scoring. The rows identify mechanisms, not a performance ranking.

| Method | Rollout | Control mechanism | Use of feedback |
|---|---|---|---|
| OPSD | Unassisted | Reference-conditioned teacher | Distribution target |
| RLSD | Unassisted | Privileged token evidence | Verifier advantage modulation |
| WS-GRPO | Unassisted | Learned prefix preferences | Shaped trajectory return |
| OC-GRPO | May be guided | Hint-assisted exploration | Context-corrected RL update |
| PAINT | Unassisted | Current rollout–reference overlap | Prefix masking / target interpolation |
| ATESD | Unassisted | Learned interval-level exposure | Teacher context allocation |
| USD | Unassisted | Global exposure / token controls | Supervision-load regulation |
| DOSD | Unassisted | Per-problem outcome FIFO | Immediate teacher context allocation |

differ from DOSD's exposure rule. Its answer-marker filter is neither a general behavioral purifier nor a proof of leakage-free learning.

**Adaptive teacher exposure.** PAINT already adjusts the visible solution prefix for the current rollout, using rollout–reference anchor overlap and sparse teacher-target interpolation (Tan & Hong, 2026). DOSD replaces reference-path alignment with a rolling record of verified outcomes: a correct alternative derivation need not resemble the supplied solution to increase its success statistic. ATESD learns an interval-level exposure controller from training-state features and delayed feedback (Han et al., 2026); USD adapts global exposure and token selection using learning load (Yang et al., 2026b). DOSD uses no learned controller, overlap score, or token-selection dual. Its problem-indexed FIFO and immediate correctness feedback distinguish it from these close comparisons, all of which are necessary baselines for evaluating its contribution.

**Privileged exploration and prefix-conditioned RL.** POPE uses privileged information to facilitate exploration on hard problems (Qu et al., 2026). PrefixRL reuses off-policy solution prefixes to obtain informative continuations, analyzing consistency under conditions on the prefixes (Setlur et al., 2026). Off-Context GRPO separates a privileged rollout policy from the unprivileged target policy and introduces importance correction for this contextual mismatch (Agrawal et al., 2026). Its adaptive variant increases hinting when unguided groups fail. In DOSD, only the *scoring teacher* receives a hint; the behavior prompt is always unprivileged. DOSD consequently has no hint-induced behavior/target context mismatch to correct, but also does not directly provide the guided exploration of these methods.

**Weak supervision for rollout efficiency.** WS-GRPO trains a preference model from outcome labels and evaluates successive reasoning prefixes to construct continuation-sensitive pseudo-rewards (Mundada et al., 2026). Those rewards are aggregated into a trajectory return before group-relative optimization. Its prefixes are *partial student trajectories*; DOSD's prefixes are *reference context for a teacher*. DOSD uses correctness to allocate that context, without training a preference model or imposing a stopping policy. Neither approach's efficiency can be assessed solely by counting online responses: their respective offline preparation and scoring costs must also be included.

# 3 Method

## 3.1 Preliminaries: On-policy self-distillation

On-policy self-distillation (OPSD; Zhao et al., 2026) starts from problem–solution pairs $\mathcal{S} = \{(x_i, r_i)\}_{i=1}^{M}$, where $x_i$ is a problem and $r_i$ is a reference solution or reasoning trace. Two conditional uses of the same language model define a student $p_S(\cdot \mid x)$, which receives only the problem, and a teacher $p_T(\cdot \mid x, r)$, which additionally receives the reference solution. The student generates a response $\hat{y} \sim p_S(\cdot \mid x)$. Given its prefix $\hat{y}_{<t}$, both policies output next-token distributions over the same vocabulary; the teacher scores the student trajectory without generating a separate response.

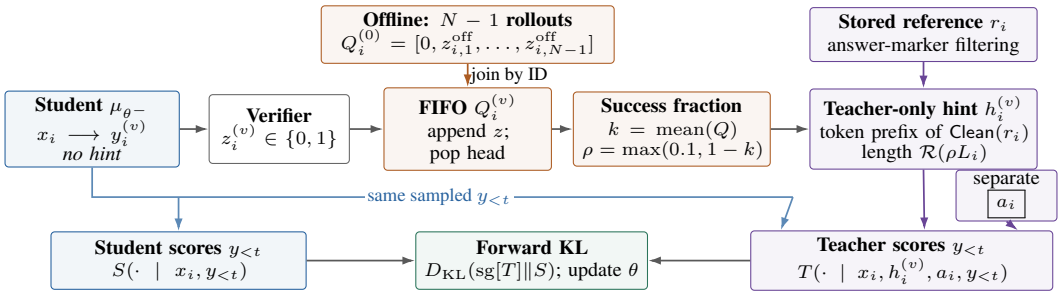



Figure 1: Overview of DOSD. Each unassisted rollout's verified outcome updates an offline-initialized FIFO before teacher scoring. The queue controls an answer-filtered token-prefix hint; a separate answer field is teacher-only. Student and teacher score the same rollout, and the forward-KL gradient updates only the student.

The OPSD objective is

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{(x,r)\sim\mathcal{S}} \mathbb{E}_{\hat{y}\sim p_S(\cdot|x)} \left[ \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} D(\text{sg}[p_T(\cdot \mid x, r, \hat{y}_{<t})] \,\|\, p_S(\cdot \mid x, \hat{y}_{<t})) \right]. \tag{1}$$

Here $D$ is a token-distribution divergence; gradients act on the student scores, with the teacher target and sampled trajectory held fixed for each update. OPSD admits different divergence and clipping choices. DOSD retains this scoring setup and changes how much reference reasoning enters the teacher context; the forward-KL instance analyzed below is one specific choice.

### 3.2 Problem Outcome Memory

Let $\mathcal{D} = \{(x_i, r_i, a_i)\}_{i=1}^{M}$ contain a question, a reference solution, and a verified final answer. The student receives only $x_i$ (with task instructions). A verifier $V(x_i, y, a_i) \in \{0, 1\}$ evaluates a completed response.

Before training, an unprivileged initialization policy produces $N-1$ responses per problem. Their outcomes initialize an ordered queue:

$$z_{i,j}^{\text{off}} = V(x_i, y_{i,j}^{\text{off}}, a_i), \qquad Q_i^{(0)} = (0, z_{i,1}^{\text{off}}, \ldots, z_{i,N-1}^{\text{off}}). \tag{2}$$

Generated responses and queues are stored separately. Training joins the original dataset and the queue artifact by $\text{id}_i$, rather than relying on row order. The artifact supplies the initial state; a mutable training-time copy accumulates subsequent outcomes.

Let $v$ count visits to problem $i$. At visit $v$, the current student behavior policy $\mu_{\theta^-}$, including its configured decoding rule, generates one complete response $y_i^{(v)}$:

$$y_i^{(v)} \sim \mu_{\theta^-}(\cdot \mid x_i), \quad z_i^{(v)} = V(x_i, y_i^{(v)}, a_i), \quad Q_i^{(v)} = \big(Q_{i,2}^{(v-1)}, \ldots, Q_{i,N}^{(v-1)}, z_i^{(v)}\big). \tag{3}$$

This update occurs before teacher scoring of $y_i^{(v)}$. The initial zero is therefore discarded before the first target is constructed: the first active window contains $N-1$ previous outcomes and the one online outcome. After $v \le N$ visits it contains $v$ online outcomes. Since these outcomes can come from slightly different checkpoints, their mean is a rolling success statistic, not an unbiased estimate of the current policy's success probability.

### 3.3 Difficulty-Privileged Teacher Context

DOSD converts the updated queue into a reference-exposure budget:

$$k_i^{(v)} = \frac{1}{N} \sum_{j=1}^{N} Q_{i,j}^{(v)}, \qquad d_i^{(v)} = 1 - k_i^{(v)}, \qquad \rho_i^{(v)} = \max\{\rho_{min}, d_i^{(v)}\}. \tag{4}$$

Here $k \in [0, 100\%]$, a $70\%$ success rate would yield $\rho = 0.3$. Low recorded success grants more reference context. This allocates the privilege information a teacher model can access to help its supervision over student response.

To form the final difficulty-priviledge information, we first obtain only the reasoning segment $u_i$ by excluding the final answer $a_i$ from $r_i$. Tokenize $u_i$ into token ids $(u_{i,1}, \ldots, u_{i,L_i})$ and define

$$b_i^{(v)} = \mathcal{R}(\rho_i^{(v)} L_i), \qquad h_i^{(v)} = \mathrm{Decode}(u_{1:b_i^{(v)}}) \tag{5}$$

where $\mathcal{R}$ is a rounding rule to make sure at least one token and no more than $L_i$ tokens reatined. Thus the budget is measured in *tokens*.

The teacher thus receives $x_i$, $h_i^{(v)}$, and a separate field of the form

```
The correct answer is \boxed{...}.
```

The student receives none of these additional fields. Even at $\rho = 0.1$, the teacher retains answer access to improve student's answer format; at $\rho = 1$, teacher receives the entire eligible reasoning segment and is instructed not to duplicate the reasoning trajectory but to rewrite on its own.

### 3.4 On-policy token distribution matching

For a generated token position $t$, let $c_{i,t} = (x_i, y_{i,<t}^{(v)})$ be the unprivileged state. The student and teacher distributions over a shared vocabulary $\mathcal{V}$ are

$$\begin{aligned} S_{i,t}(w) &= \pi_\theta(w \mid c_{i,t}), \\ T_{i,t}^{(v)}(w) &= \pi_{\bar{\theta}}(w \mid x_i, h_i^{(v)}, a_i, y_{i,<t}^{(v)}). \end{aligned} \tag{6}$$

The teacher scores the student's observed prefixes under a causal attention mask; it does not generate a second response. We specify a frozen copy of the initial student, $\bar{\theta} = \theta_0$, and an unclipped, full-vocabulary forward-KL instance of DOSD. For valid generated-token positions $\mathcal{T}_\mathcal{B}$ in a batch,

$$\mathcal{L}_\mathcal{B}(\theta) = \sum_{(i,t)\in\mathcal{T}_\mathcal{B}} \omega_{i,t} \sum_{w\in\mathcal{V}} \mathrm{sg}[T_{i,t}^{(v)}(w)] \log \frac{\mathrm{sg}[T_{i,t}^{(v)}(w)]}{S_{i,t}(w)}, \qquad \sum_{(i,t)} \omega_{i,t} = 1. \tag{7}$$

For each position, we first average over batch size, then over token counts in each generated response: $\omega_{i,t} = 1/(|\mathcal{B}|\,|y_i^{(v)}|)$. Prompt and padding tokens are excluded.

The rollout, its verifier label, the queue update, and the teacher target are all stopped-gradient. Equation (7) therefore specifies a sampled on-policy *semi-gradient* update: it does not differentiate through the distribution that generated $y$ or through the outcome-conditioned target-selection rule. For a student logit $\ell_{i,t}(w)$,

$$\frac{\partial \mathcal{L}_\mathcal{B}}{\partial \ell_{i,t}(w)} = \omega_{i,t}\big(S_{i,t}(w) - T_{i,t}^{(v)}(w)\big). \tag{8}$$

The historical correctness changes the target's context, not the sign of every token's update. In particular, a failed response is not uniformly penalized as it would be under a negative scalar advantage. This distinction is essential when comparing DOSD with RLSD.

### 3.5 Conditional attenuation of privileged-reference gradients

We analyze one scored token under the forward-KL semi-gradient in Section 3.4, fixing its student state, teacher, and behavior policy. On-policy state-distribution derivatives are outside this local calculation.

**Proposition 1** (Bounded exposure changes). *If the removed queue head is $o_i^{(v)} \in \{0, 1\}$, then*

$$k_i^{(v)} - k_i^{(v-1)} = \frac{z_i^{(v)} - o_i^{(v)}}{N}, \qquad |\rho_i^{(v)} - \rho_i^{(v-1)}| \le \frac{1}{N}. \tag{9}$$

*Holding the retained $N - 1$ outcomes fixed, let $\rho_0$ and $\rho_1$ be the budgets for an incorrect and a correct current response. Then $0 \le \rho_0 - \rho_1 \le 1/N$.*

This bound requires no stationarity assumption and controls the *input budget*, not neural predictions. The privileged-gradient concern of Yang et al. (2026a) instead requires a second-moment analysis of the teacher targets.

**Separating reference and outcome variation.** Fix the teacher's inputs $B = (c, a, q^-, \bar{\theta}, \mu_{\theta^-})$ at one scored token position, including the student-visible state $c = (x, y_{<t})$, correct answer $a$, retained $N-1$ queue bits $q^-$, frozen teacher, and student sampling configuration. Conditioning on $a$ isolates reference reasoning; the separate answer channel remains. Let $R$ be a specified ensemble of valid reasoning traces for this question, $Z$ the current correctness label, $R \perp Z \mid B$, and $p = \Pr(Z = 1 \mid B)$. The ensemble is a theoretical probe, not an assumption that the stored data have multiple references. For $m$ successes in $q^-$, define

$$\rho_z = \max\{\rho_{\min}, 1 - (m+z)/N\}, \qquad T_{r,z}(w) = \pi_{\bar{\theta}}(w \mid c, a, H_{\rho_z}(r)), \tag{10}$$

where $H_\rho(r)$ is the eligible token prefix from Equation (5). The per-position student parameter gradient is $g_{r,z} = -\sum_{w \in \mathcal{V}} T_{r,z}(w) s_w$, with $s_w = \nabla_\theta \log S_\theta(w \mid c)$. Set $\mu_z = \mathbb{E}_R[g_{R,z}]$ and $V_z = \mathbb{E}_R[\|g_{R,z} - \mu_z\|^2]$. At a fixed correctness outcome $z$, $\mu_z$ is the mean gradient over valid references, while $V_z$ measures how changing only the privileged reference varies that gradient. The deviation $g_{R,z} - \mu_z$ has zero mean but can have positive squared magnitude; expanding the token-weighted gradient retains cross-token covariances (Appendix A.5). To separate the two sources of variation, write $g_{R,Z} - \bar{g}_{\mathrm{D}} = (g_{R,Z} - \mu_Z) + (\mu_Z - \bar{g}_{\mathrm{D}})$, where $\bar{g}_{\mathrm{D}} = (1-p)\mu_0 + p\mu_1$. The first term has conditional mean zero given $Z$, so their cross term vanishes and

$$\begin{aligned} \mathcal{V}_{\mathrm{D}} &:= \mathbb{E}_{R,Z}[\|g_{R,Z} - \bar{g}_{\mathrm{D}}\|^2] \\ &= (1-p)V_0 + pV_1 + p(1-p)\|\mu_1 - \mu_0\|^2. \end{aligned} \tag{11}$$

The first two terms are reference-induced variation within the incorrect and correct branches, weighted by their probabilities. The last term measures the difference between the branches' mean gradients caused by outcome-dependent target selection. Thus reducing reference sensitivity within each branch does not by itself guarantee a reduction in total gradient variation. For comparison, set $T_{r,\mathrm{full}} = \pi_{\bar{\theta}}(\cdot \mid c, a, r)$, $g_{r,\mathrm{full}} = -\sum_w T_{r,\mathrm{full}}(w) s_w$, and $V_{\mathrm{full}} = \mathbb{E}_R\|g_{R,\mathrm{full}} - \mathbb{E}_R g_{R,\mathrm{full}}\|^2$. Equation (11) shows that DOSD lowers the centered gradient second moment exactly when $(1-p)(V_{\mathrm{full}} - V_0) + p(V_{\mathrm{full}} - V_1) > p(1-p)\|\mu_1 - \mu_0\|^2$. Prefix shortening alone does not ensure this for a neural teacher.

**A conditional mitigation guarantee.**

**Proposition 2** (Attenuation under coherent conditional targets)**.** *Suppose $Z \perp (R, W) \mid B$, and a common auxiliary teacher token $W$ satisfies $T_{r,\mathrm{full}} = P(W \mid B, R = r)$ and $T_{r,z} = P(W \mid B, H_{\rho_z}(R) = H_{\rho_z}(r))$ for $z \in \{0,1\}$. Then $\mu_0 = \mu_1 = \mathbb{E}_R g_{R,\mathrm{full}}$, and*

$$\mathcal{V}_{\mathrm{D}} = V_{\mathrm{full}} - \sum_{z=0}^{1} \Pr(Z = z \mid B)\mathbb{E}_R\|g_{R,\mathrm{full}} - g_{R,z}\|^2 \le V_{\mathrm{full}}. \tag{12}$$

The tower property gives $g_{R,z} = \mathbb{E}[g_{R,\mathrm{full}} \mid B, H_{\rho_z}(R)]$; orthogonality then gives $V_{\mathrm{full}} - V_z = \mathbb{E}_R\|g_{R,\mathrm{full}} - g_{R,z}\|^2$. Substitution into Equation (11) proves the claim. The full proof and a corresponding conditional-information inequality are in Appendix A.5.

For one SGD step $\theta^+ = \theta - \eta g_{R,Z}$, update covariance is $\eta^2 \operatorname{Cov}(g_{R,Z} \mid B)$. Zero-mean deviations can therefore affect nonlinear predictions at second order (Appendix A.5); the proposition proves *attenuation*, not elimination or improved performance. Actual neural teachers need not satisfy its coherent-target assumption, and a nonvacuous reference-variance comparison requires a specified same-question reference ensemble. Outcome-dependent hints and the separate answer channel remain possible sources of privileged dependence; Appendix A.2 details their residual effects and the limits of the guarantee.

### 3.6 Training procedure and cost

For each visit, DOSD executes: *join by ID; generate one unassisted response; verify; update the queue; construct the token prefix and separate answer field; score the same response; update the student*. Queue updates must be applied exactly once, with deterministic ownership or ordering for repeated IDs across workers. Resuming training must restore the mutable queues with the model and optimizer; reloading the original offline queues would change the algorithm.

For $M$ problems and $E$ visits per problem, initialization and training require $M(N-1)+ME$ student responses, versus $ME$ for a comparable single-rollout method without initialization. The offline

Table 2: Aggregate accuracy point estimates (%). Avg is the arithmetic mean of AIME 2024, AIME 2025, and HMMT 2025. The available record does not include seed-level uncertainty or complete run metadata.

| Method | AIME 24 | AIME 25 | HMMT 25 | Avg |
|---|---|---|---|---|
| Qwen3-1.7B | 11.71 | 9.10 | 5.26 | 8.69 |
| +SFT | 9.02 | 8.30 | 5.12 | 7.48 |
| +GRPO | 11.70 | 9.20 | 5.30 | 8.73 |
| +OPSD | 15.00 | 8.10 | 6.70 | 9.93 |
| +DOSD | 15.56 | 8.61 | 6.94 | 10.37 |
| Qwen3-4B | 23.31 | 21.30 | 10.91 | 18.51 |
| thinking | 74.10 | 64.90 | 42.00 | 60.33 |
| +SFT | 22.22 | 20.00 | 11.11 | 17.78 |
| +GRPO | 24.31 | 21.63 | 11.44 | 19.13 |
| +OPSD | 32.80 | 21.10 | 16.11 | 23.34 |
| +DOSD | 33.33 | 21.39 | 16.39 | 23.70 |
| +DOSD, thinking | 77.50 | 68.06 | 46.39 | 63.98 |
| Qwen3-8B | 26.70 | 19.44 | 10.18 | 18.77 |
| thinking | 75.80 | 65.60 | 43.90 | 61.77 |
| +SFT | 25.56 | 19.17 | 10.28 | 18.34 |
| +GRPO | 26.70 | 19.17 | 10.18 | 18.68 |
| +OPSD | 46.80 | 35.20 | 18.10 | 33.37 |
| +DOSD | 46.39 | 33.33 | 18.06 | 32.59 |
| +DOSD, thinking | 78.61 | 68.61 | 46.39 | 64.54 |

cost amortizes to $(N-1)/E$ extra responses per visit, not zero. No additional online sampling is required to estimate competence, and no separate teacher response is decoded. Teacher scoring, longer teacher-prefill contexts, verification, and offline storage nevertheless incur costs. A ring buffer with a running sum supports $O(1)$ queue updates and $O(MN)$ binary state. These accounting identities motivate reporting both online efficiency and total end-to-end compute, including initialization.

# 4 Experiments

## 4.1 Evaluation scope

**Models and benchmarks.** The result matrix covers Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, with additional thinking-mode references for the 4B and 8B models. Evaluation uses AIME 2024, AIME 2025, and HMMT 2025. Student evaluation is unassisted: reference prefixes, answers, queues, and difficulty statistics are not supplied in the evaluation prompt.

**Metric.** For a fixed decoding protocol and $K$ samples per problem, mean sampled accuracy is

$$\mathrm{Acc}(\mathcal{D}) = \frac{100}{|\mathcal{D}|K} \sum_{i \in \mathcal{D}} \sum_{j=1}^{K} V(x_i, y_{i,j}, a_i), \tag{13}$$

where $V$ is the answer verifier. This metric is distinct from pass@$K$. Avg is the unweighted arithmetic mean of the three benchmark accuracies. The aggregate record does not identify $K$, decoding parameters, verifier revision, or independent training seeds; hence the tables do not report standard deviations or confidence intervals.

**Comparison scope.** Table 2 includes Qwen3 backbones and thinking-mode references where available, together with SFT, GRPO, OPSD, and DOSD.

## 4.2 Main results

The point estimates show a qualified rather than uniform advantage over OPSD. In standard mode, DOSD reaches Avg scores of 10.37, 23.70, and 32.59 for the 1.7B, 4B, and 8B models, respectively.

Table 3: Qwen3-8B-thinking controller comparison. Accuracy is reported in percent; Avg is the arithmetic mean of the three benchmarks. The fixed-prefix condition is labeled budget-calibrated in the experiment record, but realized context budgets and uncertainty are unavailable.

| Method | AIME 24 | AIME 25 | HMMT 25 | Avg |
|---|---|---|---|---|
| Qwen3-8B-thinking | 75.80 | 65.60 | 43.90 | 61.77 |
| Full prefix (OPSD) | 77.78 | 70.56 | 45.83 | 64.72 |
| Fixed prefix, budget-calibrated | 76.39 | 66.11 | 44.17 | 62.22 |
| DOSD | 78.61 | 68.61 | 46.39 | 64.54 |

Relative to OPSD, these are changes of $+0.44$, $+0.36$, and $-0.78$ percentage points. With thinking enabled, DOSD reaches 63.98 on 4B and 64.54 on 8B, exceeding the corresponding thinking-mode references by 3.65 and 2.77 points. Without repeated seeds and paired uncertainty, these differences do not establish statistical significance.

### 4.3 Controller comparison

Table 3 compares DOSD with full-prefix OPSD and a fixed-prefix condition for Qwen3-8B-thinking. The experiment record labels the fixed-prefix condition as budget-calibrated to DOSD, but does not include the fixed ratio or realized hint-token counts. The comparison therefore tests the reported conditions but cannot independently verify exact exposure matching.

DOSD improves Avg accuracy by 2.77 points over the Qwen3-8B-thinking reference and by 2.32 points over the reported budget-calibrated fixed-prefix controller. Its Avg is 0.18 points below full-prefix OPSD: DOSD is higher on AIME 2024 ($+0.83$) and HMMT 2025 ($+0.56$), but lower on AIME 2025 ($-1.95$). These point estimates are consistent with adaptive allocation outperforming a uniform context budget while recovering nearly all of the full-prefix aggregate accuracy. They do not show that DOSD universally outperforms OPSD, nor do they isolate the queue-history contribution from other unmeasured implementation or run differences.

## 5 Conclusion and Limitations

**Conclusion.** DOSD uses a per-problem FIFO history of verified outcomes to allocate reference-prefix context to an on-policy teacher while keeping the student unassisted. Our analysis separates reference-induced variation from outcome-conditioned target selection and proves attenuation only under an explicit coherent-target assumption. DOSD is not uniformly better than OPSD across model scales, but in the Qwen3-8B-thinking comparison it exceeds the reported budget-calibrated fixed-prefix controller by 2.32 points and remains within 0.18 points of full-prefix OPSD. Thus adaptive allocation can be competitive, while reduced exposure is not automatically better than full-reference supervision.

**Limitations.** The evidence consists of aggregate point estimates without repeated seeds, per-problem predictions, $K$, complete run settings, or paired intervals, so the observed differences may reflect evaluation or training variance. Realized hint-token counts, GPU-hours, and offline-initialization costs are unavailable; exact budget matching and efficiency gains are therefore unverified. The study is limited to mathematical benchmarks and Qwen3 models, without direct adaptive-method comparisons, history-isolating controls, or reference-sensitivity measurements. DOSD also retains the teacher's answer channel, conditions the target on correctness, inherits verifier errors, and relies on theoretical assumptions that need not hold for a neural teacher. We therefore claim neither universal improvement nor elimination of privileged dependence.

## AI Use Statement

Generative AI assisted with mathematical formulation and grammar checking, experiment suggestions, and also participated in preparation of the LaTeX source.

## REFERENCES


Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Priyank Agrawal, Ankur Samanta, Shervin Ghasemlou, Boris Vidolov, Jalaj Bhandari, Kavosh Asadi, Daniel Jiang, and Aditya Modi. Off-context GRPO: Learning to reason on hard problems using privileged information. *arXiv preprint arXiv:2607.19313*, 2026. URL https://arxiv.org/abs/2607.19313v2. Version 2.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in LLM reasoning. *arXiv preprint arXiv:2605.11458*, 2026. URL https://arxiv.org/abs/2605.11458v3. Version 3.

Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How PI-conditioned teachers break self-distillation. *arXiv preprint arXiv:2608.04794*, 2026. URL https://arxiv.org/abs/2608.04794v1. Version 1.

Gagan Mundada, Zihan Huang, Rohan Surana, Sheldon Yu, Jennifer Yuntong Zhang, Xintong Li, Tong Yu, Lina Yao, Jingbo Shang, Julian McAuley, and Junda Wu. WS-GRPO: Weakly-supervised group-relative policy optimization for rollout-efficient reasoning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.17025.

Yuxiao Qu, Amrith Setlur, Virginia Smith, Ruslan Salakhutdinov, and Aviral Kumar. POPE: Learning to reason on hard problems via privileged on-policy exploration. *arXiv preprint arXiv:2601.18779*, 2026. URL https://arxiv.org/abs/2601.18779v1. Version 1.

Amrith Setlur, Zijian Wang, Andrew Cohen, Paria Rashidinejad, and Sang Michael Xie. Reuse your FLOPs: Scaling RL on hard problems by conditioning on very off-policy prefixes. *arXiv preprint arXiv:2601.18795*, 2026. URL https://arxiv.org/abs/2601.18795v2. Version 2.

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

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified OPSD: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026. URL https://arxiv.org/abs/2607.02234v1. Version 1.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022. URL https://arxiv.org/abs/2209.15189.

Alex Stein, Furong Huang, and Tom Goldstein. GATES: Self-distillation under privileged context with consensus gating. *arXiv preprint arXiv:2602.20574*, 2026. URL https://arxiv.org/abs/2602.20574v1. Version 1.

Zhiquan Tan and Yinrong Hong. PAINT: Partial-solution adaptive interpolated training for self-distilled reasoners. *arXiv preprint arXiv:2604.26573*, 2026. URL https://arxiv.org/abs/2604.26573v1. Version 1.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026a. URL https://arxiv.org/abs/2604.03128v2. Version 2.

Yongkang Yang, Zhezheng Hao, Hong Zhang, Yi Liu, Xiankun Lin, Wence Ji, Fanjunduo Wei, Jiarui Yu, Qiang Lin, Xiaoyun Liang, and Hande Dong. Matching supervision to the student's learning capacity: A unified framework for on-policy self-distillation. *arXiv preprint arXiv:2608.08176*, 2026b. URL https://arxiv.org/abs/2608.08176v1. Version 1.

Yuxiao Yang, Xiaoyun Wang, and Weitong Zhang. OGLS-SD: On-policy self-distillation with outcome-guided logit steering for LLM reasoning. *arXiv preprint arXiv:2605.12400*, 2026c. URL https://arxiv.org/abs/2605.12400v1. Version 1.

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

# A Proofs and boundaries of the analysis

## A.1 Queue dynamics

*Proof of Proposition 1.* Let $s$ be the sum of the retained $N-1$ entries. Before the update the queue sum is $s+o$, and afterwards it is $s+z$. Dividing their difference by $N$ proves the first equality in Equation (9). The map $g(k)=\max(\rho_{\min}, 1-k)$ is nonincreasing and 1-Lipschitz. Consequently,

$$|g(k^{(v)})-g(k^{(v-1)})| \le |k^{(v)}-k^{(v-1)}| = |z-o|/N \le 1/N.$$

For fixed retained history, $k_0 = s/N$ and $k_1 = (s+1)/N$. Applying the same monotonicity and Lipschitz properties yields $0 \le \rho_0 - \rho_1 \le 1/N$. If $p$ is the current conditional success probability with this history fixed, the expected exposure is $\rho_0 - p(\rho_0-\rho_1)$; its derivative with respect to $p$ is nonpositive. □

The result also clarifies the initialization. The first application of Equation (3) removes the placeholder, irrespective of whether its stored value is 0 or 1. Thus changing that placeholder alone cannot change any teacher target when the prescribed update order is respected. After $v \le N$, exactly $v$ online outcomes remain in the queue; after $N$, it is a window of the last $N$ online outcomes. These are visits to a particular problem. For example, three visits with $N=10$ still leave seven offline entries, even if the global training run has taken many optimizer steps.

For either floor or ceiling token rounding, a ratio change of at most $1/N$ changes the requested prefix length by at most $\lceil L_i/N \rceil$ tokens. This statement concerns integer budgets only. It places no continuity condition on the teacher's response to those tokens. When the eligible segment is empty, all budgets induce the same empty hint.

## A.2 Residual outcome dependence and limits of the guarantee

Fix the base information $B$ from Section 3.5 and a particular reference $R=r$. Let $F=(B,R=r)$ and $p=\Pr(Z=1 \mid F)$; under $R \perp Z \mid B$, this is the same $p$ used in Equation (11). With $\bar{g}_r = (1-p)g_{r,0} + p g_{r,1}$, varying only the current correctness outcome gives

$$\mathbb{E}_{Z|F}\|g_{r,Z} - \bar{g}_r\|^2 = p(1-p)\|g_{r,1}-g_{r,0}\|^2. \tag{14}$$

Indeed, $g_{r,0}-\bar{g}_r = -p(g_{r,1}-g_{r,0})$ and $g_{r,1}-\bar{g}_r = (1-p)(g_{r,1}-g_{r,0})$. The $1/N$ exposure-budget bound does not bound this gradient difference for an unconstrained neural teacher; identical rounded hints make it zero. This fixed-reference variation is distinct from the between-branch term $p(1-p)\|\mu_1-\mu_0\|^2$ in Equation (11), which first averages over references.

For the loss-level counterpart, set $T_z = T_{r,z}$ and $\bar{T} = (1-p)T_0 + pT_1$, and draw an auxiliary teacher token $W' \mid Z=z, F \sim T_z$.

**Proposition 3** (Outcome-conditioned target decomposition)**.** *For any student distribution $S$ with positive support on the teacher targets,*

$$\mathbb{E}_{Z|F} D_{\mathrm{KL}}(T_Z\|S) = D_{\mathrm{KL}}(\bar{T}\|S) + \mathrm{JS}_p(T_0,T_1), \tag{15}$$

*where*

$$\mathrm{JS}_p(T_0,T_1) = (1-p)D_{\mathrm{KL}}(T_0\|\bar{T}) + pD_{\mathrm{KL}}(T_1\|\bar{T}) = I(W';Z\mid F) \le h_2(p) \le \log 2. \tag{16}$$

*Here $h_2$ is binary entropy in nats; $W'$ is not the sampled student token.*

*Proof of Proposition 3.* Fix $F$ and write $\lambda_0 = 1 - p$, $\lambda_1 = p$, and $\bar{T}(w) = \sum_z \lambda_z T_z(w)$. On the support of $T_z$, insert the mixture into the likelihood ratio:

$$\sum_z \lambda_z D_{\mathrm{KL}}(T_z \| S) = \sum_z \lambda_z \sum_w T_z(w) \left[ \log \frac{T_z(w)}{\bar{T}(w)} + \log \frac{\bar{T}(w)}{S(w)} \right]$$
$$= \sum_z \lambda_z D_{\mathrm{KL}}(T_z \| \bar{T}) + D_{\mathrm{KL}}(\bar{T} \| S). \tag{17}$$

Under the explicitly constructed joint $P(Z = z, W' = w \mid F) = \lambda_z T_z(w)$, the first term is $I(W'; Z \mid F)$. Mutual information is bounded by the entropy of $Z$, so it is at most $h_2(p) \le \log 2$. The usual zero-weight convention covers $p \in \{0, 1\}$. □

For a frozen teacher and fixed scoring-state law, the Jensen–Shannon residual in Equation (15) has zero student derivative, but Equation (14) shows that sampled gradients can still vary. This distinction precludes treating a nonzero residual loss as direct evidence of a harmful mean update.

The probabilities must correspond to the scoring-state distribution used by the loss. For a response-averaged loss, one may sample a response and then a valid position uniformly. For a pooled-token loss, one instead uses token occupancy, which weights longer responses differently. The induced conditional success probability at a scored state need not agree between these conventions. Neither is generally equal to the historical queue mean.

The result is local to an optimization step with behavior and teacher frozen. Over training, both the student state distribution and the queue-dependent targets change. Hence it is not a statement that a fixed residual persists unchanged throughout training. Likewise, the target at an early prefix can depend on a future terminal outcome through $\rho_Z$: causal masking inside the teacher forward pass does not remove that side channel.

The attenuation guarantee in Proposition 2 additionally assumes coherent conditional teacher targets and a specified same-question reference ensemble. Arbitrary neural prompt responses need not satisfy that assumption; with one deterministic reference per fully identified question, conditional reference variance is vacuous without such an ensemble. DOSD also retains a minimum $10\%$ hint and the separate correct-answer channel. Consequently, the guarantee neither eliminates privileged dependence nor bounds its effect on task performance. Both reference sensitivity and outcome-dependent target variation require direct measurement (Appendix A.3).

### A.3 General privileged-context decomposition

Let $C$ be an unprivileged scoring state and $U$ a teacher-only variable under a specified joint law. For example, $U$ may include a reference realization, its exposed prefix, a separate answer, and an outcome-dependent budget. Freeze this law and the teacher kernel $T_u(w \mid c)$. Define

$$\bar{T}(w \mid c) = \mathbb{E}_{U|c} T_U(w \mid c), \qquad W \mid C = c, U = u \sim T_u(\cdot \mid c).$$

The same algebra as Equation (17) gives the standard mixture identity

$$\mathbb{E}_{C,U} D_{\mathrm{KL}}(T_U(\cdot \mid C) \| S_\theta(\cdot \mid C)) = \mathbb{E}_C D_{\mathrm{KL}}(\bar{T}(\cdot \mid C) \| S_\theta(\cdot \mid C)) + I(W; U \mid C). \tag{18}$$

This is a structural identity, related to the analysis of Yang et al. (2026a), rather than a new impossibility theorem. A nonzero second term requires conditional variation in teacher targets under the chosen law. If a dataset gives one deterministic reference and answer per fully identified question, their conditional entropy given that question is zero. Computational difficulty in reconstructing them does not turn that deterministic random variable into a nondegenerate one. Alternative-reference ensembles or random outcome-dependent budgets can supply meaningful variation, but the ensemble must be specified rather than assumed.

With a frozen teacher and fixed $(C, U)$ law, $I(W; U \mid C)$ has zero derivative with respect to student parameters. Thus the existence of a residual loss does not itself prove harmful expected gradients, inevitable divergence, or a positive gradient-variance lower bound. To make the distinction explicit, let $s_w = \nabla_\theta \log S_\theta(w \mid c)$. At a fixed state,

$$g_u = -\sum_w T_u(w) s_w, \qquad \bar{g} = -\sum_w \bar{T}(w) s_w, \qquad \delta_u = g_u - \bar{g}.$$

Although $\mathbb{E}_U \delta_U = 0$, the exact second moment contains all cross-token covariances:

$$\mathbb{E}_U \|\delta_U\|^2 = \sum_{w,w'} \mathrm{Cov}_U(T_U(w), T_U(w'))\langle s_w, s_{w'}\rangle. \tag{19}$$

Dropping the off-diagonal terms is not generally valid. If the score vectors additionally satisfy $\|s_w\| \le G$, a valid *upper* bound follows from Pinsker's inequality:

$$\mathbb{E}_U \|\delta_U\|^2 \le G^2 \mathbb{E}_U \|T_U - \bar{T}\|_1^2 \le 2G^2 \mathbb{E}_U D_{\mathrm{KL}}(T_U \| \bar{T}). \tag{20}$$

This conditional bound does not establish a lower bound, a monotone relationship across training configurations, or that DOSD reduces the right-hand side.

### A.4 When shorter context supports an information argument

There is a useful idealized interpretation of prefix restriction, but its assumptions matter. Suppose $W$, the reference, the answer $A$, and scoring states are governed by *one common joint distribution*. Condition on reference length $L$ as well, and choose nested absolute prefix budgets with $H_s$ a deterministic coarsening of $H_\ell$. Let $D = (C, A, L)$ and require coherent conditional teachers

$$T_s = P(W \mid D, H_s), \qquad T_\ell = P(W \mid D, H_\ell), \qquad T_A = P(W \mid D).$$

Then the chain rule yields

$$\begin{aligned} \mathbb{E} D_{\mathrm{KL}}(T_\ell \| T_A) - \mathbb{E} D_{\mathrm{KL}}(T_s \| T_A) &= I(W; H_\ell \mid D) - I(W; H_s \mid D) \\ &= I(W; H_\ell \mid D, H_s) \ge 0. \end{aligned} \tag{21}$$

Thus, in this coherent model, discarding reference information cannot increase that particular conditional-information quantity. Conditioning on length prevents a percentage-based prefix rule from silently revealing information through the unknown original length.

Actual neural teachers queried with different prompts are not guaranteed to be the conditional distributions of this common joint, and DOSD's budget also depends on the response outcome. Equation (21) therefore must not be applied as a proven monotonicity result for DOSD's neural targets. For example, even if two budgets differ by $1/N$, an unconstrained teacher could return $(1-\epsilon, \epsilon)$ for one prompt and $(\epsilon, 1-\epsilon)$ for the other. At equal branch weights their Jensen–Shannon divergence approaches $\log 2$ as $\epsilon \to 0$, regardless of $N$. This counterexample rules out an unconditional reduction claim for arbitrary neural targets.

### A.5 Gradient-variance attenuation under coherent targets

We give the full single-token derivation behind Proposition 2. All expectations below condition on the fixed base information $B$ in Section 3.5. Student score vectors $s_w$ are consequently fixed. For a fixed outcome branch $z$, inserting $g_{R,z} = -\sum_w T_{R,z}(w) s_w$ into the definition of $V_z$ and expanding the squared norm gives

$$V_z = \sum_{u,w \in \mathcal{V}} \mathrm{Cov}_R(T_{R,z}(u), T_{R,z}(w))\langle s_u, s_w\rangle. \tag{22}$$

This retains the cross-token terms; the diagonal terms alone are neither the exact parameter-gradient variance nor a universal lower bound. For $\bar{g}_{\mathrm{D}} = (1-p)\mu_0 + p\mu_1$, conditioning on $Z$ and expanding $g_{R,z} - \bar{g}_{\mathrm{D}} = (g_{R,z} - \mu_z) + (\mu_z - \bar{g}_{\mathrm{D}})$ removes the cross term because $\mathbb{E}_R(g_{R,z} - \mu_z) = 0$. The remaining between-branch contribution is

$$(1-p)\|\mu_0 - \bar{g}_{\mathrm{D}}\|^2 + p\|\mu_1 - \bar{g}_{\mathrm{D}}\|^2 = p(1-p)\|\mu_1 - \mu_0\|^2,$$

which proves Equation (11) without assuming independent teacher-token coordinates.

For the coherent-teacher claim, let $H_z = H_{\rho_z}(R)$ and suppose $T_{r,\mathrm{full}} = P(W \mid B, R = r)$ and $T_{r,z} = P(W \mid B, H_z = H_z(r))$ under one joint law. Since $g_{R,\mathrm{full}} = -\mathbb{E}[s_W \mid B, R]$ and $H_z$ is a deterministic function of $R$, the tower property yields

$$g_{R,z} = -\mathbb{E}[s_W \mid B, H_z] = \mathbb{E}[g_{R,\mathrm{full}} \mid B, H_z].$$

Write $\mu = \mathbb{E}_R g_{R,\text{full}}$. Conditional-expectation orthogonality makes $g_{R,z} - \mu$ orthogonal to $g_{R,\text{full}} - g_{R,z}$. Therefore

$$V_{\text{full}} = V_z + \mathbb{E}_R \|g_{R,\text{full}} - g_{R,z}\|^2, \qquad \mu_z = \mu. \tag{23}$$

Because $Z \perp (R, W) \mid B$, its two branches use the same conditional reference law and $\bar{g}_{\text{D}} = \mu$. Averaging Equation (23) over $Z$ proves Equation (12). In fact, the analogous identity holds for covariance matrices: if $\Sigma_{\text{full}} = \text{Cov}(g_{R,\text{full}} \mid B)$ and $\Sigma_{\text{D}} = \text{Cov}(g_{R,Z} \mid B)$, then

$$\Sigma_{\text{full}} - \Sigma_{\text{D}} = \sum_{z=0}^{1} \Pr(Z = z \mid B) \mathbb{E}_R \left[ (g_{R,\text{full}} - g_{R,z})(g_{R,\text{full}} - g_{R,z})^\top \right] \succeq 0. \tag{24}$$

Taking traces recovers the proposition. This is a conditional, one-step comparison with the *same answer channel* and full eligible reasoning as the comparator, not a claim of improved task performance.

The same idealized joint law gives an information counterpart. Independence implies $I(W; Z \mid B) = 0$, and conditioning on $Z$ followed by data processing for $H_z = f_z(R)$ gives

$$I(W; Z, H_{\rho_Z}(R) \mid B) = \sum_{z=0}^{1} \Pr(Z = z \mid B) I(W; H_{\rho_z}(R) \mid B) \le I(W; R \mid B). \tag{25}$$

The bound includes the selector $Z$ for analysis even though DOSD does not pass $Z$ as a separate teacher field. It conditions on the correct answer $a \in B$, so it does not bound the information carried by that separate answer field. If $R$ is deterministic after conditioning on the fully identified question, both sides of this inequality are zero; a non-degenerate same-question reference ensemble must be specified for this comparison to quantify reference variation.

To connect gradient covariance to the student distribution, let $\theta^+ = \theta - \eta g_{R,Z}$ and $\theta_* = \theta - \eta \bar{g}_{\text{D}}$. For any smooth test-time student-token probability $f(\theta)$ with locally bounded third derivative, Taylor expansion around $\theta_*$ gives

$$\mathbb{E}_{R,Z} f(\theta^+) - f(\theta_*) = \frac{\eta^2}{2} \operatorname{tr}\left(\nabla_\theta^2 f(\theta_*) \Sigma_{\text{D}}\right) + O\left(\eta^3 \mathbb{E} \|g_{R,Z} - \bar{g}_{\text{D}}\|^3\right). \tag{26}$$

The first-order residual vanishes, but the second-order term need not. Equation (24) reduces the conditional covariance in every parameter direction under the coherent model; it does not order the signed prediction shift because the Hessian may be indefinite. Moreover, real multi-step optimization, adaptive optimizers, and moving on-policy states are outside this local result.

Finally, fixing a particular reference $r$ and varying only the current outcome gives Equation (14). If $\|s_w\| \le G$ for all vocabulary items, then Pinsker's inequality applied to the binary mixture in Equation (16) yields

$$p(1-p)\|g_{r,1} - g_{r,0}\|^2 \le p(1-p) G^2 \|T_{r,1} - T_{r,0}\|_1^2 \le 2G^2 \, \text{JS}_p(T_{r,0}, T_{r,1}).$$

A further $O(N^{-2})$ bound would require an additional local condition $\|T_{r,1} - T_{r,0}\|_1 \le L_\rho |\rho_1 - \rho_0|$ with controlled $L_\rho$. Token truncation and an unconstrained neural teacher do not supply this regularity automatically. Thus the theorem gives a principled route to mitigation, while the retained hint, separate answer, and outcome-dependent target remain possible sources of privileged dependence.

## B Implementation Contract and Audit Requirements

This appendix records requirements for aligning the mathematical procedure with an implementation; it does not claim that the new-cluster code has been independently audited.

**Data artifacts.** Every queue record must have a unique `problem_id`, exactly $N$ binary entries, and an explicit oldest-to-newest order. Dataset and queue identifiers must be checked for missing, duplicate, or mismatched records before training. The separate generation artifact should record the offline checkpoint, decoding settings, verifier version, and original responses so that the initialization labels can be audited. Shuffling the original dataset must not change the queue attached to any problem.

**Atomic visit semantics.** The current outcome is incorporated exactly once before the corresponding teacher pass. Re-scoring a response or taking additional optimization steps on it must not append its outcome repeatedly. If a batch contains the same problem more than once, the implementation needs a deterministic serial order or a documented equivalent ownership scheme. A model checkpoint should be paired with queue contents, visit counters, and data-loader state; otherwise resumed training is not equivalent to continuation.

**Prompt and token alignment.** The teacher and student score identical generated token IDs, with the appropriate offset for their different prompt lengths. Prompt tokens and padding are masked from the loss, and no future generated tokens are visible through the teacher's causal mask. The teacher prompt still depends retrospectively on correctness through its prefix budget. Audit the final-answer extraction, answer-marker preprocessing, tokenizer, prefix rounding, total context limit, and empty-hint behavior. A truncation fallback must not silently expose the full answer-bearing reference or truncate away the separate answer field. Reference text without a recognizable answer marker requires an explicit audit policy, not an assumption of semantic answer removal.

**Run manifest.** The manifest should fix $N$, $\rho_{\min}$, offline and student initial checkpoints, teacher update policy, trainable parameters, sampling parameters, rollout horizon, KL direction and clipping, distillation temperature, reduction weights, batch schedule, seeds, and verifier. These choices are not determined by the queue rule. Any departure from the fixed-teacher, unclipped-forward-KL instance in Section 3.4 must be reflected in the manuscript, particularly when interpreting Equation (15) as a loss decomposition.

**Audit quantities.** Relevant audit quantities include pre- and post-update queue sums, current correctness, reference-prefix ratio, actual hint-token count, offline fraction, and empty-hint frequency. Matched-state diagnostics can compare alternative prefix budgets and valid references with the same final answer, while delayed, shuffled, and reversed controls can isolate timing and allocation effects. These diagnostics are absent from the available experiment record and are not used to support the empirical claims in this paper.

## C Scope of the Experimental Record

The numerical evidence available for this paper is limited to the aggregate benchmark accuracies in Tables 2 and 3. No additional ablation, efficiency, robustness, or reference-sensitivity results are included.

**Aggregate evaluation data.** The record provides one accuracy value for each method–benchmark pair on AIME 2024, AIME 2025, and HMMT 2025. It does not include per-problem predictions, sample identifiers, the number of samples per problem, evaluation seeds, or verifier and decoding revisions. Consequently, paired resampling and seed-level uncertainty cannot be reconstructed from the available artifacts.

**Controller terminology.** Full-prefix OPSD exposes the complete eligible reference-reasoning segment to the teacher. The fixed-prefix condition uses a single prefix policy for all problems and is labeled budget-calibrated in the supplied result record. DOSD instead selects a problem- and visit-dependent prefix using the updated outcome queue. The record does not contain the fixed ratio, realized hint-token counts, or a token-by-token budget audit, so exact exposure matching is not independently verified here.

**Unreported resource and mechanism measurements.** Offline initialization requires $N - 1$ sampled responses per training problem, whereas each training visit uses one online student rollout. The available record does not report $N$, visit distributions, generated or teacher-processed tokens, GPU-hours, or queue-state traces. It also does not contain current-only, frozen-window, delayed-update, shuffled-allocation, reverse-rank, answer-channel, or queue-length controls. We therefore make no empirical claim about end-to-end efficiency, the isolated value of historical memory, or the degree of reference-dependence reduction.