PAST: PRIVILEGED ADAPTATION FROM COMPLETE STUDENT TRAJECTORIES FOR ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) uses a privileged teacher to supervise a reasoning model on prefixes sampled from its own rollouts. Yet each rollout also reveals how the student's response unfolds and whether it succeeds, hindsight that standard OPSD does not use to form the teacher. We introduce Privileged Adaptation from Student Trajectories (PAST)[1], which treats each completed student trajectory as additional privileged information for the OPSD teacher while leaving the student's distillation prefixes unchanged. PAST preserves the student's next-token distribution on correct trajectories and adapts the teacher on failed trajectories toward verified success while keeping it close to the student. We characterize what a teacher informed by the trajectory can transfer to a prefix-only student. Forward-KL distillation projects the teacher distributions to their conditional arithmetic mean given the prefix. This projection separates variation across trajectories that remains privileged from the mean policy shift available to the student. For correct trajectories, the unclipped population objective also has the frozen student as an ideal distributional fixed point. Across three mathematical reasoning benchmarks, PAST improves the Avg@12 macro average over Vanilla OPSD by 5.6 percentage points. A 2×2 factorial study shows gains from complete trajectories and teacher adaptation, while trajectory removal and shuffling confirm that the adapted teacher uses the matching hindsight context.

1 INTRODUCTION

On-policy distillation trains a reasoning model on prefixes sampled from its own rollouts, reducing the mismatch between training and inference while retaining token-level teacher supervision (Agarwal et al., 2024; Gu et al., 2024). On-policy self-distillation (OPSD) makes this interface self-contained. A privileged copy of the model teaches its unprivileged policy on those prefixes. Yet, by the time supervision is constructed, the rollout has also revealed the student's complete response and whether it succeeds. Standard OPSD uses task privilege to teach on the prefixes, but does not adapt the teacher from how the student's response unfolded (Zhao et al., 2026a).

Completed responses already support several forms of self-improvement. SD-Zero conditions a reviser on the response and its outcome, H2SD routes successful and failed responses through different hindsight mechanisms, and RSTG supplies teacher guidance to failed zero-variance RL groups (He et al., 2026; Cai et al., 2026; Han et al., 2026). Trainable teacher distillation provides a complementary mechanism by adapting teacher modules during autoregressive distillation (Rao et al., 2026). These developments motivate a broader question. How can the current student's complete trajectory serve as student-specific privilege for an adaptive OPSD teacher while preserving the original prefix-only student interface? At each supervised prefix, the rest of the trajectory provides such privilege to the teacher while remaining unavailable to the deployed student.

Using complete trajectories as privilege does not automatically produce useful supervision. Future-dependent targets can conflict at the same prefix and cancel under distillation (Zhao et al., 2026b; Wang et al., 2026). Outcome matters as well. Supervision on correct responses should preserve the

[1]An anonymized implementation is available at `https://anonymous.4open.science/r/PAST-F135/README.md`.

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.

student's behavior, whereas supervision on failed responses should move toward verified success without drifting beyond what the student can learn.

We introduce *Privileged Adaptation from Student Trajectories* (PAST), which uses each complete student response as additional privilege for the OPSD teacher. Each cycle freezes the student, samples fresh responses, and adapts one teacher based on final answer correctness. Full-vocabulary distribution matching preserves the student's next-token distribution on correct responses. On failed responses, optimization searches for continuations that pass the verifier while regularization keeps the teacher close to the student. PAST then distills the adapted teacher once on the original student prefixes. Teacher continuations never replace these prefixes, and the next cycle resamples from the updated student.

Our analysis characterizes what a teacher conditioned on the trajectory can transfer to a prefix-only student. Forward-KL distillation projects these teacher distributions to their conditional arithmetic mean given the prefix (Lin, 1991; Banerjee et al., 2005). This separates privileged variation that the student cannot reproduce from the mean policy shift it can learn, explaining why teacher use of hindsight need not improve the student. On correct trajectories, the unclipped population objective has the frozen student as an ideal distributional fixed point.

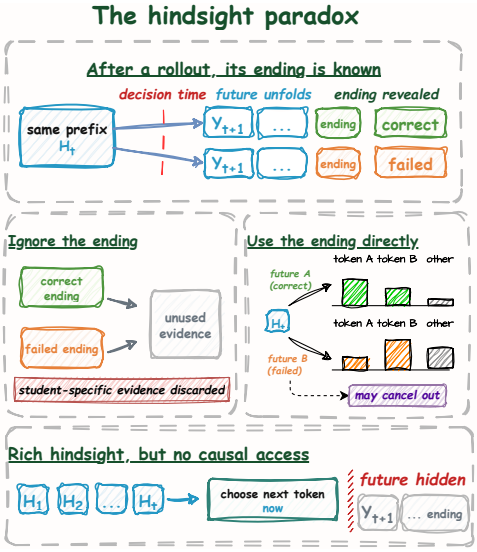

Figure 1: Complete rollouts reveal useful hindsight, but it cannot be copied directly into a prefix-only policy. PAST first adapts the teacher, then distills back to the original prefix.

Our contributions are summarized below.

- We propose PAST, which adapts one privileged teacher from complete correct and failed student trajectories, then distills it onto the student's unchanged on-policy prefixes.
- We characterize forward-KL causal projection and separate privileged teacher variation from the policy shift available to a prefix-only student.
- Across three mathematical reasoning benchmarks, PAST improves the Avg@12 macro average over Vanilla OPSD by 5.6 percentage points. Factorial controls assign positive contributions to trajectory access and teacher adaptation, while trajectory removal and shuffling confirm that the teacher uses the hindsight context.

2 Preliminaries and Problem Setup

PAST combines verifiable policy optimization with on-policy self-distillation. We first define the two training interfaces, then isolate the information change introduced by a complete student trajectory.

2.1 Verifiable On-Policy Learning

At training cycle k, let $(X, Z) \sim \mathcal{D}$ denote a task and its task privilege, such as a reference solution authorized by the training protocol. The frozen student p_k does not observe Z and generates a complete response

$$Y = (Y_1, \ldots, Y_T) \sim p_k(\cdot \mid X). \tag{1}$$

A verifier returns the trajectory-level outcome

$$R = V(X, Y) \in \{0, 1\}. \tag{2}$$

The verifier may use answer information authorized by the task protocol. It supplies neither token-level labels nor intermediate supervision.

Group Relative Policy Optimization (GRPO) converts such outcomes into an on-policy update (Shao et al., 2024). Given G responses $Y^{(g)}$ sampled from an old policy $\mu_{\theta_{\text{old}}}$, let $R_g = V(X, Y^{(g)})$ and

$$\widehat{A}_g = \frac{R_g - \overline{R}}{s_R + \epsilon_A}, \qquad \rho_{g,t}(\theta) = \frac{\mu_\theta(Y_t^{(g)} \mid X, Y_{<t}^{(g)})}{\mu_{\theta_{\text{old}}}(Y_t^{(g)} \mid X, Y_{<t}^{(g)})}, \tag{3}$$

where $\overline{R}$ and s_R are the group mean and standard deviation. When $s_R = 0$, the group-relative advantage is zero. Omitting an optional reference-policy penalty, the loss is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{T_g}\sum_{t=1}^{T_g}\min\Big\{\rho_{g,t}\widehat{A}_g, \text{clip}(\rho_{g,t}, 1-\epsilon, 1+\epsilon)\widehat{A}_g\Big\}\right]. \tag{4}$$

Every token in one response shares the same outcome-derived advantage. PAST uses this group-relative signal only to adapt its teacher on failed student trajectories.

2.2 On-Policy Distillation

On-policy distillation evaluates a teacher on states generated by the student (Agarwal et al., 2024; Gu et al., 2024). At token position t, the state seen by the student is

$$H_t = (X, Y_{<t}, t). \tag{5}$$

For a teacher distribution $q(\cdot \mid H_t)$ and a candidate student $\pi(\cdot \mid H_t)$, the forward-KL objective is

$$\mathcal{L}_{\text{OPD}}(\pi; q) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} D_{\text{KL}}(q(\cdot \mid H_t) \,\|\, \pi(\cdot \mid H_t))\right], \tag{6}$$

where the expectation follows the student rollout distribution and the teacher is a stop-gradient target. The term *on-policy* refers to the states H_t, not to trajectories sampled from the teacher.

2.3 Privileged On-Policy Self-Distillation

On-policy self-distillation (OPSD) forms the teacher and student from the same model under different information sets (Zhao et al., 2026a). Vanilla OPSD supplies the teacher with task privilege Z while the student observes only H_t.

$$q_k^{\text{OPSD}}(a \mid H_t, Z), \qquad p_k(a \mid H_t). \tag{7}$$

Its student update instantiates Eq. (6) with $q = q_k^{\text{OPSD}}$. The sampled response determines the prefixes on which both policies are evaluated, but its unobserved suffix is not part of the Vanilla OPSD teacher context.

2.4 Trajectory-Privileged Teacher and Causal Student

PAST changes how this privileged teacher is formed. It augments the task privilege with the complete student trajectory

$$U = (Y, Z), \tag{8}$$

and adapts one teacher $q_\phi(a \mid H_t, U)$. The outcome R selects the adaptation objective but is not included in U. Thus the teacher can condition on the suffix of Y beyond H_t, while a deployable student $\pi(a \mid H_t)$ remains prefix-only at inference. Appendix B compares this interface with the closest on-policy distillation protocols.

After adaptation, PAST freezes the teacher and defines

$$q_U(a \mid H_t) = q_{\phi_{k+1}}(a \mid H_t, U). \tag{9}$$

The student is trained on the original rollout states with

$$\mathcal{L}_{S,k}(\pi) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} D_{\text{KL}}(q_U(\cdot \mid H_t) \,\|\, \pi(\cdot \mid H_t))\right]. \tag{10}$$

Optimizing this objective produces p_{k+1}, and the next cycle draws fresh rollouts from the updated student. Section 3 defines how correct and failed outcomes adapt the teacher. Section 4 characterizes the supervision that survives projection to the prefix-only student.

3 METHOD

PAST turns each fresh student rollout batch into one teacher update followed by one student update. The completed rollout becomes teacher privilege, the verifier outcome selects the adaptation objective, and the adapted teacher supervises the student's original prefixes. Correct and failed trajectories train the same teacher. Figure 2 summarizes this cycle.

3.1 COMPLETE STUDENT TRAJECTORIES AS TEACHER PRIVILEGE

At cycle k, the frozen student p_k generates $Y \sim p_k(\cdot \mid X)$ without privileged information. PAST gives the teacher the same task privilege Z as Vanilla OPSD and adds the complete student response Y. The external outcome R selects the training objective but is not included in the teacher input. Both outcome branches therefore adapt one conditional distribution $q_\phi(\cdot \mid H_t, U)$ under the same information interface. We initialize the teacher from the student, $q_0 \leftarrow p_0$, so their initial policy difference is zero. Appendix B gives the exact input serialization.

3.2 PRESERVING CORRECT STUDENT BEHAVIOR

A verified correct trajectory should train the teacher without replacing behavior that already succeeds. PAST therefore matches the frozen student's full next-token distribution on the observed trajectory rather than fitting only its sampled tokens. For distributions p and q at one position, define the OPSD-style clipped contribution

$$\widetilde{D}_\tau(p\|q) = \sum_{v\in\mathcal{V}} \min\left\{ p(v)\log\frac{p(v)}{q(v)}, \tau \right\}. \tag{11}$$

The correct-trajectory loss is

$$\mathcal{L}_{+,i} = \frac{1}{T_i}\sum_{t=1}^{T_i} \widetilde{D}_{\tau_{\mathrm{OPSD}}}(p_k(\cdot \mid H_{i,t})\|q_\phi(\cdot \mid H_{i,t}, U_i)). \tag{12}$$

The threshold and token mask match the corresponding Vanilla OPSD student update. Training uses the clipped quantity for stability and logs the unclipped full-vocabulary KL. No hard-label cross-entropy is added.

3.3 ADAPTING ON FAILED STUDENT TRAJECTORIES

A failed trajectory calls for a teacher update toward success that remains compatible with the frozen student. The teacher samples complete continuations under U_i, and the verifier assigns each continuation a binary reward. A group containing both outcomes supplies the group-relative policy signal $\mathcal{L}_{\mathrm{GRPO},g}$. In this instantiation of Eq. (4), the trainable policy is the privileged teacher q_ϕ, and the old policy is its frozen snapshot q_{ϕ_k} from the start of the teacher update. For each successful continuation C_j in group g, PAST measures its exact proximity to the frozen student

$$K_j = \frac{1}{|C_j|}\sum_s D_{\mathrm{KL}}(q_\phi(\cdot \mid G_{j,s}, U_i)\|p_k(\cdot \mid G_{j,s})), \tag{13}$$

where $G_{j,s}$ is the teacher continuation prefix at position s. Let $\mathcal{S}_g$ contain the successful continuations and let m_g^{mix} indicate a group with both outcomes. An active failed group minimizes

$$\mathcal{L}_{-,g} = m_g^{\mathrm{mix}}\mathcal{L}_{\mathrm{GRPO},g} + \beta_{\mathrm{KL}}\frac{1}{|\mathcal{S}_g|}\sum_{j\in\mathcal{S}_g} K_j, \qquad \beta_{\mathrm{KL}} = 0.05. \tag{14}$$

Mixed groups combine success seeking with student proximity. All-success groups retain only the proximity term. Adaptive sampling adds continuations after an initial all-failure draw and skips a trajectory only when a full retry also fails. This rule changes sampling cost rather than the stated population objective. Appendix B gives the complete sampler.

PAST first averages examples within each active branch and then averages the active branches

$$\mathcal{L}_T = \frac{m_+\overline{\mathcal{L}}_+ + m_-\overline{\mathcal{L}}_-}{m_+ + m_-}, \tag{15}$$

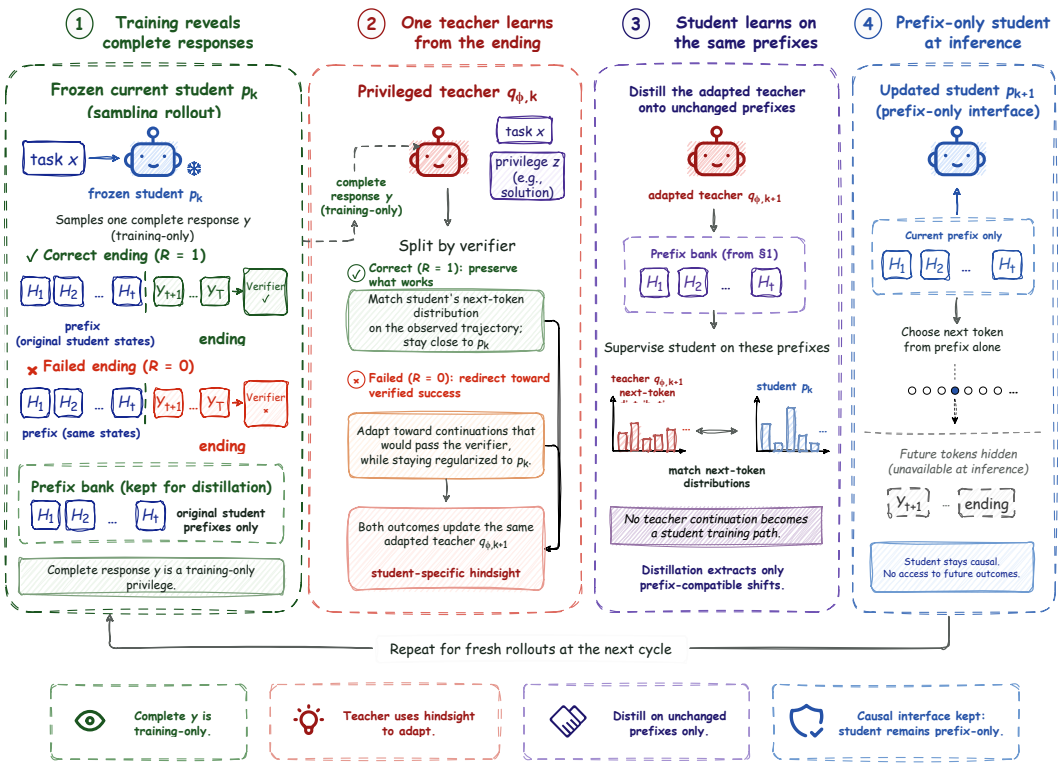

Figure 2: PAST adapts a privileged teacher from complete student responses and outcomes, then distills it on the student's original prefixes. Correct trajectories preserve the current policy. Failed trajectories direct verified improvement with regularization toward the student, while the student remains prefix-only.

where m_+ and m_- indicate whether the corresponding branch contains a valid training term. The observed ratio of correct to failed trajectories therefore does not set the branch weights implicitly.

3.4 Causal Student Distillation

The student receives the adapted teacher's supervision only on states generated by the frozen student. After one teacher update, PAST freezes $q_{\phi_{k+1}}$ and performs one Vanilla OPSD student update on the original states H_t. The implementation applies Eq. (11) in the forward direction $q_{\phi_{k+1}} \| \pi$ with the baseline vocabulary, threshold, mask, temperature, and reduction. Teacher-generated continuations from the failed branch are not student training paths. The updated policy becomes p_{k+1}, cycle-local rollouts are released, and the next cycle samples from the new student. Appendix B gives the complete update and snapshot lifecycle.

4 Theory

PAST gives its teacher access to a complete student trajectory but requires the deployed student to act from its current prefix. We characterize the exact causal target of forward-KL distillation, show why even a perfect privileged teacher can provide no useful student update, and give sufficient conditions under which a teacher change survives projection as a local improvement direction.

Fix one training cycle. Let (H, U) follow the distribution induced by the frozen student's on-policy rollouts, where H is a student prefix and U contains the complete trajectory and task privilege. Define

$$q_U(a \mid H) = q_\phi(a \mid H, U), \qquad \overline{q}(a \mid H) = \mathbb{E}[q_U(a \mid H) \mid H]. \tag{16}$$

The results use the unclipped population KL under the support and integrability conditions stated in Appendix C.

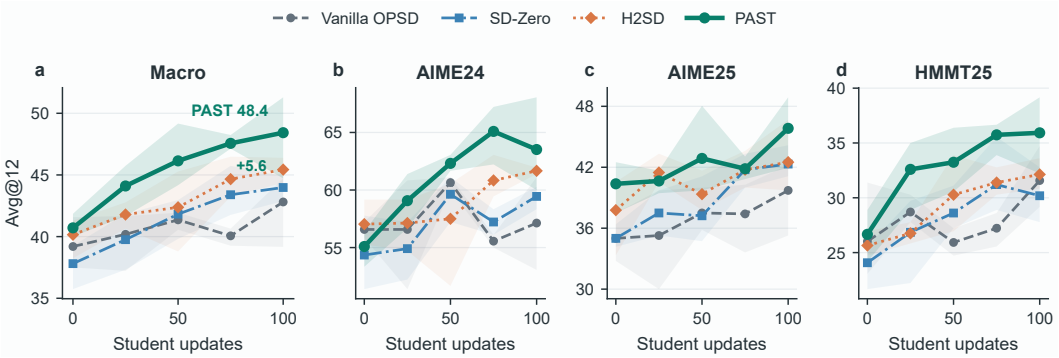

Figure 3: Avg@12 over student updates. Lines show three-seed means, bands show seed ranges, and Macro equally weights the three tasks.

Proposition 1 (Privileged-to-causal projection)**.** *For every prefix-only policy* $\pi(a \mid H)$*,*

$$\begin{aligned}\mathbb{E}[D_{\mathrm{KL}}(q_U\|\pi)] = \mathbb{E}[D_{\mathrm{KL}}(q_U\|\overline{q})] \\ + \mathbb{E}[D_{\mathrm{KL}}(\overline{q}\|\pi)]\,.\end{aligned} \tag{17}$$

Consequently, $\overline{q}$ *is the unique minimizer wherever it has positive support.*

Forward-KL distillation retains the conditional arithmetic mean of the teacher distributions conditioned on trajectories. The first term in Eq. (17) is variation that no prefix-only policy can reproduce. The second is the remaining approximation error to the causal target $\overline{q}$.

Corollary 2 (Projection onto a restricted student family)**.** *Let* Π *be any class of prefix-only policies for which the displayed expectations are finite. For every* $\pi \in \Pi$*,*

$$\begin{aligned}&\mathbb{E}[D_{\mathrm{KL}}(q_U\|\pi)] - \inf_{\rho\in\Pi} \mathbb{E}[D_{\mathrm{KL}}(q_U\|\rho)] \\ &= \mathbb{E}[D_{\mathrm{KL}}(\overline{q}\|\pi)] - \inf_{\rho\in\Pi} \mathbb{E}[D_{\mathrm{KL}}(\overline{q}\|\rho)].\end{aligned} \tag{18}$$

The two objectives have the same minimizing sequences and, whenever a minimum is attained, the same minimizers over Π*. Restricting the student family changes how closely it can represent* $\overline{q}$*, but not the causal target selected by forward-KL distillation. The residual error therefore reflects the approximation limit of* Π *rather than a different target induced by privileged variation.*

Corollary 3 (Privilege use and distillable shift)**.** *Let* $A \sim q_U(\cdot \mid H)$*. Then*

$$\mathbb{E}[D_{\mathrm{KL}}(q_U\|\overline{q})] = I(A; U \mid H), \tag{19}$$

and, relative to the frozen student p_k*,*

$$\mathbb{E}[D_{\mathrm{KL}}(q_U\|p_k)] = I(A; U \mid H) + \mathbb{E}[D_{\mathrm{KL}}(\overline{q}\|p_k)]. \tag{20}$$

Equation (20) separates teacher change into privileged variation and the causal shift available to the student. Privileged variation can consume the entire teacher–student divergence without moving the student target.

Proposition 4 (Perfect privileged teachers need not transfer)**.** *For every horizon* $T \geq 2$*, there is a binary verification problem, a prefix-only student* p_k*, and a teacher that uses the student's complete trajectory such that the teacher succeeds with probability one, while* $\overline{q} = p_k$ *at every supervised prefix. Exact distillation leaves the student unchanged, and its success probability is* 2^{1-T}*.*

The gap can therefore approach one as the response grows. Teacher success alone gives no nontrivial lower bound on student improvement. Appendix C gives the construction, in which the teacher uses a future student token to coordinate an otherwise successful response while every projected next-token target remains uniform.

We next identify a positive transfer condition. Let $Q_k(H, a) \in [0, 1]$ be the probability of verifier success after taking action a at H and continuing with the frozen student. For any prefix-only distribution r, define its frozen-student value shift

$$g_k(r; H) = \sum_a \big(r(a \mid H) - p_k(a \mid H)\big) Q_k(H, a). \tag{21}$$

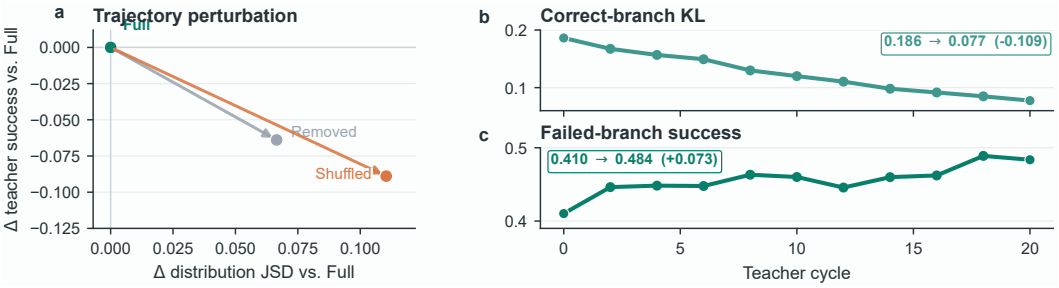

Figure 4: Matching trajectories improve teacher success and compatibility (a). Correct-branch KL falls (b) as failed-branch success rises (c).

Proposition 5 (Value-aligned causal transfer)**.** *For almost every prefix H,*

$$g_k(\overline{q}; H) = \mathbb{E}[g_k(q_U; H) \mid H]. \tag{22}$$

Thus changes across trajectories may cancel in distribution, but any positive mean shift under the frozen student's value function survives causal projection.

The two outcome branches make this condition method-specific. Let $\eta(H) = P(R = 0 \mid H)$ and let $\overline{q}_+$ and $\overline{q}_-$ denote the conditional mean teachers on correct and failed trajectories. Suppose

$$\begin{aligned} \delta_+(H) &= \mathbb{E}[D_{\mathrm{KL}}(p_k \| q_U) \mid H, R = 1], \\ \gamma_-(H) &= g_k(\overline{q}_-; H), \qquad \delta_S(H) = D_{\mathrm{KL}}(\overline{q} \| \pi). \end{aligned} \tag{23}$$

Corollary 6 (Outcome-conditioned transfer margin)**.** *For any prefix-only student π,*

$$g_k(\pi; H) \geq \eta\gamma_- - (1-\eta)\sqrt{\delta_+/2} - \sqrt{\delta_S/2}. \tag{24}$$

The projected update has positive frozen-student value whenever the failed-branch margin exceeds the preservation and distillation errors on the right-hand side.

The correct branch makes $\delta_+ = 0$ at its ideal unclipped distributional optimum. A mean correct-branch objective at most δ implies mean total variation at most $\sqrt{\delta/2}$. The failed branch must supply $\gamma_- > 0$, not merely high teacher success. Its proximity term targets the needed student compatibility. Appendix C proves a compatibility bound for entire trajectories and connects the regularizer on successful paths to this condition.

Finally, the value margin has a rollout-level implication. For the conservative interpolation $\pi_\epsilon = (1-\epsilon)p_k + \epsilon\pi$ in a finite-horizon problem,

$$\frac{d}{d\epsilon} J(\pi_\epsilon)\Big|_{\epsilon=0} = \sum_t \mathbb{E}_{H_t \sim p_k}[g_k(\pi; H_t)]. \tag{25}$$

If the right-hand side is positive, sufficiently small $\epsilon > 0$ improves verifier success. The theorem identifies the transferable quantity and the positive margin required for local improvement. The implemented update targets this direction through pointwise clipping and a shared-parameter policy. The trajectory perturbations in Section 5.4 test privilege use, while the factorial and objective studies test whether teacher adaptation produces a better causal student rather than only a more successful privileged teacher.

5 Experiments

We evaluate student improvement and how trajectory access and teacher adaptation contribute to it.

5.1 Setup

We train Qwen3-1.7B in thinking mode for 100 student updates with seeds 17, 29, and 43. Evaluation uses AIME 2024, AIME 2025, and HMMT 2025 at temperature 1.0 with 12 samples per problem. We report task Avg@12 and their equally weighted macro average. All methods share the student-update count and training protocol. Appendix D gives the complete protocol, uncertainty analysis, and checks.

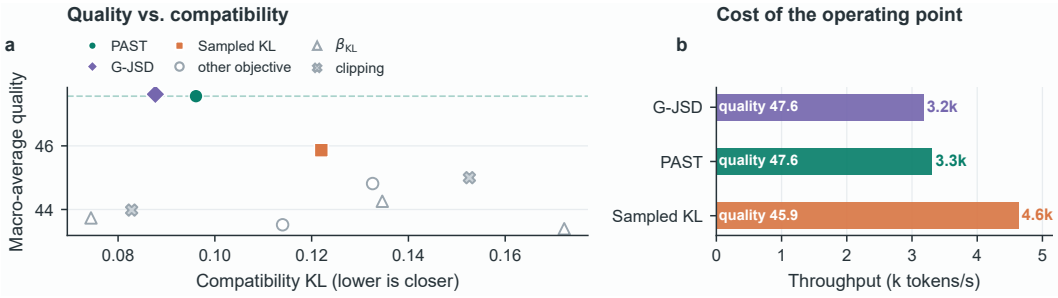

Figure 5: PAST and generalized JSD retain high quality at low compatibility KL, while sampled-token KL trades quality for throughput.

5.2 PAST Improves the Causal Student

PAST produces the strongest final student in the comparison set, reaching an Avg@12 macro average of 48.426 compared with 42.809 for Vanilla OPSD. This is a 5.617-point improvement, and PAST leads on each of the three benchmarks (Table 1). PAST also exceeds H2SD in every training seed; Appendix E reports the seed-level results and uncertainty analysis.

Table 1: Final Avg@12 after 100 updates. PAST has the highest macro average and exceeds Vanilla OPSD by 5.617 points.

Method	AIME24	AIME25	HMMT25	Macro
Base	53.056	34.907	27.870	38.611
SFT	55.000	37.222	27.407	39.877
GRPO	55.463	38.056	27.963	40.494
Vanilla OPSD	57.130	39.722	31.574	42.809
SD-Zero	59.444	42.315	30.185	43.981
H2SD	61.667	42.500	32.130	45.432
PAST	**63.519**	**45.833**	**35.926**	**48.426**

PAST's advantage develops over training. Its macro average rises at every checkpoint and progressively separates from Vanilla OPSD (Figure 3). The HMMT25 gain appears early; AIME24 and AIME25 vary more across checkpoints, but PAST finishes with the highest mean on all three tasks.

5.3 Factorial Attribution of the PAST Gain

The 2×2 design separates trajectory access from teacher adaptation. The cell without either factor is Vanilla OPSD. Trajectory-only changes the teacher input; teacher-only adapts without the completed response; PAST combines the matching trajectory with adaptation.

Neither factor alone reproduces the gain. Trajectory access changes the teacher context but adds only 0.555 macro points, while adaptation without the completed response adds 0.092 points. Each condition lowers at least one task. PAST instead reaches 48.426 and improves all three. The macro interaction is $+4.970$ points and is positive on every task. This pattern supports complementary roles: the matching trajectory supplies hindsight about the student, and adaptation converts it into a more useful target.

Table 2: PAST combines trajectory access and teacher adaptation to produce the strongest student.

Trajectory	Trained teacher	AIME24	AIME25	HMMT25	Macro
No	No	57.130	39.722	31.574	42.809
No	Yes	59.074	38.056	31.574	42.901
Yes	No	59.537	41.111	29.444	43.364
Yes	Yes	**63.519**	**45.833**	**35.926**	**48.426**

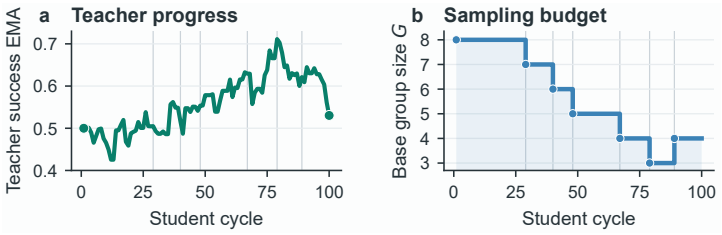

Figure 6: Adaptive sampling reduces its base group size as teacher-success EMA rises over 100 student cycles.

5.4 The Teacher Uses Matching Trajectory Privilege

The teacher depends on the trajectory matching the current response. Teacher success is its continuation pass rate, while distribution JSD measures next-token change. Removing the trajectory tests the student context; shuffling across problems retains the field but breaks its correspondence with the problem and prefix, separating matched hindsight from an additional input alone.

Removing the trajectory lowers verifier success by 0.064, while shuffling it causes the larger success drop of 0.089 and raises JSD by 0.110 (Figure 4a). The matching condition also gives the lowest student perplexity, with format validity above 0.96 in all conditions. During adaptation, correct-branch KL decreases and failed-branch success increases (Figures 4b and 4c), showing that the two objectives move in their intended directions. The factorial comparison connects this teacher adaptation to final student performance.

5.5 Joint Outcome-Conditioned Adaptation Performs Best

Training jointly on correct and failed trajectories leads every task, reaching 48.426 macro versus 45.864 for failed-only and 44.908 for correct-only training (Table 7). Its 2.562-point advantage over the stronger single branch supports joint adaptation. Both branches train the same teacher, allowing supervision from successful student behavior to accompany the search for verified continuations on failed trajectories. The combined objective supplies both signals before each student update.

5.6 Objective Robustness, Sampling Efficiency, and Scale

PAST remains effective across alternative distribution objectives and reduces teacher sampling cost, while the 4B extension shows task-dependent gains.

Full-vocabulary KL and generalized JSD yield comparable student performance, whereas sampled-token KL improves throughput at a quality cost (Figure 5). Adaptive sampling uses 17.6% fewer teacher samples, 850 instead of 1,032, and slightly improves the macro average from 46.728 to 47.284 (Table 9). It reduces routine sampling after teacher success stabilizes and adds continuations for failed groups that still lack a successful path (Figure 6). In the one-seed Qwen3-4B extension, PAST raises the macro average from 47.685 to 48.519, with gains on AIME25 and HMMT25 and a lower AIME24 score (Table 11). Appendices E and F report the full ablations, optimization diagnostics, scale results, and training-cost accounting.

6 Conclusion

PAST uses completed on-policy responses as teacher privilege while retaining the causal student interface. It preserves correct behavior and adapts failed trajectories toward verifier success before distilling on the student's original prefixes. Our analysis identifies the causal projection of this privileged teacher and conditions for improving verified success. Across three mathematical reasoning benchmarks, PAST improves Avg@12 over Vanilla OPSD by 5.6 points. The factorial study establishes the final-student gain from combining trajectory access and teacher adaptation, while trajectory perturbations show that the adapted teacher uses the matching hindsight context.

REFERENCES

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

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

Qiye Cai, Yichuan Ma, Peiji Li, Yongkang Chen, Qipeng Guo, Yicheng Zou, Linyang Li, Xiaocheng Feng, and Bing Qin. H2sd: Hybrid hindsight self-distillation. *arXiv preprint arXiv:2607.18955*, 2026.

Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3430–3437, 2020. doi: 10.1609/aaai.v34i04.5746.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, pp. 6621–6642, 2024.

Wojciech Marian Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant M. Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. *arXiv preprint arXiv:1902.02186*, 2019.

Ken Ding. HDPO: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. In *International Conference on Learning Representations*, 2024.

Zhuowen Han, Jinwei Xiao, Zhengxi Lu, Renren Jin, Zhiyuan Yao, Yuxin Liu, Hongyan Hao, Yueqing Sun, Yu Yang, Qi Gu, Xunliang Cai, and Deyi Xiong. Distill where you fail: Recovering learning signals of negative rl-groups from adaptive teacher guidance. *arXiv preprint arXiv:2608.00782*, 2026.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-STaR: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, 2024.

Yu Li, Shu Hong, and Tian Lan. Localizing credit at the divergence: Path-conditioned self-distillation for llm reasoning. *arXiv preprint arXiv:2606.15576*, 2026.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133, 2020.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026.

David Lopez-Paz, Leon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Jun Rao, Xuebo Liu, Zepeng Lin, Liang Ding, Jing Li, Dacheng Tao, and Min Zhang. Exploring and enhancing the transfer of distribution in knowledge distillation for autoregressive language models. *Knowledge-Based Systems*, 337:115382, 2026.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

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.

Guobin Shen, Lei Huang, Xiang Cheng, Chenxiao Zhao, Jindong Li, Dongcheng Zhao, and Xing Yu. From generic correlation to input-specific credit in on-policy self distillation. *arXiv preprint arXiv:2605.11613*, 2026.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Zhiquan Tan and Yinrong Hong. Paint: Partial-solution adaptive interpolated training for self-distilled reasoners. *arXiv preprint arXiv:2604.26573*, 2026.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026. doi: 10.18653/v1/2026.findings-acl.1276.

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. *arXiv preprint arXiv:2601.18734*, 2026a.

Xuyang Zhao, Liting Zhang, Zichen Xu, Zhihu Wang, Caiyue Xu, Shiwan Zhao, and Qicheng Li. Is more privileged information better? from solution traces to problem-solving structure in self-distilled reasoning. *arXiv preprint arXiv:2608.01589*, 2026b.

A Related Work

Autoregressive and on-policy distillation. Knowledge distillation transfers predictive distributions from a teacher to a student, and sequence-level distillation instead transfers teacher-generated sequences (Hinton et al., 2015; Kim & Rush, 2016). Policy distillation studies state-conditioned action distributions, while interactive imitation learning motivates collecting supervision on states visited by the learner (Rusu et al., 2015; Czarnecki et al., 2019; Ross et al., 2011). Imitation-based and on-policy language-model distillation combine these ideas by training on student-generated states (Lin et al., 2020; Agarwal et al., 2024). MiniLLM and DistiLLM study divergence choice and efficient optimization for autoregressive language models (Gu et al., 2024; Ko et al., 2024), and prefix OPD reduces training cost by truncating the supervised rollout (Zhang et al., 2026). PAST retains this student-state interface and changes how the teacher is formed.

Context-conditioned and privileged self-distillation. Context distillation trains an unconditioned policy to recover behavior induced by information available only to a teacher (Snell et al., 2022). OPSD supplies a shared-parameter teacher with a verified solution and evaluates it on student prefixes (Zhao et al., 2026a); SDFT and OPCD similarly distill demonstration- or context-conditioned behavior on policy (Shenfeld et al., 2026; Ye et al., 2026). Learning using privileged information and generalized distillation provide the broader information-asymmetry framework (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). Within this framework, PAST uses the matching completed student trajectory as additional teacher-only context.

Complete-trajectory hindsight and outcome routing. Iterative self-improvement already reuses model-generated behavior in several ways. STaR trains on successful self-generated rationales, ReST^{EM} filters generated solutions with binary feedback, and V-STaR uses correct and incorrect solutions to train a verifier (Zelikman et al., 2022; Singh et al., 2023; Hosseini et al., 2024). SPIN and self-rewarding language models further improve across cycles through self-generated responses or feedback (Chen et al., 2024; Yuan et al., 2024). Complete responses also provide more direct hindsight supervision. SD-Zero revises a completed response before distilling the revision (He et al., 2026). HSD conditions the teacher on a successful peer rollout and localizes supervision near the divergence from a failed path (Li et al., 2026). H2SD routes correct and failed trajectories through distinct supervision mechanisms (Cai et al., 2026), while HDPO and RSTG target all-failure or zero-variance negative groups with privileged teacher guidance (Ding, 2026; Han et al., 2026). PAST learns one trajectory-privileged OPSD teacher across both outcome regimes. Correct trajectories preserve the student's successful distribution, failed trajectories train the teacher toward verifier success under student regularization, and the adapted teacher supervises the original causal prefixes. The factorial controls show that matching-trajectory access and teacher adaptation work together to produce the final student gain.

Trainable teachers and supervision design. Online KD and OKD establish that teachers or teacher modules can be trained during distillation (Chen et al., 2020; Rao et al., 2026). More directly, π-Distill jointly trains a privileged teacher and an unprivileged student from successful action information in agentic environments (Penaloza et al., 2026). Other OPSD variants alter the privileged representation or select signals by problem-solving structure, partial solutions, position reliability, token teachability, or input-specific credit (Zhao et al., 2026b; Tan & Hong, 2026; Liu et al., 2026; Wang et al., 2026; Shen et al., 2026). Analyses of thinking models further show that privileged context can reshape branching and self-correction behavior (Kaur et al., 2026).

B Method and Protocol Details

B.1 Comparison of Student-Facing Interfaces

Table 3 separates the source of the teacher distribution from the states used for the student update. All listed distillation methods train on prefixes sampled from the student. Their main differences lie in teacher construction, privileged context, and the role of outcome feedback.

Table 3: Comparison of training interfaces for the student. PAST uses the completed response to adapt the teacher while retaining the original student prefixes for distillation.

Method	Teacher construction	Teacher context beyond the prefix	Outcome role	Student update states
OPD (Agarwal et al., 2024; Gu et al., 2024)	Fixed external teacher	No additional privilege	Not required	Student prefixes
OPSD (Zhao et al., 2026a)	Stop-gradient privileged self-teacher	Task-side privilege Z	Not required	Student prefixes
SD-Zero (He et al., 2026)	Self-revision-trained teacher, frozen during distillation	Complete response and outcome-conditioned instruction	Enters the teacher context	Student prefixes
H2SD (Cai et al., 2026)	Fixed self-teacher with offline hints	Verified response or external hint	Selects the success or failure objective	Student prefixes
PAST	Cycle-adapted privileged self-teacher	Task privilege Z and complete response Y	Selects teacher adaptation but is excluded from U	Original student prefixes

B.2 Teacher Input Serialization

All teacher phases use one user message and the model's native chat template. The serialized fields appear in the following order:

```
Problem: {problem}

Please reason step by step, and put your final answer within \boxed{}.

=== Student Attempt Begin ===
{student_output}
=== Student Attempt End ===

=== Reference Solution Begin ===
{reference_solution}
=== Reference Solution End ===

The student attempt above may be correct or incorrect. Use it as hindsight
context for the student's reasoning state. Maintain the student's established
reasoning style and presentation whenever they are compatible with a correct
solution. Use the reference solution to ensure correctness, but do not copy or
paraphrase the reference solution. Now solve the original problem through your
own reasoning, and put the final answer within \boxed{}.
```

The student attempt is always produced by the current unprivileged student. Correct and failed trajectories use the same message; correctness is kept outside the prompt and controls only routing and verification. In the correct branch and student distillation phase, the assistant segment is teacher-forced on the original student response. In the failed branch, the teacher samples the assistant segment autoregressively.

B.3 Adaptive Failed-Branch Sampling

Let $G_{\max}$ be the fixed group size of the GRPO baseline and let $G_{\text{base}} \in [1, G_{\max}]$ be the current base draw. Warm-up starts at $G_{\text{base}} = G_{\max}$. Once the recent teacher success rate stabilizes above the configured threshold, the controller decreases G_{base} one step at a time, down to one. For each failed student trajectory, PAST applies the following state machine.

1. Draw G_{base} teacher continuations.
2. If at least one succeeds, classify the realized group as mixed or all-success and compute Eq. (14).
3. If all fail, add one continuation at a time until the first success or until the group reaches $G_{\max}$.

4. If a size-$G_{\max}$ group remains all-failure, redraw one complete size-$G_{\max}$ group. Skip the trajectory only when this retry also contains no success.

A mixed group uses relative advantages and successful-path KL. An all-success group, including the single-success case, has zero relative advantage and uses only successful-path KL. An all-failure group contributes neither GRPO nor student-proximity KL.

B.4 Loss Reduction and Clipping Diagnostics

For N_+ verified correct trajectories and a set $\mathcal{G}_-^{\mathrm{act}}$ of active failed groups, the branch means are

$$\overline{\mathcal{L}}_+ = \frac{1}{N_+} \sum_{i:R_i=1} \mathcal{L}_{+,i}, \qquad \overline{\mathcal{L}}_- = \frac{1}{|\mathcal{G}_-^{\mathrm{act}}|} \sum_{g \in \mathcal{G}_-^{\mathrm{act}}} \mathcal{L}_{-,g}. \tag{26}$$

Only existing means enter Eq. (15); the optimizer step is skipped when neither branch is active. Correct-teacher and student-distillation losses clip individual vocabulary contributions at the Vanilla OPSD threshold. Failed-path KL remains exact. Training logs retain the clipped loss, unclipped KL, clip fraction, removed positive mass, failed-path KL quantiles, and non-finite counts.

B.5 Student and Teacher Snapshot Lifecycle

PAST uses one frozen base model with separate student and teacher LoRA adapters. At the start of cycle k, the student adapter is the read-only policy p_k and the teacher adapter is q_k. The student adapter generates the rollout; only the teacher adapter is trainable during teacher adaptation. PAST then freezes q_{k+1} and makes the student adapter trainable for one logical update from p_k to p_{k+1}. Gradient accumulation may split either logical update into microbatches but does not add optimizer steps or replay rollouts.

Checkpoints are written only at complete cycle boundaries and contain both adapters, both optimizer and scheduler states, the adaptive-sampling controller, data position, random-number-generator states, and configuration identifiers. An interrupted partial cycle is discarded and restarted from the preceding boundary. Full-vocabulary reference logits are recomputed under the appropriate frozen adapter and released after each microbatch rather than stored across phases.

C Complete Theory and Proofs

C.1 Conditions and Conditioning Measure

All expectations in Section 4 are taken under the joint distribution induced by the task distribution, the frozen student's on-policy rollout, the prefix-selection rule, and the adapted teacher's next-token distribution. For almost every prefix H, the conditional teacher distributions $q_U(\cdot \mid H)$ are defined on the common vocabulary $\mathcal{V}$. We assume that the displayed KL divergences and conditional expectations are finite. Uniqueness holds on actions with positive mass under the corresponding target distribution. These conditions let conditional expectation and finite-vocabulary summation be interchanged below.

The value results additionally assume a finite-horizon generation problem with terminal verifier reward in $[0, 1]$. For a frozen student p_k, $Q_k(H, a)$ is the expected terminal reward after taking a at H and following p_k thereafter. Conditional branch means are needed only when the corresponding event has positive probability. Terms multiplied by a zero branch probability are defined as zero.

C.2 Proof of the Causal Projection

Proof of Proposition 1. For any prefix-only π, insert $\overline{q}$ into the log ratio

$$\log \frac{q_U(A \mid H)}{\pi(A \mid H)} = \log \frac{q_U(A \mid H)}{\overline{q}(A \mid H)} + \log \frac{\overline{q}(A \mid H)}{\pi(A \mid H)}. \tag{27}$$

Taking expectations gives the first term in Eq. (17). For the second term, conditioning on H replaces $q_U(a \mid H)$ by $\mathbb{E}[q_U(a \mid H) \mid H] = \overline{q}(a \mid H)$, which yields $\mathbb{E}[D_{\mathrm{KL}}(\overline{q} \| \pi)]$. Nonnegativity establishes optimality, and equality requires $\pi = \overline{q}$ wherever $\overline{q}$ has positive support. □

Proof of Corollary 2. Proposition 1 writes the privileged-teacher objective as

$$\mathbb{E}[D_{\mathrm{KL}}(q_U\|\pi)] = \mathbb{E}[D_{\mathrm{KL}}(q_U\|\overline{q})] + \mathbb{E}[D_{\mathrm{KL}}(\overline{q}\|\pi)].$$

The first term is independent of π. Subtracting the infimum over $\rho \in \Pi$ gives Eq. (18). The excess risks are identical pointwise, so the objectives have the same minimizing sequences and the same minimizers whenever a minimum is attained. If $\overline{q} \in \Pi$, the remaining KL is minimized uniquely by $\overline{q}$ on its support. □

Proof of Corollary 3. Under the joint law $P(A, U \mid H) = P(U \mid H)q_U(A \mid H)$, the conditional marginal of A is $\overline{q}(A \mid H)$. The definition of conditional mutual information gives Eq. (19). Applying Proposition 1 with $\pi = p_k$ gives Eq. (20). □

Whenever the expected teacher–student KL is positive, Eq. (20) also gives the descriptive fraction

$$\frac{\mathbb{E}[D_{\mathrm{KL}}(\overline{q}\|p_k)]}{\mathbb{E}[D_{\mathrm{KL}}(q_U\|p_k)]} = 1 - \frac{I(A;U \mid H)}{\mathbb{E}[D_{\mathrm{KL}}(q_U\|p_k)]}. \tag{28}$$

This ratio is an interpretation of the decomposition, not an objective optimized explicitly by PAST.

C.3 A Perfect Privileged Teacher Need Not Transfer

Proof of Proposition 4. Fix $T \geq 2$. A response is a binary string $C = (C_1, \ldots, C_T)$, and the verifier accepts exactly the two constant strings

$$V(C) = \mathbf{1}\{C_1 = \cdots = C_T\}. \tag{29}$$

Let the frozen student generate independent fair bits. Its success probability is therefore $2/2^T = 2^{1-T}$.

Draw the student's complete training trajectory $Y \sim p_k$ and include it in U. Let $B = Y_T$, which is available to the privileged teacher but is absent from every student prefix $H_t = Y_{<t}$. Define the teacher to output B deterministically at every position,

$$q_U(a \mid H_t) = \mathbf{1}\{a = B\}. \tag{30}$$

When this teacher generates a response under the fixed privilege U, it produces $(B, \ldots, B)$ and succeeds with probability one.

For every supervised position $t \leq T$, the fair bit $B = Y_T$ is independent of $H_t = Y_{<t}$. Hence

$$\overline{q}(a \mid H_t) = \mathbb{E}[\mathbf{1}\{a = B\} \mid H_t] = \tfrac{1}{2} = p_k(a \mid H_t). \tag{31}$$

The exact forward-KL projection is the original student at every supervised prefix, so distillation makes no update. Moreover, $A = B$ under the teacher and B remains uniform given H_t, which gives $I(A;U \mid H_t) = \log 2$ while $D_{\mathrm{KL}}(\overline{q}\|p_k) = 0$. The privileged teacher is perfect, yet its advantage over the causal student approaches one as T grows. □

The construction uses the future trajectory as a coordination variable. Its success is a property of repeatedly conditioning on the same unavailable bit, not of a next-token change that a prefix-only student can retain.

C.4 Value-Aligned Causal Transfer

Proof of Proposition 5. Condition on H. Since $Q_k(H, a)$ and $p_k(a \mid H)$ do not depend on U,

$$\mathbb{E}[g_k(q_U; H) \mid H] = \sum_a \left(\mathbb{E}[q_U(a \mid H) \mid H] - p_k(a \mid H)\right) Q_k(H, a) \tag{32}$$

$$= \sum_a \left(\overline{q}(a \mid H) - p_k(a \mid H)\right) Q_k(H, a) \tag{33}$$

$$= g_k(\overline{q}; H). \tag{34}$$

Thus conditional averaging preserves the mean of every bounded prefix-action value shared across privilege conditions. Distributional cancellation erases value only when positive and negative value shifts cancel as well. □

Let $\eta = P(R = 0 \mid H)$. The conditional barycenter obeys

$$\overline{q} = (1 - \eta)\overline{q}_+ + \eta\overline{q}_-, \tag{35}$$

and linearity gives

$$g_k(\overline{q}; H) = (1 - \eta)g_k(\overline{q}_+; H) + \eta g_k(\overline{q}_-; H). \tag{36}$$

At the ideal correct-branch optimum $\overline{q}_+ = p_k$, the full projected shift is $\overline{q} - p_k = \eta(\overline{q}_- - p_k)$. In particular,

$$\mathrm{TV}(\overline{q}, p_k) = \eta\,\mathrm{TV}(\overline{q}_-, p_k). \tag{37}$$

Failed-branch signal is therefore attenuated by its conditional frequency even before student approximation error is introduced.

Proof of Corollary 6. Because $Q_k(H, \cdot) \in [0, 1]$, the difference in its expectation under two action distributions is at most their total variation. Convexity of total variation, Pinsker's inequality, and Jensen's inequality give

$$\begin{aligned} g_k(\overline{q}_+; H) &\geq -\,\mathrm{TV}(\overline{q}_+, p_k) && (38)\\ &\geq -\mathbb{E}[\mathrm{TV}(q_U, p_k) \mid H, R = 1] && (39)\\ &\geq -\sqrt{\delta_+(H)/2}. && (40) \end{aligned}$$

Using Eq. (36) and $\gamma_- = g_k(\overline{q}_-; H)$ yields

$$g_k(\overline{q}; H) \geq \eta\gamma_- - (1 - \eta)\sqrt{\delta_+/2}. \tag{41}$$

The same bounded-value argument and Pinsker's inequality give

$$|g_k(\pi; H) - g_k(\overline{q}; H)| \leq \mathrm{TV}(\pi, \overline{q}) \leq \sqrt{D_{\mathrm{KL}}(\overline{q}\|\pi)/2}. \tag{42}$$

Substituting $\delta_S = D_{\mathrm{KL}}(\overline{q}\|\pi)$ proves Eq. (24). □

C.5 Compatibility Between Privileged and Student Continuations

The failed teacher is evaluated under its own privileged continuation, whereas g_k evaluates the teacher's next-token distribution followed by the frozen student. A continuation-distance condition connects these quantities.

Proposition 7 (Continuation compatibility)**.** *Fix (H, U). Let $\mathbb{Q}_U$ be the distribution of a complete continuation generated by q_U. Let $\mathbb{M}_U$ draw its first action from $q_U(\cdot \mid H)$ and all later actions from p_k. Write $s_U = \mathbb{E}_{\mathbb{Q}_U}[V]$ and $v_k = \mathbb{E}_{p_k}[V \mid H]$. Then*

$$g_k(q_U; H) \geq s_U - v_k - \sqrt{D_{\mathrm{KL}}(\mathbb{Q}_U\|\mathbb{M}_U)/2}. \tag{43}$$

Moreover, the trajectory KL has the chain-rule form

$$D_{\mathrm{KL}}(\mathbb{Q}_U\|\mathbb{M}_U) = \mathbb{E}_{C\sim\mathbb{Q}_U}\left[\sum_{s>t} D_{\mathrm{KL}}(q_U(\cdot \mid G_s)\|p_k(\cdot \mid G_s))\right], \tag{44}$$

where t is the position following H and G_s is the generated prefix at position s.

Proof. The hybrid success probability is

$$\mathbb{E}_{\mathbb{M}_U}[V] = \sum_a q_U(a \mid H)Q_k(H, a) = v_k + g_k(q_U; H). \tag{45}$$

Since $V \in [0, 1]$, its expectation differs under $\mathbb{Q}_U$ and $\mathbb{M}_U$ by at most their total variation. Pinsker's inequality gives

$$s_U - \mathbb{E}_{\mathbb{M}_U}[V] \leq \mathrm{TV}(\mathbb{Q}_U, \mathbb{M}_U) \leq \sqrt{D_{\mathrm{KL}}(\mathbb{Q}_U\|\mathbb{M}_U)/2}, \tag{46}$$

which proves Eq. (43). The two trajectory laws use the same first-action distribution. Applying the KL chain rule to all later positions gives Eq. (44). □

Proposition 7 formalizes the role of student proximity. High privileged success gives a positive compatible value margin only when the teacher's advantage exceeds its continuation mismatch. The implemented quantity in Eq. (13) evaluates the same tokenwise divergence on sampled paths that pass the verifier. Because that loss is conditioned on success rather than the unconditional trajectory KL in Eq. (44), it encourages compatibility but does not certify this bound globally.

C.6 Local Rollout Improvement

Proposition 8 (Local policy improvement)**.** *Let $J(r)$ denote expected terminal verifier reward under a finite-horizon prefix-only policy r. For any prefix-only π, define $\pi_\epsilon = (1-\epsilon)p_k + \epsilon\pi$. Then Eq. (25) holds. If its right-hand side is strictly positive, there exists $\epsilon_0 > 0$ such that $J(\pi_\epsilon) > J(p_k)$ for every $\epsilon \in (0, \epsilon_0)$.*

Proof. Because the horizon and vocabulary are finite, $J(\pi_\epsilon)$ is a finite sum of products of action probabilities and is differentiable at $\epsilon = 0$. Differentiating one policy factor at a time leaves p_k on every preceding and subsequent factor. The contribution at position t is

$$\mathbb{E}_{H_t \sim p_k}\left[\sum_a \big(\pi(a \mid H_t) - p_k(a \mid H_t)\big) Q_k(H_t, a)\right] = \mathbb{E}_{H_t \sim p_k}[g_k(\pi; H_t)]. \tag{47}$$

Summing over positions proves Eq. (25). A differentiable function with a strictly positive derivative at zero is larger than its value at zero throughout some sufficiently small right neighborhood, which proves the second claim. □

This result concerns a conservative policy-space interpolation. It does not identify a finite parameter-space gradient step with π_ϵ, nor does it establish improvement when the occupancy-weighted margin is nonpositive.

C.7 Forward and Reverse Projection

Forward and reverse KL aggregate privileged targets differently. Proposition 1 gives the arithmetic barycenter

$$\arg\min_\pi \mathbb{E}[D_{\mathrm{KL}}(q_U \| \pi)] = \mathbb{E}[q_U \mid H]. \tag{48}$$

For reverse KL, direct Lagrangian optimization gives

$$\arg\min_\pi \mathbb{E}[D_{\mathrm{KL}}(\pi \| q_U)] \propto \exp(\mathbb{E}[\log q_U \mid H])\,. \tag{49}$$

PAST uses the divergence direction that preserves the conditional mean soft target. Reverse KL instead suppresses actions that are not supported consistently across hindsight conditions.

C.8 Correct-Branch Stability and the Clipping Boundary

Proposition 9 (Correct trajectories are ideal fixed points)**.** *On verified correct trajectories, consider the unclipped population objective*

$$\mathcal{L}_+(q) = \mathbb{E}[D_{\mathrm{KL}}(p_k \| q) \mid R = 1]. \tag{50}$$

Its distributional optimum is $q = p_k$ on the supported prefixes. If $\mathcal{L}_+(q) \le \delta$, then

$$\mathbb{E}[\mathrm{TV}(p_k, q) \mid R = 1] \le \sqrt{\delta/2}. \tag{51}$$

Proof of Proposition 9. Nonnegativity of KL gives the unique supported distributional optimum $q = p_k$. For the approximate statement, Pinsker's inequality gives $\mathrm{TV}(p_k, q) \le \sqrt{D_{\mathrm{KL}}(p_k \| q)/2}$ at each condition. Jensen's inequality and the assumed mean KL bound yield Eq. (51). □

For a vocabulary distribution p and threshold $\tau > 0$, the pointwise-clipped objective in Eq. (11) equals the true KL whenever

$$q(v) \ge p(v)\exp[-\tau/p(v)] \quad \text{for every } v \text{ with } p(v) > 0. \tag{52}$$

The equality point $q = p$ lies inside this region, so the clipped and unclipped objectives have the same local gradient and Hessian around the correct fixed point. The clipped quantity is not a global divergence. For example, with $p = (0.9, 0.1)$, $q = (0.001, 0.999)$, and $\tau = 0.05$, the true KL is 5.892, whereas the clipped sum is $-0.180 < 0$. No global KL claim in this paper uses the clipped objective.

C.9 Adaptive-Sampling Identities

Suppose independent teacher continuations succeed with probability $s \in (0, 1)$ and the maximum group size is G. A fixed group is all-failure with probability $(1 - s)^G$, all-success with probability s^G, and mixed otherwise. If PAST begins with b samples and adds one sample after each all-failure prefix, the expected number of first-stage samples is

$$\mathbb{E}[N_1] = b + \sum_{j=b}^{G-1} (1 - s)^j. \tag{53}$$

When a complete all-failure group triggers one independent size-G retry, the total expected sample count and final skip probability are

$$\mathbb{E}[N] = b + \sum_{j=b}^{G-1} (1 - s)^j + G(1 - s)^G, \qquad P(\text{skip}) = (1 - s)^{2G}. \tag{54}$$

These identities explain the sampling reduction but do not make the success-masked, retry-conditioned GRPO estimator unbiased for an unconditional population objective.

D Full Experimental Protocol

D.1 Training and Evaluation

The main experiments use Qwen3-1.7B in thinking mode. Each method receives 100 logical student updates and is evaluated at checkpoints 0, 25, 50, 75, and 100. Training seeds are 17, 29, and 43. Evaluation covers AIME 2024, AIME 2025, and HMMT 2025 with temperature 1.0 and 12 generations per problem. Avg@12 is computed within each task, and the macro average weights the three tasks equally. The primary uncertainty analysis uses 10,000 stratified paired-bootstrap replicates with bootstrap RNG seed 29. Within each training-seed and benchmark stratum, paired item-level Avg@12 differences are resampled with replacement. Each replicate averages within strata and then equally weights the nine stratum means. We report the 2.5th and 97.5th percentiles as the 95% confidence interval.

The verifier parses the boxed final answer and compares it with the authorized task answer. Truncated, malformed, or unparseable responses are verifier failures. The resulting signal evaluates final-answer correctness rather than intermediate reasoning. All main comparisons share the student-update count and evaluation protocol. Table 12 reports teacher sampling costs separately.

D.2 Protocol and Semantic Checks

Table 4 summarizes repeated checks performed before the main run, including the observed end-to-end parity difference alongside semantic, numerical, and recovery checks.

D.3 Full-Vocabulary Implementation

Full-vocabulary KL is computed in PyTorch under the corresponding frozen adapter. Table 5 compares the training implementation with an FP32 reference. The main BF16 path has gradient cosine

Table 4: Protocol, numerical, and recovery checks, each repeated eight times.

Check	Observed mean	Passed	Repeats
Model and thinking mode	1.000000	8	8
Training-data protocol	1.000000	8	8
Evaluation protocol	1.000000	8	8
Batch-loss absolute error	0.000000	8	8
Batch-gradient cosine	0.999989	8	8
End-to-end paired difference	-0.102349	8	8
CPU semantic tests	1.000000	8	8
Resume and RNG recovery	1.000000	8	8

0.999457 and no non-finite events. Sequence chunking and the fused kernel are implementation alternatives; they do not change the stated objective.

The system records token IDs, masks, response lengths, verifier outcomes, failure types, cycle identifiers, and sampled-action log probabilities. Full-vocabulary reference distributions are recomputed for each microbatch and are not persisted across phases. Checkpoint recovery restores both adapters, optimizers, schedulers, sampling-controller state, data position, and random-number-generator states at a complete cycle boundary.

E Additional Results and Diagnostics

E.1 Learning Curves

Figure 7 reports changes from each method's own checkpoint-0 evaluation. This normalization isolates within-run learning and is not used for the absolute comparison in Table 1.

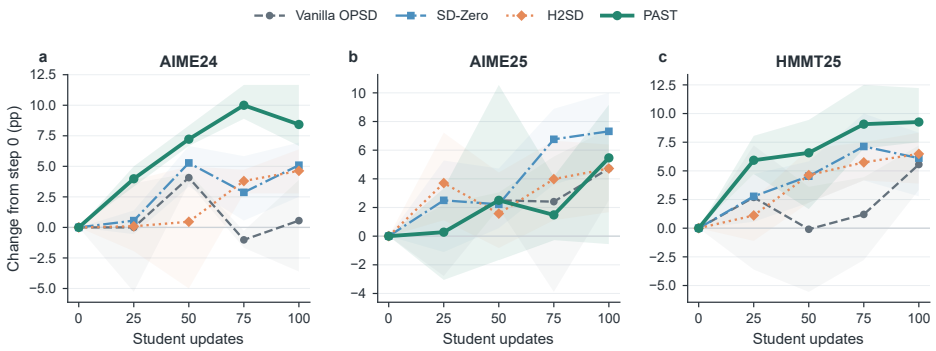

Figure 7: Avg@12 change from each method's own checkpoint 0. Lines show the mean over three training seeds and shaded regions show the seed range.

Table 6 gives the final macro average for each training seed. PAST has the highest value in every seed, exceeding H2SD by 4.908, 3.704, and 0.370 points. The primary stratified paired bootstrap estimates the PAST–Vanilla OPSD difference at 5.617 points, with a 95% confidence interval of $[2.592, 8.333]$.

E.2 Correct and Failed Branches

Joint training has the strongest point estimate in every task and improves the macro average by 2.562 points over the stronger single branch (Table 7).

Table 5: Numerical accuracy and system measurements for full-vocabulary KL implementations.

Implementation	Loss error	Grad. cosine	GB	Tokens/s	Non-finite
FP32 reference	0.000000	0.999986	72.305	1854	0
BF16 full vocabulary	0.002978	0.999457	61.718	3306	0
Sequence chunking	0.002525	0.999558	49.967	2881	0
Fused kernel	0.003780	0.999196	55.156	3709	0

Table 6: Final Avg@12 macro average by training seed.

Method	Seed 17	Seed 29	Seed 43	Mean
Vanilla OPSD	44.722	44.537	39.167	42.809
SD-Zero	42.870	44.908	44.167	43.981
H2SD	46.389	45.555	44.352	45.432
PAST	**51.297**	**49.259**	**44.722**	**48.426**

Table 7: Correct-only, failed-only, and joint teacher adaptation.

Training branch	AIME24	AIME25	HMMT25	Macro
Correct only	60.093	43.611	31.019	44.908
Failed only	62.037	43.333	32.222	45.864
Full PAST	**63.519**	**45.833**	**35.926**	**48.426**

Figure 8 shows that the correct-branch KL decreases from 0.1951 to 0.0715. Failed-branch success peaks at 0.4709 and ends at 0.3652, showing that success can fluctuate even while the preservation loss decreases; the clipping fraction and gradient norm remain finite throughout this trace.

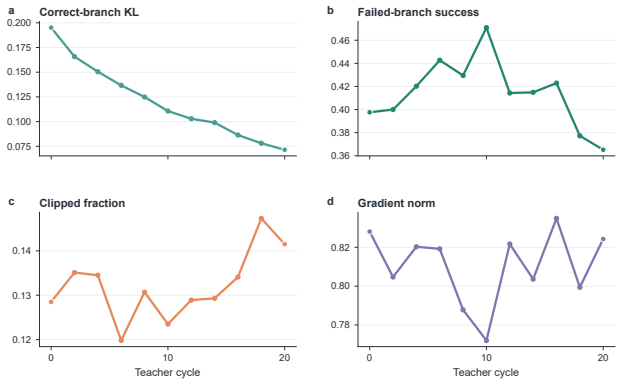

Figure 8: Teacher-training dynamics in E04. Correct-branch KL decreases, while failed-branch success and optimization diagnostics vary across cycles.

E.3 Distribution Objectives

The full-vocabulary reference and generalized JSD give similar quality from one seed, whereas sampled-token KL is cheaper but lower in this run. The regularization coefficient on failed paths controls a visible trade-off between success and compatibility. Removing it raises teacher success but also produces the largest compatibility KL. These comparisons support the full-vocabulary objective as a strong operating point for quality and compatibility.

Table 8: One-seed objective and regularization ablations. Compatibility is measured by KL to the frozen student.

Objective	Macro	Teacher success	Compat. KL	GB	Tokens/s
Full-vocabulary reference	47.562	0.7323	0.0961	61.741	3311
Sampled-token KL	45.864	0.7294	0.1220	43.804	4636
Generalized JSD	**47.623**	0.7169	0.0877	61.949	3175
KL + CE	44.815	0.7447	0.1326	60.909	3373
Top-k approximation	43.519	0.7223	0.1140	48.069	4199
$\beta_{\mathrm{KL}} = 0$	43.395	**0.7855**	0.1722	61.390	3331
$\beta_{\mathrm{KL}} = 0.01$	44.259	0.7480	0.1346	61.781	3293
$\beta_{\mathrm{KL}} = 0.1$	43.735	0.6984	0.0744	62.005	3265
Failed-path clipping	43.981	0.6823	0.0828	61.637	3281
Correct clipping off	45.000	0.7517	0.1526	61.857	3359

E.4 Adaptive Sampling

Adaptive sampling reduces teacher generation while reaching a slightly higher observed macro average. It uses 850 teacher samples and 743,366 generated tokens, compared with 1,032 samples and 862,342 tokens for sampling with a fixed size.

The same accounting gives 1,250 versus 1,432 verifier calls and 0.0148 versus 0.0172 wall-clock hours. Relative to fixed-size sampling, the adaptive controller therefore reduces teacher samples by

Table 9: Adaptive versus fixed-size teacher sampling.

Sampler	Macro	Samples	Tokens	GPU-hours
Adaptive	**47.284**	**850**	**743,366**	**0.0392**
Fixed-8	46.728	1,032	862,342	0.0457

17.6%, verifier calls by 12.7%, generated tokens by 13.8%, GPU-hours by 14.2%, and wall-clock time by 13.9%. In the exported 100-cycle E10 trace, the controller reaches $G_{\mathrm{base}} = 3$ and records no final skipped groups.

E.5 Privilege and Optimization Diagnostics

Table 10 complements Figure 4 with compatibility and format diagnostics. Matching trajectories produce the highest teacher success and the lowest student perplexity. Format validity remains high in all conditions, so the shuffled-trajectory degradation is not explained by malformed outputs alone.

Table 10: Teacher behavior under trajectory perturbations. Successful-path KL and student perplexity measure compatibility with the frozen student.

Condition	Success	Success-path KL	Student PPL	Format pass	JSD
Matching trajectory	**0.7333**	0.4243	**5.8081**	0.9806	**0.0199**
Trajectory removed	0.6694	**0.3645**	6.3699	**0.9833**	0.0864
Cross-problem shuffle	0.6444	0.5047	7.0097	0.9611	0.1303

Figure 9 reports the corresponding cycle-level optimization trace. Correct-branch KL decreases from 0.1862 to 0.0775, and failed-branch success rises from 0.4102 to 0.4836. The mean skipped-group rate is 0.0692 and its maximum is 0.0892, quantifying how often this trace lacks an active failed-branch update.

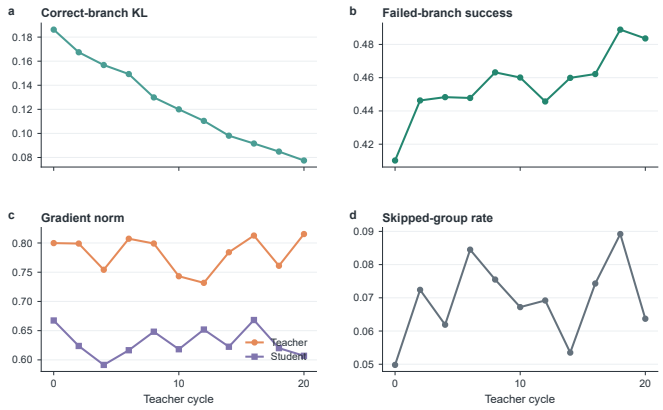

Figure 9: Cycle-level correct-branch KL, failed-branch success, gradient norms, and skipped-group rate.

E.6 One-Seed 4B Extension

PAST has the highest macro point estimate in the Qwen3-4B extension evaluated with one seed. It improves AIME25 and HMMT25 over Vanilla OPSD, while Vanilla OPSD remains stronger on AIME24 (Table 11). This variation across tasks motivates a broader scale study with multiple seeds.

Table 11: One-seed Qwen3-4B extension.

Method	AIME24	AIME25	HMMT25	Macro
Vanilla OPSD	**66.111**	45.000	31.944	47.685
Trajectory input only	64.444	42.222	**34.444**	47.037
PAST	62.778	**48.611**	34.167	**48.519**

F Cost and Reproducibility

F.1 Training and Deployment Cost

Table 12 reports training cost through generated teacher work, verifier calls, runtime, and peak memory, together with deployed parameter count.

Table 12: Reported training-account measurements and deployment size. Teacher samples are additional continuations generated specifically for teacher adaptation.

Method	T. samples	Verifier	Tokens	GPU-h	Wall-h	Peak GB	Deploy B
Vanilla OPSD	0	400	220,476	0.0090	0.0034	53.837	1.7
SD-Zero	0	13,456	7,548,980	0.3667	0.1400	60.701	1.7
H2SD	0	3,200	1,914,827	0.0971	0.0371	61.734	1.7
PAST	850	1,250	743,366	0.0392	0.0148	62.971	1.7

Within PAST, adaptive sampling reduces both samples and tokens relative to fixed-8 sampling, as reported in Table 9. Full-vocabulary KL increases peak memory relative to sampled-token approximations; sequence chunking provides a lower-memory implementation while preserving close numerical agreement with the reference (Table 5).

F.2 Reproducibility Boundary

Each complete cycle checkpoint stores the student and teacher LoRA adapters; their optimizer, scheduler, and mixed-precision states; the next cycle identifier; logical-step counters; adaptive-sampling state; data cursor; and Python, PyTorch, CUDA, and rollout-sampler random-number-generator states. It also records the base checkpoint revision, tokenizer, chat template, verifier, and full training configuration. Resume begins at the next rollout phase. Partial-cycle rollouts, logits, gradients, and verifier buffers are discarded rather than promoted to a completed state.

Autoregressive student and teacher sampling use the corresponding named LoRA adapter in the rollout backend. Full-vocabulary distributions and all backward passes are computed in the training backend. Adapter version and cycle identifiers are checked when updated weights are synchronized. The CPU test backend implements the same phase and routing semantics without running large-model inference.

Limitations

The current evaluation centers on Qwen3-1.7B across three competition-level mathematical reasoning benchmarks. The one-seed 4B extension provides an initial view beyond this setting, and a broader study across model families and scales is a natural next step. PAST currently uses final-answer verification, retaining the simple verifier interface shared by RLVR methods. Reliable process feedback could be incorporated into the same framework for adapting the teacher when available. Our analysis characterizes the population target, while the appendix reports skip rates and group composition for the finite sample implementation. These diagnostics provide a concrete basis for exploring estimators that further trade teacher sampling cost against coverage of successful paths.