ACT: SELECTIVE TEMPORAL FEEDBACK FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense token-level supervision on student-generated trajectories, yet dense supervision can still be poorly allocated. Two positions with similar local teacher–student disagreement may have different trajectory contexts: the disagreement may quickly dissipate at one position but persist into the future at another. Exploiting this temporal context requires distinguishing *where* to intervene from *what* feedback to apply: position selection should preserve disagreement before sign cancellation, whereas feedback construction should retain its direction and net effect. Based on this distinction, we introduce Advantage Credit Transfer (ACT), which decouples position selection from feedback construction. Over a short future window, ACT uses unsigned future mismatch to select a sparse set of positions and mixes signed future feedback into the shaped directional coefficients at those positions, leaving all other shaped directional coefficients unchanged. ACT reuses the teacher and student log probabilities already computed by OPD, requiring no additional teacher queries, trajectory sampling, or value network. Controlled experiments show that future-mismatch selection outperforms alternative selectors at equal coverage, signed future feedback outperforms unsigned replacements under a fixed gate, and selective correction outperforms uniform allocation under matched total correction magnitude. With a 1.5B student, ACT achieves the highest macro average among the evaluated methods across five benchmarks spanning mathematical, coding, and scientific reasoning, under both math-domain and multi-domain distillation.

1 INTRODUCTION

On-policy distillation (OPD) trains on trajectories generated by the student, with a stronger teacher scoring the prefixes actually visited by the student (Agarwal et al., 2024; Gu et al., 2024; Lu, 2025). This paradigm mitigates exposure bias in offline sequence distillation (Kim & Rush, 2016) and has become a practical route for transferring reasoning capabilities to smaller models (Guo et al., 2025; He et al., 2025). Each sampled token receives a raw sampled log-ratio coefficient $A_t = \log p(y_t \mid s_t) - \log q_\theta(y_t \mid s_t)$: positive values reinforce the current action, whereas negative values suppress it. Dense supervision, however, does not imply well-allocated supervision. The same raw coefficient may correspond to very different trajectory neighborhoods: one position may be followed by persistent teacher–student disagreement, while another may be followed by a rapid return to agreement. A purely local update treats these cases identically, whereas the subsequent trajectory may help determine where the limited correction budget should be concentrated.

Prior work improves OPD along two directions. Methods such as EOPD, TrOPD, TIDE, and REOPOLD focus on the reliability of uncertain positions or extreme local feedback (Jin et al., 2026; Xing et al., 2026; Yu et al., 2026; Ko et al., 2026; Oh et al., 2026; Yang et al., 2026); methods such as MiniLLM, TOPD, DASH, PCSD, and SOPD exploit future returns, near-future divergence, adaptive temporal horizons, persistent support, or step-level long-horizon targets (Gu et al., 2024; Jiang & Ferraro, 2026; Hou et al., 2026; Lv et al., 2026; Sun et al., 2026). These strands expose a design axis that is rarely made explicit: the statistic used to locate high-disagreement positions need not be the one used to update them. Signed temporal aggregation preserves correction direction but may obscure strong opposing signals through positive–negative cancellation; unsigned aggregation preserves disagreement magnitude but cannot determine whether the current action should be reinforced or suppressed. We therefore ask what position selection and feedback construction should retain.

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.

We propose Advantage Credit Transfer (ACT), which separates the two decisions and constructs statistics with distinct roles. Unsigned future mismatch $M = K|D|$ selects *where to correct*, while signed future feedback $F = KD$ determines *what correction to transfer*. Specifically, ACT mixes F into the shaped directional coefficient only at the top-ρ positions under M, leaving the others unchanged: $D' = [I + \beta\,\mathrm{diag}(g)(K - I)]D$. Unlike signed aggregation, M preserves disagreement lost from F through cancellation, thereby identifying trajectory regions with little net feedback but substantial teacher–student disagreement.

To explain why later feedback can act on earlier actions, we decompose each raw sampled log-ratio coefficient into a predictable state mean and an action-sampling residual. The score-function identity shows that the local update is driven by the current action residual, whereas a later state's conditional mean can carry information about an earlier action. We then test whether this distinction is useful. At equal coverage, future-mismatch gating outperforms random, local-magnitude, and teacher-entropy selection; under matched total correction magnitude, selective allocation also substantially outperforms uniform allocation. Figure 1 shows ACT's consistent advantage over representative single-domain baselines on all five benchmarks. These results indicate that future disagreement locates positions worth intervening on, while signed future feedback provides a complementary correction direction.

Our contributions are summarized as follows: (1) we introduce ACT, a sparse temporal coefficient update that explicitly separates position selection from feedback construction without additional teacher queries, trajectory sampling, or a value network; (2) we analyze ACT through cancellation information and score-function credit assignment, clarifying what unsigned future mismatch preserves and how future-dependent selection changes the expected update; and (3) we conduct unified evaluation across multiple teachers and training distributions and validate selective temporal feedback through selector, feedback-source, and matched-budget experiments.

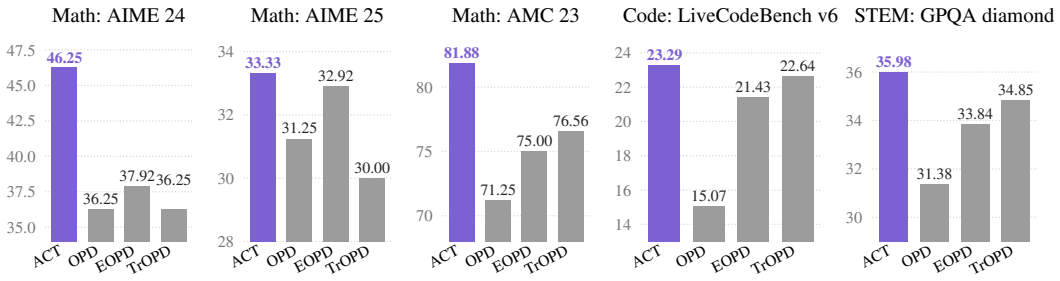

Figure 1: Performance comparison of ACT and representative single-domain baselines. All models are trained from the same DeepSeek-R1-Distill-Qwen-1.5B student with Skywork-OR1-Math-7B as teacher. ACT is highlighted in purple; baselines are shown in gray.

2 Background

2.1 On-policy distillation

Let $x \sim \mathcal{D}$, $y = (y_1, \ldots, y_T) \sim q_\theta(\cdot \mid x)$, and $s_t = (x, y_{<t})$. A fixed teacher p scores the same response at the prefixes actually visited by the student. The sequence-level reverse-KL objective is

$$\mathcal{J}_{\mathrm{RKL}}(\theta) = \mathbb{E}_{x,y\sim q_\theta}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(q_\theta(\cdot \mid s_t) \| p(\cdot \mid s_t))\right]. \tag{1}$$

A commonly used local OPD surrogate is

$$\mathcal{L}_{\mathrm{OPD}} = -\frac{1}{N}\sum_{t} \mathrm{sg}(A_t) \log q_\theta(y_t \mid s_t), \quad A_t = \log p(y_t \mid s_t) - \log q_\theta(y_t \mid s_t). \tag{2}$$

This surrogate treats sampled prefixes as fixed and uses the detached token-level log ratio A_t as its update coefficient, thereby omitting future log-ratio terms from the full sequence-level gradient (Appendix A.1). ACT retains this local surrogate and uses trajectory context to allocate correction.

3 SELECTIVE TEMPORAL FEEDBACK

3.1 LOCAL SHAPING AND TEMPORAL OPERATOR

Define the student-excess log ratio $\delta_t = \log q_\theta(y_t \mid s_t) - \log p(y_t \mid s_t) = -A_t$, and use the one-sided Hellinger-shaped directional coefficient

$$D_t = \psi_\tau(\delta_t) := \begin{cases} -\delta_t, & \delta_t \le \tau, \\ -2(1 - e^{-\delta_t/2}), & \delta_t > \tau. \end{cases} \tag{3}$$

The second branch lies in $[-2, 0]$, preserving the suppressive direction while limiting the influence of extreme student-excess samples. The ablation that uses only this local shaping without temporal feedback is denoted ACT-Directional (Minka et al., 2005; Nowozin et al., 2016; Yu et al., 2026).

Building on this shaped directional coefficient, ACT constructs two statistics from the same discounted future window to answer where temporal correction should be applied and what correction should be transferred.

For horizon H and discount γ, define the normalized forward-looking temporal kernel

$$K_{tj} = \frac{\gamma^{j-t}\, \mathbb{I}[t \le j < \min(t+H, n+1)]}{\sum_{u=t}^{\min(t+H-1,n)} \gamma^{u-t}}, \tag{4}$$

From this kernel we form two statistics: $M_t = \sum_j K_{tj}|D_j|, \quad F_t = \sum_j K_{tj} D_j$. **Unsigned future mismatch** M_t aggregates the magnitude of shaped directional coefficients in the window, measuring total disagreement intensity regardless of sign. **Signed future feedback** F_t retains coefficient directions and gives the net feedback within the window. The former identifies where to intervene, while the latter constructs the temporal correction at that position.

A key informational distinction separates these two scores. Write $P_t = \sum_j K_{tj}[D_j]_+$ and $Q_t = \sum_j K_{tj}[-D_j]_+$ for the weighted positive and negative feedback masses. Then

$$M_t = P_t + Q_t, \qquad F_t = P_t - Q_t, \qquad M_t - |F_t| = 2\min(P_t, Q_t). \tag{5}$$

The gap $M_t - |F_t|$ is the *cancellation mass*: disagreement removed when positive and negative coefficients are summed. When the teacher alternates between supporting and suppressing student actions within a window, strong positive and negative feedback can cancel, making $|F_t|$ small even when M_t remains large. Ranking by M_t rather than $|F_t|$ therefore avoids discarding high-disagreement trajectory regions merely because their net feedback is small. Section 6.3 evaluates this design empirically.

For each response, the future-mismatch selector ranks positions by M_t and induces the binary gate $g_t = \mathbb{I}[M_t$ is in the per-response top-$\rho]$. Per-response ranking prevents long responses from consuming another response's intervention budget.

3.2 SPARSE COEFFICIENT UPDATE

The full update admits a compact operator form. Let $D = (D_1, \ldots, D_n)^\top$ and $g = (g_1, \ldots, g_n)^\top$. Then

$$D' = W_g D, \qquad W_g = I + \beta\, \mathrm{diag}(g)(K - I). \tag{6}$$

For $0 \le \beta \le 1$, each row of W_g is a convex combination of the identity row and the kernel row, making W_g *row-stochastic*. This guarantees $\min_j D_j \le D'_t \le \max_j D_j$: the corrected directional coefficient cannot leave the range of the original shaped directional coefficients. With a fixed gate, the map is also non-expansive in the sup norm, which bounds its sensitivity to perturbations in the input (Appendix B).

Equation 6 also has a direct operational interpretation: the gate g determines which positions receive correction, while the window kernel K determines which coefficients contribute to that correction. They separately control *where to modify* and *what information to transfer*. Thus, ACT is not a simple nonnegative reweighting of the shaped directional coefficient. At a selected position with $D_t = 0$, ACT can still introduce a nonzero correction through $D'_t = \beta F_t$. When $D_t F_t < 0$ and the opposing future evidence is sufficiently strong, it can also reverse the sign of the local correction (Appendix D).

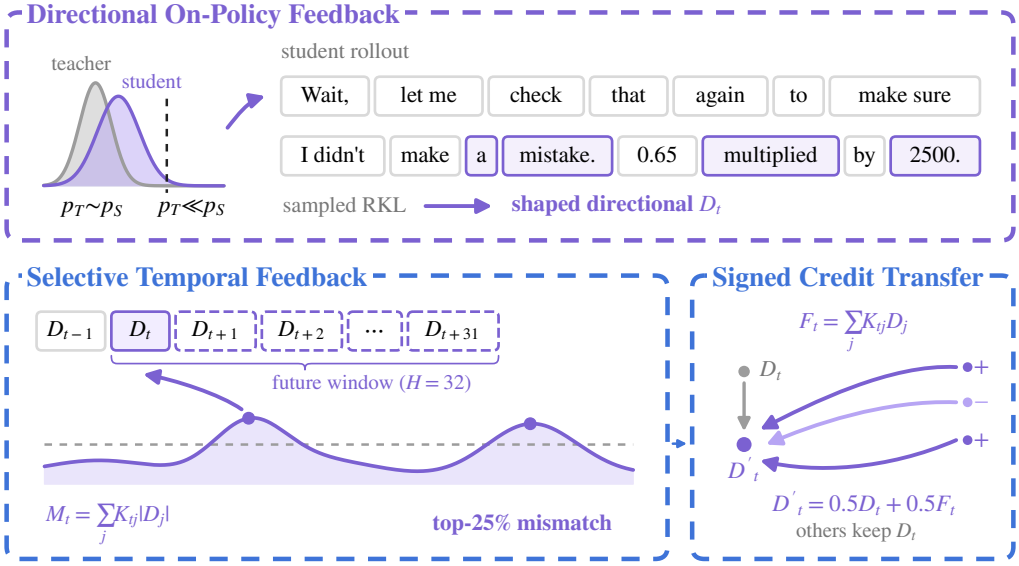

Figure 2: ACT overview. **Top:** the raw sampled log-ratio coefficient A_t is transformed into the shaped directional coefficient D_t to limit extreme suppressive feedback. **Bottom left:** unsigned future mismatch M_t selects the top-25% positions. **Bottom right:** signed future feedback F_t updates only the selected coefficients, while all other positions retain D_t.

3.3 Full Objective

The sampled branch is optimized with $\mathcal{L}_{\text{ACT}} = -\frac{1}{N}\sum_t \text{sg}(D'_t) \log q_\theta(y_t \mid s_t)$. Consistent with EOPD, we additionally compute teacher entropy $H^p_t = -\sum_v p(v \mid s_t) \log p(v \mid s_t)$ and apply top-k forward KL when $H^p_t > \tau_H$:

$$\mathcal{L} = \mathcal{L}_{\text{ACT}} + \alpha \frac{1}{N} \sum_t \mathbb{I}[H^p_t > \tau_H] D^{(k)}_{\text{KL}}(p_t \| q_{\theta,t}). \tag{7}$$

4 What Information Should Travel Backward?

Why can feedback at later positions be used to correct an earlier action? We first use a conditional-mean decomposition to distinguish the information in local and future feedback, and then examine how position selection changes the expected update.

4.1 Local Action Residual and Future State Mean

For the analysis, write the raw sampled log-ratio coefficient as $r_t = \log p(y_t \mid s_t) - \log q_\theta(y_t \mid s_t) = A_t$. Let $\mathcal{F}_t$ denote the history immediately before sampling y_t, and $z_t = \nabla_\theta \log q_\theta(y_t \mid s_t)$. Decompose

$$r_t = \mu_t + \epsilon_t, \qquad \mu_t = \mathbb{E}[r_t \mid \mathcal{F}_t] = -D_{\text{KL}}(q_\theta(\cdot \mid s_t) \| p(\cdot \mid s_t)). \tag{8}$$

Here μ_t is a property of the current state—the negative local KL divergence—while ϵ_t captures which action was sampled.

Proposition 1 (Local action and future state)**.** *Under on-policy sampling and finite second moments,*

$$\mathbb{E}[z_t r_t] = \mathbb{E}[z_t \epsilon_t], \qquad \mathbb{E}[z_t r_j] = \mathbb{E}[z_t \mu_j] \quad (j > t). \tag{9}$$

The first identity follows from the standard score-function baseline property: μ_t is determined by the history before y_t is sampled, so its product with z_t has zero expectation (Sutton et al., 1998). For $j > t$, however, the future state s_j already contains y_t. Its conditional mean μ_j can therefore reflect the consequence of the earlier action, while the residual introduced by sampling y_j vanishes in expectation (Schulman et al., 2015). The proof appears in Appendix A.

4.2 SELECTION CHANGES THE EXPECTED CORRECTION

The ACT gate is computed from the same future trajectory as the feedback. For any $j > t$,

$$\mathbb{E}[z_t g_t r_j] = \mathbb{E}[z_t g_t \mu_j] + \mathbb{E}[z_t \operatorname{Cov}(g_t, r_j \mid \mathcal{F}_j)]\,. \tag{10}$$

The second term captures selection–feedback coupling. Stop-gradient cuts only the differentiation path through g_t; it does not remove the statistical dependence induced by the shared trajectory. Accounting for the local feedback displaced by mixing, the expected change from the local sampled update decomposes into three terms (Appendix A.3):

$$\mathbb{E}[U_{\mathrm{ACT}} - U_{\mathrm{local}}] = S + B - L, \tag{11}$$

where S denotes future *state evidence*, namely the contribution of the conditional mean $\mu_j^D = \mathbb{E}[D_j \mid \mathcal{F}_j]$ of the shaped coefficient at later positions to the earlier update; B denotes *selection coupling*, the covariance between gate weights and shaped feedback; and L is the *local feedback displaced* by the mixture. Products of the conditional standard deviations of gate weights and shaped feedback bound B, providing a diagnostic for when retrospective selection can materially alter the update direction.

Observed distinction between local and future mismatch. Figure 3 compares local magnitude $|D_t|$ with the future-only mismatch M_t^+. The latter excludes the current token to avoid mechanical correlation from the shared term. Across 12,288 measured tokens, the Pearson and Spearman correlations are 0.22 and 0.36. This shows that local magnitude and future context provide different descriptive information, but it implies neither independence nor labels of correctness or causal intervention value. Section 6.3 further tests through selector ablations whether this distinction is useful for training.

Why magnitude and sign play different roles. Figure 3 also visualizes the cancellation identity in Equation 5: opposing coefficients cancel, leaving some windows with large M_t but small $|F_t|$. At 25% coverage, the measured Jaccard overlap between the top-M_t and top-$|F_t|$ sets is 0.55, showing that the rankings are related but not interchangeable.

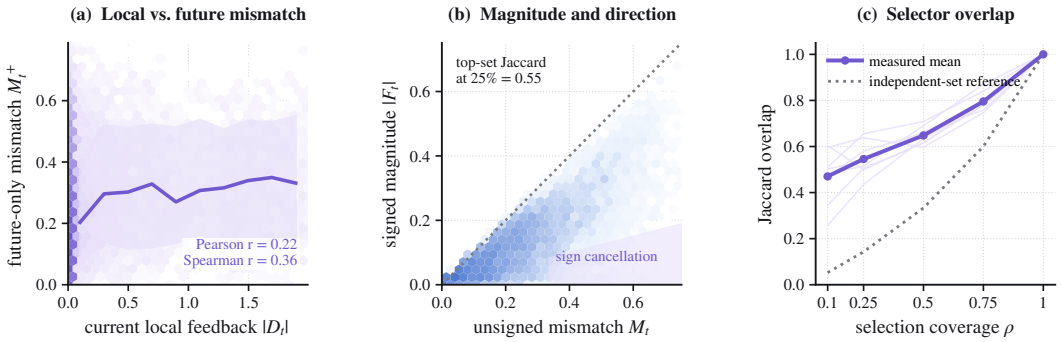

Figure 3: Three complementary views of temporal allocation. Left: local feedback versus future-only mismatch M_t^+. Center: cancellation allows unsigned mismatch M_t to exceed signed magnitude $|F_t|$. Right: Jaccard overlap between equal-coverage top-M_t and top-$|F_t|$ sets.

5 EXPERIMENTAL SETUP

Models and training. Our principal setting distills Skywork-OR1-Math-7B (He et al., 2025) into DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025). We train for 100 steps on the math subset of OpenThoughts3 (Guha et al., 2026) to ensure convergence. Each step uses 128 prompts and samples four student rollouts per prompt; the maximum prompt and response lengths are 2,048 and 8,096 tokens, respectively. The student model is optimized with Adam (Kingma & Ba, 2015) at a learning rate of 5×10^{-6}. All methods share the student initialization, data order, rollout budget, and optimization configuration. The multi-domain setting keeps the student fixed and replaces the teacher with Skywork-OR1-7B. Its training data contain 849,920 math, 243,952 code, and 99,696 science examples; the optimizer, 100-step budget, batch size, rollout count, and response length match the math setting. All experiments run on a cluster of 32 NVIDIA H20 GPUs.

Table 1: Single-domain distillation results (Avg@8). All methods use DeepSeek-R1-Distill-Qwen-1.5B as the student and Skywork-OR1-Math-7B as the teacher. Best and second-best distilled-method results are shown in bold and underline.

Method	AIME 24	AIME 25	AMC 23	LiveCodeBench v6	GPQA diamond	Avg.
Student	28.75	23.33	68.12	19.71	33.65	34.71
Teacher	66.67	53.33	91.88	37.50	47.10	59.30
Single-Domain Distillation						
OPD	36.25	31.25	71.25	15.07	31.38	37.04
EOPD	37.92	<u>32.92</u>	75.00	21.43	33.84	40.22
Entropy OPD 20%	31.67	32.50	74.69	14.71	30.74	36.86
REOPOLD	36.67	31.25	74.69	17.86	33.02	38.70
Outlier-FKL	34.58	30.83	76.25	<u>23.21</u>	34.53	39.88
TrOPD	36.25	30.00	76.56	22.64	34.85	40.06
TIDE	<u>39.58</u>	31.67	78.12	21.36	32.70	40.69
DASH	38.75	28.33	77.19	21.64	34.34	40.05
PCSD	38.33	30.00	76.56	17.21	32.45	38.91
MiniLLM	38.75	29.58	75.31	23.14	34.66	40.29
ACT-Directional	<u>39.58</u>	29.58	<u>79.69</u>	22.14	<u>35.61</u>	<u>41.32</u>
ACT	**46.25**	**33.33**	**81.88**	**23.29**	**35.98**	**44.15**

Baselines. We compare sampled reverse-KL OPD, EOPD (Jin et al., 2026), Entropy OPD, which retains only the highest-entropy 20% of sampled tokens, TrOPD (Xing et al., 2026), reward-clipped REOPOLD (Ko et al., 2026), Outlier-FKL (Xing et al., 2026), TIDE (Yu et al., 2026), PCSD (Lv et al., 2026), and a MiniLLM-style normalized-future ACT control (Gu et al., 2024). Another control is *DASH-style weighting* (Hou et al., 2026): its detached DASH coefficient recursion reweights the ACT-Directional sampled branch while retaining the EOPD branch. ACT-Directional combines Equation 3 with EOPD but disables temporal feedback, thereby isolating the gain from the temporal component.

Evaluation. We evaluate on five benchmarks: AIME24 and AIME25 (Zhang & Math-AI Team, 2024; 2025), the 40-problem Evalchemy subset of AMC23 (Mathematical Association of America, 2023; Raoof et al., 2025), LiveCodeBench v6 (Jain et al., 2025), and GPQA-Diamond (Rein et al., 2024). We sample eight responses per question with temperature 0.6 and top-$p = 0.95$. Math generations use 65,536 tokens; LCB and GPQA use 32,768 tokens. We report **Avg@8**, the mean correctness over eight samples, for each benchmark and their equal-weight five-benchmark macro average, denoted **Macro Avg@8**.

6 Results

6.1 Math-domain distillation

Table 1 shows that ACT attains the highest Macro Avg@8 among the compared distillation methods, outperforming OPD, EOPD, TIDE, DASH, and ACT-Directional by $+7.11$, $+3.93$, $+3.46$, $+4.10$, and $+2.83$ points, respectively. ACT-Directional is the most direct control: the two methods share the shaped directional coefficient, EOPD auxiliary branch, and remaining training settings, differing only in temporal gating and mixing. Relative to the ACT-Directional baseline, ACT improves all five benchmarks: $+6.67$ on AIME24, $+3.75$ on AIME25, $+2.19$ on AMC23, $+1.14$ on LCB v6, and $+0.38$ on GPQA-Diamond. The gains are not confined to competition mathematics. LCB and GPQA also improve over both ACT-Directional and OPD, indicating that the gain is not restricted to a particular answer format.

6.2 Multi-domain distillation

Table 2 evaluates the setting in which both the teacher and the training distribution change. ACT reaches a Macro Avg@8 of 39.74, outperforming OPD, EOPD, and full TrOPD by $+6.06$, $+2.42$,

Table 2: Multi-domain distillation results (Avg@8). All methods use DeepSeek-R1-Distill-Qwen-1.5B as the student and Skywork-OR1-7B as the teacher.

Method	AIME 24	AIME 25	AMC 23	LiveCodeBench v6	GPQA diamond	Avg.
Student	28.75	23.33	68.12	19.71	33.65	34.71
Teacher	64.17	47.92	92.19	41.93	49.12	59.06
Multi-Domain Distillation						
OPD	29.17	24.17	60.31	23.00	31.76	33.68
EOPD	38.33	23.75	67.19	23.93	33.40	37.32
REOPOLD	29.58	29.58	62.81	22.57	33.65	35.64
Outlier-FKL	32.08	26.67	67.81	23.50	33.84	36.78
TrOPD	30.42	27.08	60.00	23.21	34.03	34.95
ACT	**38.75**	**30.42**	**71.25**	**24.21**	**34.09**	**39.74**

and $+4.79$ points, respectively, while ranking first on all five benchmarks. This provides evidence that the benefit of selective temporal feedback is not limited to math-only training: even after adding code and science data with different trajectory structures, the future-mismatch gate continues to identify useful intervention positions.

6.3 Analysis and Ablation Studies

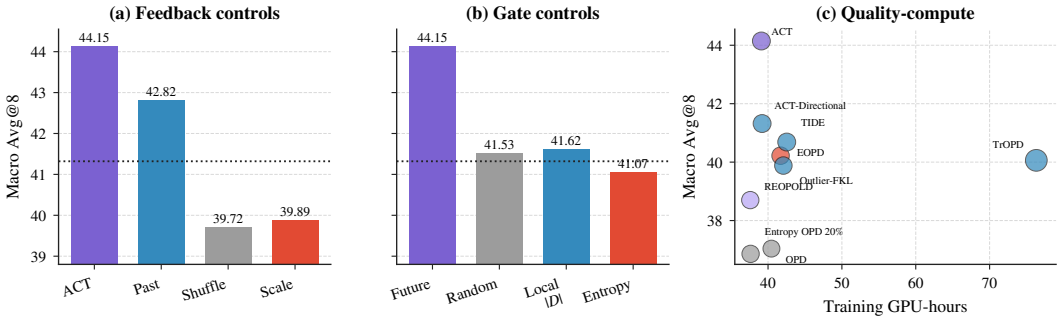

Figure 4: Mechanism controls and quality–compute comparison under a fixed 25% selection budget. Left: future feedback outperforms past, shuffled, and scale-only controls. Center: future-mismatch gating outperforms random, local- magnitude, and teacher-entropy gates. Right: Macro Avg@8 versus measured step-40–100 training GPU-hours; bubble area encodes peak memory.

Future feedback is not equivalent to scaling. Does future feedback help only by shrinking the local update? To answer this question, we fix the future-mismatch gate and change only the feedback assigned to selected positions. *Past* uses a discounted past-window average; *Shuffle* preserves the current-token contribution but permutes future residuals across selected positions; and *Scale* retains only the current-token contribution of the normalized window, removing neighboring feedback. Figure 4(a) shows that replacing future aggregation with past aggregation reduces Macro Avg@8 by 1.33 points. Shuffling or removing neighboring feedback reduces it by 4.42 and 4.26 points, respectively, both below the ACT-Directional baseline. The Scale control is inconsistent with a pure coefficient-shrinkage explanation. The Shuffle control further indicates that performance depends not only on the marginal distribution of future feedback, but also on its alignment with the current position. This agrees with the theory: the future conditional mean μ_j can carry action-consequence information (Proposition 1), whereas permutation breaks the corresponding prefix dependence.

Future mismatch provides a stronger selection criterion. To test the role of the selection criterion under a fixed position budget, we hold future aggregation fixed and vary only the gate; every gate in Figure 4(b) selects exactly 25% of valid response positions. The future-mismatch gate reaches 44.15, whereas random, local-$|D_t|$, and teacher-entropy selection reach 41.53, 41.62, and 41.07, respectively. The latter three remain close to ACT-Directional, while future selection yields a further 2.83-point

gain. Thus the trajectory information used by future mismatch to locate intervention positions is not fully captured by local magnitude or uncertainty.

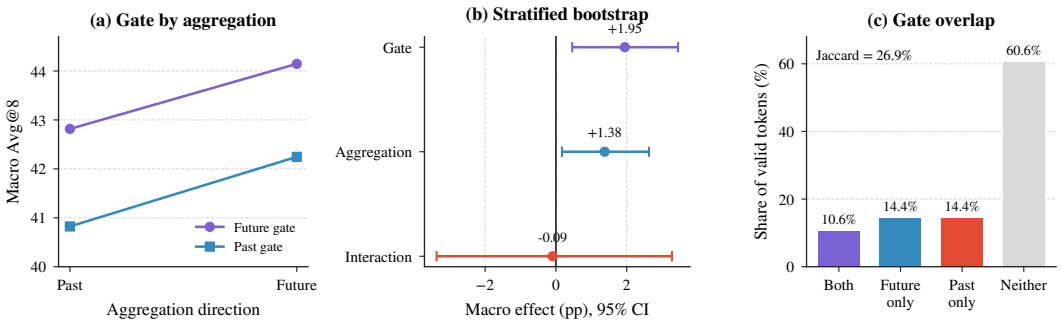

Figure 5: Gate source, aggregation direction, and selector overlap. Left: future gating improves both aggregation choices. Center: stratified question bootstrap gives positive gate and aggregation main effects with near-zero interaction. Right: future/past gate allocation; Jaccard overlap is 26.9%.

Gate source and aggregation direction. The preceding experiments vary either the gate or the feedback, but do not fully cross their temporal sources. We therefore use a 2×2 factorial design (Figure 5): the gate uses either future or past information, and aggregation likewise uses either a future or past window. Future gating improves the two aggregation columns by 1.90 and 1.99 points; future aggregation improves the two gate rows by 1.33 and 1.42 points. The interaction point estimate is near zero, although its confidence interval is wide. The best combination uses both future gating and future aggregation, and the gate main effect has the larger point estimate. These results support retaining different information for the two roles: disagreement magnitude for position selection and signed net evidence for feedback construction. An ablation on feedback sign further validates the latter role: with the future-mismatch gate fixed, replacing signed F_t with unsigned M_t or $|F_t|$ reduces the five-benchmark Macro Avg@8 from 44.15 to 40.97 and 40.12, respectively (Appendix F.1).

Selective allocation under a matched budget. Could the future gate simply increase the total correction magnitude? To distinguish correction amount from position allocation, we construct a Uniform control. For each sequence, define the unscaled temporal correction $\Delta_t = \beta(F_t - D_t)$ and choose the detached coefficient $c = \frac{\sum_t g_t^F |\Delta_t|}{\sum_t |\Delta_t| + \varepsilon}$. This control applies $c\Delta_t$ at every valid position, matching the selective method's aggregate L_1 correction while distributing the budget across the entire response. Uniform reaches 40.34, while ACT reaches 44.15. The key difference is positional coverage: ACT modifies 25% of tokens, whereas Uniform modifies every valid token. Concentrating correction at future-mismatch positions is therefore more effective than uniformly distributing the same budget.

Gate overlap. We further quantify the difference between the positions selected by the two gates. Figure 5(c) shows that future and past gates each cover 25% of valid positions, but only 42.4% of future-gated positions are also selected by the past gate, giving a Jaccard overlap of 26.9%. The gates share a core set, yet allocate nearly three fifths of their budgets to different positions. The positions unique to the future gate yield the stronger correction, showing that the two gates differ substantially in position allocation.

6.4 Efficiency and Inference Budget

ACT applies a normalized temporal operator to already-computed token coefficients and therefore requires no additional teacher evaluation or rollout. Figure 4(c) shows that, over the common step-40–100 interval, its measured training cost remains close to ACT-Directional and EOPD; TrOPD is more expensive because it additionally generates teacher prefixes. At inference time, we impose common token caps to control for differences in generation budget. Figure 6 shows that ACT retains its advantage across caps. On AIME24, it reaches 42.50 Avg@8 at 16k, compared with 34.17 for EOPD and 31.67 for OPD; on AMC23, it reaches 78.44 at 16k and 81.88 at 64k. Comparisons under matched caps show that ACT's overall gain does not arise from a larger inference budget.

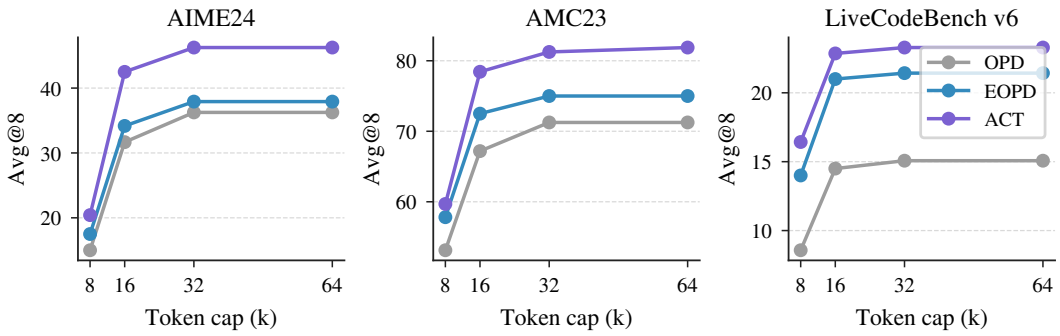

Figure 6: Prefix evaluation under common token caps.

7 Related Work

Knowledge distillation. Knowledge distillation (Hinton et al., 2015) transfers a teacher model's knowledge to a smaller student through soft targets (Sanh et al., 2019; Jiao et al., 2020). For language models, sequence-level distillation (Kim & Rush, 2016) trains on teacher-generated outputs, but the resulting exposure bias degrades long-horizon generation. On-policy distillation addresses this problem by training on student-generated trajectories, with GKD (Agarwal et al., 2024) and MiniLLM (Gu et al., 2024) as foundational approaches. Recent work further investigates the success conditions, reliability, and long-horizon scaling challenges of this paradigm (Li et al., 2026; Song & Zheng, 2026; Oh et al., 2026; Yang et al., 2026).

Reliable local correction. The per-token coefficient in sampled OPD is an unbounded one-sample estimate, and its reliability degrades when teacher and student distributions diverge (Oh et al., 2026; Yang et al., 2026). EOPD combines reverse KL with entropy-triggered forward KL to stabilize high-uncertainty positions (Jin et al., 2026). TrOPD partitions tokens by supervision reliability and adds teacher-prefix off-policy guidance (Xing et al., 2026). TIDE separates student-excess from student-deficit mismatch with Hellinger-shaped coefficients (Yu et al., 2026). REOPOLD clips unreliable rewards (Ko et al., 2026).

Temporal and trajectory-aware supervision. Several methods use information beyond the current token. MiniLLM converts future log ratios into sequence-level training signals (Gu et al., 2024); TOPD identifies reasoning forks through near-future disagreement and optimal transport (Jiang & Ferraro, 2026); DASH adaptively adjusts the supervision horizon through a detached coefficient recursion (Hou et al., 2026); and PCSD aggregates persistent local teacher support with an adaptive decay window (Lv et al., 2026). ACT differs primarily not in whether it uses a future window, but in the role assigned to temporal information, separating *where temporal correction is applied* from *what correction is transferred*.

8 Conclusion

Dense supervision does not guarantee that correction is well allocated. ACT uses a short future window, applying unsigned future mismatch for position selection and signed future feedback for correction construction, thereby explicitly separating the two roles of temporal information. Experiments show that future information contributes to both position selection and feedback aggregation, with effects validated across math, code, and science benchmarks and in both single-domain and multi-domain distillation. Overall, future disagreement both helps the student allocate a limited correction budget and provides effective correction feedback at selected positions.

AI use statement

Generative AI tools were used only to assist with the translation and language polishing of parts of this manuscript. All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the content of this work.

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*, volume 2024, pp. 21246–21263, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. 2026:108059–108130, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

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

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, et al. Dash: Divergence-adaptive supervision horizons for on-policy self-distillation of reasoning models. *arXiv preprint arXiv:2608.06243*, 2026.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Yuxuan Jiang and Francis Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance. *arXiv preprint arXiv:2606.00305*, 2026.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning*, 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.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Kevin Lu. On-policy distillation. Thinking Machines Lab: Connectionism, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

Chunji Lv, Yangguang Wei, Junlin Liu, Yang Gao, Ming Liu, Xinming Wang, Jinyang Wu, Guoren Wang, and Changsheng Li. Pcsd: Persistent consistency for self-distillation in agentic reinforcement learning. *arXiv preprint arXiv:2608.01837*, 2026.

Mathematical Association of America. American mathematics competitions: AMC 2023. Competition problems, 2023.

Tom Minka et al. Divergence measures and message passing. 2005.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Minjae Oh, Gyubin Choi, Sangjun Song, Yunho Choi, and Yohan Jo. KL for a KL: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Negin Raoof, Etash Kumar Guha, Ryan Marten, Jean Mercat, Eric Frankel, Sedrick Keh, Hritik Bansal, Georgios Smyrnis, Marianna Nezhurina, Trung Vu, Zayne Rea Sprague, Mike A Merrill, Liangyu Chen, Caroline Choi, Zaid Khan, Sachin Grover, Benjamin Feuer, Ashima Suvarna, Shiye Su, Wanjia Zhao, Kartik Sharma, Charlie Cheng-Jie Ji, Kushal Arora, Jeffrey Li, Aaron Gokaslan, Sarah M Pratt, Niklas Muennighoff, Jon Saad-Falcon, John Yang, Asad Aali, Shreyas Pimpalgaonkar, Alon Albalak, Achal Dave, Hadi Pouransari, Greg Durrett, Sewoong Oh, Tatsunori Hashimoto, Vaishaal Shankar, Yejin Choi, Mohit Bansal, Chinmay Hegde, Reinhard Heckel, Jenia Jitsev, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. Evalchemy, June 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *Conference on Language Modeling*, 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

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

Changhui Sun, Lanbo Liu, Hang Lei, Tong Ling, Jiahang Xie, Zhiyong Zheng, Yujia Wang, Hao Liu, Feng Xiao, Lu Liu, et al. Step-level on-policy distillation: Interpolating between on-policy distillation and supervised fine-tuning. *arXiv preprint arXiv:2608.16333*, 2026.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-opd: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 2026.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement. *arXiv preprint arXiv:2608.09836*, 2026.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2024, 2024.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025, 2025.

A THEORETICAL ANALYSIS OF ACT

This section develops the formal analysis of ACT's update structure. We separate three questions: which action receives future feedback (Section A.1); what information the unsigned selector preserves (Section A.2); and how same-trajectory selection changes the expected correction (Section A.3). Section B bounds the sensitivity of the resulting update.

A.1 SCORE-FUNCTION IDENTITIES AND FUTURE STATE INFORMATION

Lemma 1 (Score-function decomposition)**.** *With on-policy sampling and* $r_t = \log p(y_t \mid s_t) - \log q_\theta(y_t \mid s_t)$*, define* $\mu_t = \mathbb{E}[r_t \mid \mathcal{F}_t]$ *and* $\epsilon_t = r_t - \mu_t$*. Then*

$$\mu_t = -D_{\mathrm{KL}}(q_\theta(\cdot \mid s_t) \| p(\cdot \mid s_t)), \qquad \mathbb{E}[z_t r_t] = \mathbb{E}[z_t \epsilon_t],$$

and for $j > t$*,* $\mathbb{E}[z_t r_j] = \mathbb{E}[z_t \mu_j]$*. The full finite-horizon sequence divergence* J *satisfies*

$$-\nabla_\theta J = \mathbb{E} \sum_t z_t \sum_{j \geq t} r_j. \tag{12}$$

Proof. The normalized student distribution gives $\mathbb{E}[z_t \mid \mathcal{F}_t] = 0$; its conditional expectation of the log ratio gives μ_t. A predictable μ_t therefore vanishes against z_t. For $j > t$, z_t is measurable at time j, and $\mathbb{E}[\epsilon_j \mid \mathcal{F}_j] = 0$, proving the future identity. Finally, with $R = \sum_j r_j$,

$$\nabla_\theta \mathbb{E}[R] = \mathbb{E}\left[R \sum_t z_t - \sum_t z_t \right].$$

The last score sum has zero mean. Terms $z_t r_j$ with $j < t$ vanish by the same baseline argument, yielding the sequence-gradient formula. □

The first identity is the familiar state-baseline property. The second gives the temporal interpretation used by ACT: the future state mean μ_j can depend on the earlier action because that action is part of the future prefix. ACT's selected normalized window is a designed surrogate for these future state terms, not the full reward-to-go estimator.

Selection–feedback coupling. Let g_t be any detached gate that may depend on the sampled future. For $j > t$, substitute $r_j = \mu_j + \epsilon_j$:

$$\mathbb{E}[z_t g_t r_j] = \mathbb{E}[z_t g_t \mu_j] + \mathbb{E}[z_t \operatorname{Cov}(g_t, r_j \mid \mathcal{F}_j)].$$

The second term is a selection–feedback coupling. Stop-gradient removes a differentiation path through g_t, but it does not remove this statistical dependence. By conditional Cauchy–Schwarz, the coupling is bounded by

$$\|\mathbb{E}[z_t g_t \epsilon_j]\| \leq \mathbb{E}\left[\|z_t\| \sqrt{\operatorname{Var}(g_t \mid \mathcal{F}_j) \operatorname{Var}(r_j \mid \mathcal{F}_j)} \right].$$

A.2 WHAT UNSIGNED SELECTION RETAINS: CANCELLATION THEORY

The choice between unsigned mismatch $M_t = (K|D|)_t$, net magnitude $|F_t| = |(KD)_t|$, and signed support $F_t = (KD)_t$ as gate scores is central to ACT's design. The following result characterizes exactly what these three rankings preserve and discard.

For one row of K, let

$$P_t = \sum_j K_{tj}[D_j]_+, \qquad Q_t = \sum_j K_{tj}[-D_j]_+.$$

These are the weighted positive and negative feedback masses within the window.

Proposition 2 (Mismatch intensity, net support, and cancellation). *With $M_t = (K|D|)_t$ and $F_t = (KD)_t$,*

$$M_t = P_t + Q_t, \qquad F_t = P_t - Q_t, \qquad M_t - |F_t| = 2\min(P_t, Q_t). \tag{13}$$

Thus $M_t \geq |F_t|$, with equality precisely when at least one of P_t, Q_t vanishes (i.e., the window is uniformly signed). If the window kernel contains at least two positive weights, neither F_t nor $|F_t|$ alone determines M_t in general.

Proof. Use $D_j = [D_j]_+ - [-D_j]_+$ and $|D_j| = [D_j]_+ + [-D_j]_+$ with kernel linearity to obtain the first two identities. For $P, Q \geq 0$, $P + Q - |P - Q| = 2\min(P, Q)$, proving the cancellation gap. For non-identifiability, take two positive kernel weights $u, v > 0$ and set their coefficients to a/u and $-a/v$ with all others zero. Then $F_t = 0$ but $M_t = 2a > 0$, whereas the all-zero window has both statistics zero. Positive a can be chosen small enough to respect the implemented coefficient bound. □

The cancellation gap $2\min(P_t, Q_t)$ is the total feedback mass in the window that vanishes under signed aggregation. ACT ranks positions by total mismatch *before* cancellation; a signed-support ranking (as used in PCSD (Lv et al., 2026)) or a net-magnitude ranking discards exactly this quantity. The counterexample in the proof shows that a window with $F_t = 0$ can have arbitrarily large M_t: the two selectors are informationally non-equivalent, not merely numerically different.

Corollary 1 (Sign symmetry of the temporal map). *For a fixed kernel and consistent tie-breaking, replacing D by $-D$ leaves both $K|D|$ and $|KD|$ unchanged while reversing KD. The ACT temporal map therefore satisfies*

$$g(-D) = g(D), \qquad W_{g(-D)}(-D) = -W_{g(D)}D.$$

This sign-equivariance means ACT does not preferentially select one sign of shaped feedback: the gate responds to disagreement *intensity*, not direction. A signed-support ranking generally changes under this reversal. The property applies to the temporal map on D; the inherited one-sided shaping ψ_τ is not an odd function of the raw log ratio.

A.3 Selection-induced drift and directional alignment

The temporal correction includes both added future feedback and the local contribution displaced by mixing. For $j > t$, define

$$c_{tj} = \frac{\beta g_t K_{tj}}{N}, \qquad a_t = \frac{\beta g_t (1 - K_{tt})}{N}, \qquad \mu_j^D = \mathbb{E}[D_j \mid \mathcal{F}_j].$$

Let $\bar{c}_{tj} = \mathbb{E}[c_{tj} \mid \mathcal{F}_j]$.

Theorem 1 (State, selection, and local-retention decomposition). *The expected change from the local sampled update is*

$$\mathbb{E}[U_{\mathrm{ACT}} - U_{\mathrm{local}}] = S + B - L, \tag{14}$$

where

$$\begin{aligned} S &= \sum_{t<j} \mathbb{E}[z_t \bar{c}_{tj} \mu_j^D], \\ B &= \sum_{t<j} \mathbb{E}[z_t \operatorname{Cov}(c_{tj}, D_j \mid \mathcal{F}_j)], \\ L &= \sum_t \mathbb{E}[z_t a_t D_t]. \end{aligned} \tag{15}$$

Furthermore, $\|B\| \leq R$, where

$$R = \sum_{t<j} \mathbb{E}\left[\|z_t\| \sqrt{\operatorname{Var}(c_{tj} \mid \mathcal{F}_j) \operatorname{Var}(D_j \mid \mathcal{F}_j)} \right].$$

Proof. Because the future kernel is upper triangular and row-normalized,

$$U_{\mathrm{ACT}} - U_{\mathrm{local}} = \sum_{t<j} z_t c_{tj} D_j - \sum_t z_t a_t D_t.$$

For $j > t$, z_t is $\mathcal{F}_j$-measurable. Consequently,

$$\mathbb{E}[z_t c_{tj} D_j] = \mathbb{E}[z_t\{\bar{c}_{tj}\mu_j^D + \mathrm{Cov}(c_{tj}, D_j \mid \mathcal{F}_j)\}].$$

Summing proves the decomposition. Conditional Cauchy–Schwarz bounds each covariance by the product of its conditional standard deviations; the triangle inequality gives the bound on R. □

The three terms have distinct interpretations. S captures future *state evidence*: the contribution of the conditional mean μ_j^D of the shaped coefficient at later positions to the earlier-action update. B captures *selection coupling*: the covariance between the gate weight c_{tj} and shaped feedback D_j, arising because both depend on the same trajectory. L captures the *local feedback displaced* by mixing. The last term matters: ACT is not an unrestricted reward-to-go addition.

Corollary 2 (Directional alignment condition)**.** *Fix a unit comparison direction v (e.g., a normalized negative gradient of a separately specified evaluation loss at the current parameters). The incremental directional alignment is*

$$v^\top \mathbb{E}[U_{\mathrm{ACT}} - U_{\mathrm{local}}] = v^\top (S - L) + v^\top B.$$

Since $|v^\top B| \le R$,

$$v^\top (S - L) > R \implies v^\top \mathbb{E}[U_{\mathrm{ACT}} - U_{\mathrm{local}}] > 0. \tag{16}$$

Likewise $v^\top (S - L) < -R$ guarantees negative incremental alignment.

This sufficient condition connects the selection coupling to a declared direction of improvement rather than assigning it an inherently positive role. When the future state signal S minus the displaced local term L dominates the covariance bound, the temporal correction is guaranteed to improve alignment regardless of the coupling's sign.

A.4 Conditional-mean feedback and Rao–Blackwellization

Lemma 2 (Single-lag projection)**.** *For $j > t$ let $X = z_t c_{tj} D_j$ have finite second moment, and let $X_{\mathrm{RB}} = \mathbb{E}[X \mid \mathcal{F}_j] = z_t\{\bar{c}_{tj}\mu_j^D + \mathrm{Cov}(c_{tj}, D_j \mid \mathcal{F}_j)\}$. Then X_{RB} preserves the mean and obeys*

$$\mathrm{Cov}(X) - \mathrm{Cov}(X_{\mathrm{RB}}) = \mathbb{E}[\mathrm{Cov}(X \mid \mathcal{F}_j)] \succeq 0.$$

If c_{tj} is $\mathcal{F}_j$-measurable, this reduces to $X_{\mathrm{RB}} = z_t c_{tj} \mu_j^D$, with

$$\mathbb{E}\|X\|^2 - \mathbb{E}\|X_{\mathrm{RB}}\|^2 = \mathbb{E}\big[c_{tj}^2 \|z_t\|^2 \,\mathrm{Var}(D_j \mid \mathcal{F}_j)\big]. \tag{17}$$

Proof. Conditional expectation gives the projected variable and preserves its mean. Decompose $X = X_{\mathrm{RB}} + (X - X_{\mathrm{RB}})$. The conditional residual is orthogonal in second moment to $\mathcal{F}_j$-measurable quantities. The law of total covariance proves the matrix inequality. In the predictable case, take its trace and factor out $z_t c_{tj}$. □

This is a standard projection guarantee specialized to ACT's setting. Replacing the sampled D_j by its conditional mean μ_j^D under the retrospective gate is not generally this projection, because replacing only D_j by its conditional mean does not remove the same-trajectory selection dependence in c_{tj}. Single-lag reductions also do not imply a smaller variance for the entire temporal sum, whose cross-lag covariances can change. The result motivates future work on conditional-mean feedback while the method evaluated here uses the direct sampled window.

A.5 HYBRID SHAPING AND ITS CONDITIONAL MEAN

For the all-action Hellinger shaping function $h(r) = 2(\exp(r/2) - 1)$, with $r = \log p_v - \log q_v$,

$$\mathbb{E}_{v\sim q}[h(r)] = 2\left(\sum_v \sqrt{p_v q_v} - 1\right) = -2H^2(p, q) \in [-2, 0],$$

where $H^2(p, q) = 1 - \sum_v \sqrt{p_v q_v}$ is the squared Hellinger distance.

Lemma 3 (Hybrid conditional mean)**.** *Let $\delta(v) = -r(v)$. For ACT's hybrid coefficient $D(v) = \psi_\tau(\delta(v)) = \psi_\tau(-r(v))$ with threshold τ,*

$$-D_{\mathrm{KL}}(q\|p) \le \mu^D := \mathbb{E}_q[D] \le -2H^2(p, q) \le 0.$$

More explicitly,

$$\mu^D = -D_{\mathrm{KL}}(q\|p) + \mathbb{E}_q\left[\mathbb{I}[r < -\tau]\{2(e^{r/2} - 1) - r\}\right].$$

Proof. Convexity gives $r \le h(r) = 2(e^{r/2} - 1)$, and the hybrid obeys $r \le \psi_\tau(-r) \le h(r)$. Since $\mathbb{E}_q[h(r)] = -2H^2(p, q)$, expectations give the sandwich bound. Adding and subtracting r on the shaped branch yields the explicit correction term. □

The hybrid's conditional mean lies between the negative reverse KL (which is unbounded) and the negative squared Hellinger distance (which is bounded in $[-2, 0]$). The correction term is nonnegative and measures how much the shaping raises severe-excess coefficients toward the Hellinger bound. This characterizes what μ_j^D in Theorem 1 actually summarizes at each future state.

A.6 EFFECTIVE WINDOW GEOMETRY

For $h = \min(H, n - t + 1)$, define the kernel concentration width and the coefficient retained on the current token:

$$h_{\mathrm{eff}} = \frac{\left(\sum_{j=0}^{h-1} \gamma^j\right)^2}{\sum_{j=0}^{h-1} \gamma^{2j}}, \qquad \omega_{\mathrm{self}} = 1 - \beta + \frac{\beta}{\sum_{j=0}^{h-1} \gamma^j}.$$

The remaining coefficient share is $1 - \omega_{\mathrm{self}}$. At $H = 32, \gamma = 0.95, \beta = 0.5$, the interior kernel has $h_{\mathrm{eff}} \approx 26.34$ and retains about 53.10% on the current coefficient. At the last token, retention becomes 100%.

Figure 7 visualizes both quantities as the number of remaining valid tokens varies.

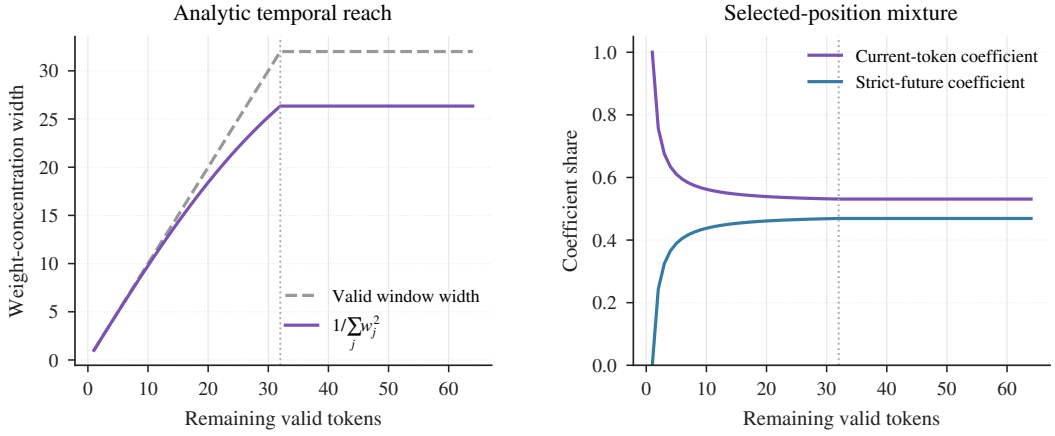

Figure 7: Analytical supervision geometry under $H = 32, \gamma = 0.95, \beta = 0.5$. Left: valid window width and kernel concentration width. Right: the coefficient shares assigned to the current token and strict-future tokens at a selected position. The vertical dotted line marks the full 32-token horizon.

The left panel quantifies the finite reach of temporal information: even with $H = 32$, the discounted kernel has an effective width of about 26 tokens and saturates once 32 valid tokens remain. This is consistent with the sensitivity analysis (Table 7), where extending the horizon to $H = 64$ does not improve performance. The right panel shows that a selected position retains a majority current-token coefficient in the interior and continuously reverts to the unmixed shaped directional coefficient near the response boundary. Thus, ACT introduces finite-range future information without replacing its current-position feedback, and boundary renormalization exactly recovers the local objective at the final position.

B Stability and Sensitivity

Theorem 2 (Range preservation, non-expansiveness, and intervention bound)**.** *For $0 \le \beta \le 1$, W_g is row-stochastic. Therefore*

$$\min_j D_j \le D'_t \le \max_j D_j, \qquad \|W_g D - W_g \widetilde{D}\|_\infty \le \|D - \widetilde{D}\|_\infty$$

for a fixed gate. At most $b = \lceil \rho n \rceil$ coefficients change, and

$$\|D' - D\|_1 \le \beta \sum_{t:g_t=1} \sum_j K_{tj} |D_j - D_t|. \tag{18}$$

For one response, if $|D_t| \le C$, $\|z_t\| \le Z_0$, and $N = n$, then $\|U_{\mathrm{ACT}} - U_{\mathrm{local}}\| \le 2\beta(b/n)CZ_0$.

Proof. Row t of W_g is the convex combination $(1-\beta g_t)e_t^\top + \beta g_t K_{t:}$. This proves range preservation and fixed-gate non-expansiveness in the sup norm. Also $D'_t - D_t = \beta g_t \sum_j K_{tj}(D_j - D_t)$. Taking absolute values proves Equation 18, and multiplying each difference by its score norm proves the last bound. □

In particular, the implemented lower bound -2 from Hellinger shaping survives temporal mixing. Row normalization does not imply column normalization or conservation of $\sum_t D_t$.

Proposition 3 (Top-budget optimality and gate-margin stability)**.** *The top-b mask maximizes $g^\top M$ over $g \in [0,1]^n$, $\mathbf{1}^\top g = b$. For $0 < b < n$, let $\Delta = M_{(b)} - M_{(b+1)}$, with scores sorted decreasingly. On the same response mask and kernel, if $\|D - \widetilde{D}\|_\infty \le \varepsilon$ and $\Delta > 2\varepsilon$, the selected set is unchanged and*

$$\|W_{g(D)} D - W_{g(\widetilde{D})} \widetilde{D}\|_\infty \le \varepsilon.$$

Proof. Moving selection mass from a lower-scoring entry to an unsaturated higher-scoring entry cannot decrease $g^\top M$; repeated exchanges yield top-b. Further, $\|K|D| - K|\widetilde{D}|\|_\infty \le \varepsilon$. The selected-versus-unselected boundary remains strictly ordered when $\Delta > 2\varepsilon$. Apply the fixed-gate bound from Theorem 2. □

The first part shows that ACT's selection is optimal for the specified mismatch statistic—no other binary mask with the same budget captures more total mismatch. The margin certificate provides an empirical stability diagnostic: when the gap Δ between the last selected and first unselected positions is large relative to coefficient perturbations, the gate is robust to small changes in the input.

C ALGORITHM

Algorithm 1: Selective temporal feedback for OPD

1. Sample a batch of student responses $y \sim q_{\theta_{\text{old}}}(\cdot \mid x)$.
2. Evaluate student log probabilities, teacher log probabilities, and teacher top-k distributions on the sampled responses.
3. Compute the shaped directional coefficient D_t using Equation 3.
4. Compute F_t and M_t with the normalized forward-looking temporal kernel (Equation 4).
5. Select the top-ρ positions by M_t independently for each response.
6. Form D'_t using Equation 6.
7. Add entropy-triggered top-k forward KL and update the student with Equation 7.

The window is normalized by the number and discount weights of valid tokens remaining in each response. Near the end of a response, the shorter valid window is therefore renormalized rather than padded with zeros. If only the current token remains, $F_t = D_t$, so no temporal change occurs. The same identity holds everywhere when $H = 1$, recovering the local objective.

D OPERATOR SEPARATION OF SELECTION AND FEEDBACK

The operator form in Equation 6 enables a precise characterization of the two intervention axes. Let Z have row $z_t^\top$. For any fixed gate g and aggregation kernel K,

Proposition 4 (Two distinct intervention axes)**.** *On a common fixed sample batch,*

$$U(g, K) - U(\widetilde{g}, K) = \frac{\beta}{N} Z^\top \operatorname{diag}(g - \widetilde{g})(K - I)D, \tag{19}$$

$$U(g, K) - U(g, \widetilde{K}) = \frac{\beta}{N} Z^\top \operatorname{diag}(g)(K - \widetilde{K})D. \tag{20}$$

The joint difference of differences is

$$\frac{\beta}{N} Z^\top \operatorname{diag}(g - \widetilde{g})(K - \widetilde{K})D.$$

This gives an exact gradient-level counterpart to the factorial design in Section 6.3: the gate changes *where* correction is injected; the aggregation changes *which coefficients form* the correction. Equal gate counts do not make these operations interchangeable.

Beyond nonnegative rescaling. If $D_t = 0$, any multiplicative reweighting $a_t D_t = 0$, whereas ACT gives $D'_t = \beta g_t F_t$. For a selected position with $D_t \neq 0$, a sign reversal occurs when $D_t F_t < 0$ and $\beta > |D_t|/(|D_t| + |F_t|)$. This separates the coefficient map from nonnegative scaling of local feedback.

E SUPERVISION COMPONENTS AND COMPONENT LADDER

Table 3 separates ACT's supervision into three branches:

The completed math-domain component ladder is 37.04 (OPD) $\rightarrow$ 40.22 (EOPD) $\rightarrow$ 41.32 (ACT-Directional) $\rightarrow$ 44.15 (ACT) Macro Avg@8. Each step adds one component: EOPD adds entropy-triggered FKL, ACT-Directional adds Hellinger shaping, and ACT adds mismatch-based position selection and signed temporal mixing.

F COMPLETE MECHANISM CONTROLS

Table 4 shows that future gating with future feedback achieves the highest Macro Avg@8, 44.15. Uniformly distributing the same feedback yields only 40.34, and random, local-magnitude, and

Table 3: ACT supervision branches on a fixed rollout. Auxiliary-statistic sizes exclude model activations, logits, and optimizer state.

Branch	Gradient-defining term	Applied at	Auxiliary size
Local sampled	$-\operatorname{sg}(D_t)\log q_t$	All valid positions	$O(N)$
Temporal change	$-\operatorname{sg}[\beta g_t(F_t - D_t)]\log q_t$	Top-ρ positions	$O(N)$
Teacher support	$\alpha h_t D_{\mathrm{KL}}(p_t^{\mathcal{K}} \| q_t^{\mathcal{K}})$	$h_t = \mathbb{I}[H(p_t) > \tau_H]$	$O(Nk)$

Table 4: Temporal and gate controls under the common Macro Avg@8 protocol. Every method uses the same training budget and benchmarks.

Gate	Feedback	A24	A25	AMC	LCB	GPQA	Avg.
Future	Future	46.25	33.33	81.88	23.29	35.98	44.15
Future	Past	42.08	27.50	83.75	23.50	37.25	42.82
Future	Shuffled future	32.50	30.00	79.69	21.64	34.79	39.72
Future	Scale only	34.17	29.17	77.81	23.07	35.23	39.89
Random	Future	41.67	27.92	80.31	22.71	35.04	41.53
Local $\lvert D\rvert$	Future	41.67	31.25	77.50	21.71	35.98	41.62
Teacher entropy	Future	38.75	32.50	76.25	22.93	34.91	41.07
Past	Future	38.33	32.92	82.19	23.57	34.22	42.25
Past	Past	36.67	30.42	79.69	23.00	34.34	40.82
Uniform	Future	37.08	30.00	77.81	22.07	34.72	40.34

teacher-entropy gates also trail future-mismatch gating. Table 5 further gives positive gate and aggregation main effects, 1.95 (95% interval $[0.46, 3.45]$) and 1.38 ($[0.17, 2.63]$), while the interaction interval crosses zero. This supports treating *where to select* and *what to aggregate* as two independently contributing axes.

F.1 Feedback sign ablation

The main ACT design deliberately assigns different semantics to the two window statistics: $M_t = (K|D|)_t$ determines *where* to intervene, whereas $F_t = (KD)_t$ determines *what signed correction* to apply. We test whether this distinction is necessary by keeping the future-mismatch gate, selection ratio, horizon, discount, and mixing strength fixed, and replacing F_t at selected positions with either M_t or $|F_t|$:

$$D'_t = \begin{cases} (1-\beta)D_t + \beta Z_t, & g_t = 1, \\ D_t, & g_t = 0, \end{cases} \qquad Z_t \in \{F_t, M_t, |F_t|\}.$$

Thus the three temporal variants differ only in the feedback quantity assigned under the same M_t-based gate.

Table 6 shows that both unsigned replacements substantially degrade the five-benchmark Macro Avg@8: M_t loses 3.18 points and $|F_t|$ loses 4.03 points relative to ACT. They also fall below ACT-Directional by 0.35 and 1.20 points, respectively. The degradation is broad rather than being caused by one outlier benchmark: relative to signed ACT, M_t is lower on all five tasks, with the largest drops on AIME24 (-6.67) and AIME25 (-5.42); $|F_t|$ is likewise lower throughout and loses 10.83 points on AIME24.

This result identifies a functional asymmetry between selection and feedback. Both M_t and $|F_t|$ are nonnegative. Using either as the feedback target therefore erases whether the future window calls for increasing or suppressing the sampled action. In particular, when $F_t < 0$, replacing it by $|F_t|$ reverses the temporal direction; when a window contains substantial positive and negative masses, replacing F_t by M_t converts cancellation—evidence of conflicting local signals—into uniformly positive support. With $\beta = 0.5$, either replacement can attenuate or even reverse a negative shaped directional coefficient at a selected position. The unsigned statistics remain useful for ranking intervention sites because sign cancellation should not hide contested windows, but they are poorly matched to the

Table 5: Stratified question-bootstrap contrasts for the gate-by-aggregation factorial.

Contrast	Estimate	95% low	95% high
Gate main effect	+1.95	+0.46	+3.45
Aggregation main effect	+1.38	+0.17	+2.63
Interaction	−0.09	−3.37	+3.28
Future gate: future vs. past aggregation	+1.33	−0.71	+3.49
Past gate: future vs. past aggregation	+1.42	−0.58	+3.42
Future aggregation: future vs. past gate	+1.90	−0.37	+4.21
Past aggregation: future vs. past gate	+1.99	−0.12	+4.16

Table 6: Ablating the sign of the feedback assigned at positions selected by the future-mismatch gate. All temporal variants use $H = 32, \gamma = 0.95, \rho = 0.25, \beta = 0.5$. Δ is the five-benchmark Macro Avg@8 difference from signed ACT.

Feedback at selected positions	A24	A25	AMC	LCB	GPQA	Avg.	Δ
Signed F_t (ACT)	46.25	33.33	81.88	23.29	35.98	44.15	0.00
Unsigned mismatch M_t	39.58	27.92	80.31	22.57	34.47	40.97	−3.18
Unsigned net magnitude $\lvert F_t \rvert$	35.42	28.75	80.00	21.57	34.85	40.12	−4.03
No temporal feedback	39.58	29.58	79.69	22.14	35.61	41.32	−2.83

semantic role of a correction coefficient. The ablation therefore supports the intended division of labor: use unsigned mismatch for *where*, and preserve signed feedback for *what*.

F.2 Per-task views of the temporal controls

Figure 8 augments the aggregate comparison with paired question-level effects and optimization traces. Several patterns reinforce ACT's central insight. First, past aggregation retains most of the gain on AMC23 (+4.06 over ACT-Directional) and LCB v6 (+1.36), indicating that future gating still provides useful position allocation when the aggregation window is changed from future to past. Second, the shuffled-future control falls *below* ACT-Directional on AIME24 (−7.08), showing that arbitrarily permuting future feedback degrades downstream performance; both the aggregate statistic and its alignment with the current position matter. Third, the gradient-norm trajectories during training remain closely matched for all four temporal controls despite their 4.43-point spread in downstream accuracy. This observation does not support attributing the gain to a change in overall update scale and is consistent with a contribution from position allocation.

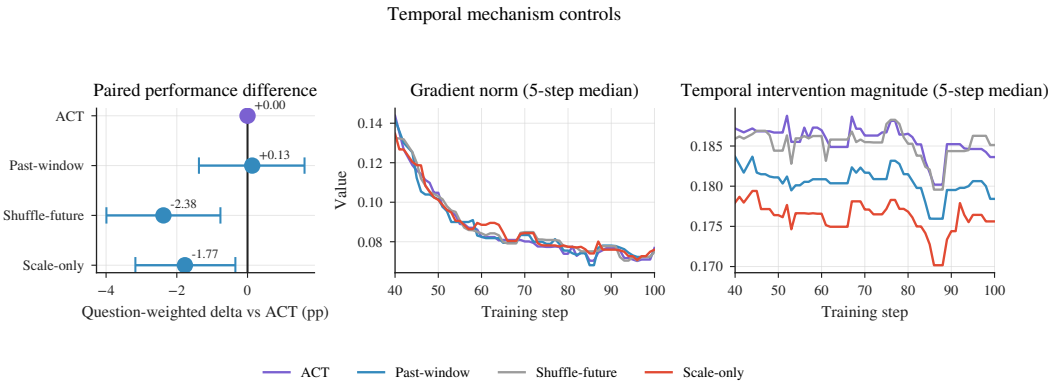

Figure 8: Temporal controls with paired downstream differences, gradient norm, and temporal intervention magnitude. Error bars in the left panel are 95% question-bootstrap intervals.

Figure 9 provides the per-task breakdown for the gate-selector ablation. None of the three alternative selectors matches the future-mismatch gate on every individual task. Teacher-entropy selection is the

closest alternative on AIME25 but drops -7.50 on AIME24 and -5.63 on AMC23. The random gate performs comparably on GPQA and LCB, suggesting that these benchmarks are more tolerant of imprecise selection; the differences are concentrated on mathematical reasoning tasks, indicating that long multi-step trajectories are more sensitive to position selection.

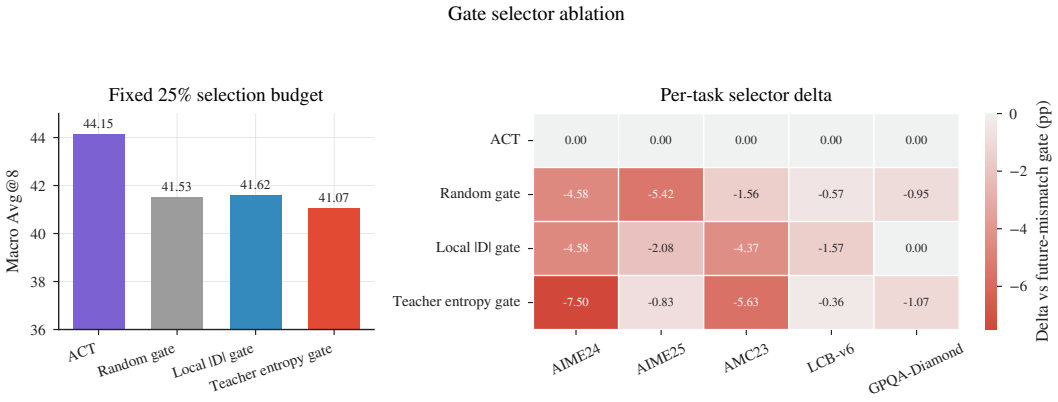

Figure 9: Gate-selector ablation under a fixed 25% token budget. The heat map shows each selector's per-task difference from the future-mismatch gate.

G Selection Ratio, Horizon, and Mixing Sensitivity

We compare selection ratios $\rho \in \{0, 0.25, 0.5, 1.0\}$, horizons $H \in \{8, 32, 64\}$, and mixing strengths $\beta \in \{0.25, 0.5, 1.0\}$, varying one temporal hyperparameter at a time.

Table 7: Single-factor sensitivity of ACT (Avg@8). Default: $\rho = 0.25, H = 32, \beta = 0.5$. The $\rho = 0$ endpoint is ACT-Directional.

Sweep	Setting	A24	A25	AMC	LCB	GPQA	Avg.
Selection ratio	$\rho = 0$	39.58	29.58	79.69	22.14	35.61	41.32
Selection ratio	$\rho = 0.25$	46.25	33.33	81.88	23.29	35.98	44.15
Selection ratio	$\rho = 0.5$	42.08	32.08	80.94	22.86	33.21	42.23
Selection ratio	$\rho = 1.0$	38.33	31.67	80.63	22.93	34.28	41.57
Horizon	$H = 8$	41.67	32.92	80.63	22.07	35.35	42.53
Horizon	$H = 32$	46.25	33.33	81.88	23.29	35.98	44.15
Horizon	$H = 64$	40.83	33.33	79.69	21.93	35.80	42.32
Mixing strength	$\beta = 0.25$	38.75	34.58	79.69	23.57	33.08	41.93
Mixing strength	$\beta = 0.5$	46.25	33.33	81.88	23.29	35.98	44.15
Mixing strength	$\beta = 1.0$	39.17	34.17	79.06	22.57	34.15	41.82

Table 7 has a clear peak at the default configuration. Both extremes of selection ratio ($\rho = 0$ and $\rho = 1.0$) remain close to the ACT-Directional baseline, indicating that the benefit of temporal feedback depends on selective allocation. At $\rho = 1.0$, Macro Avg@8 is only 0.25 points above $\rho = 0$, while AIME24 decreases from 39.58 to 38.33. When every position receives temporal correction, the gate no longer selects among positions; every position instead mixes its shaped directional coefficient with the window average. This result is consistent with the uniform-matched control in Section 6.3. The horizon sensitivity shows diminishing returns beyond 32 tokens ($H = 64$ loses 1.82 points), consistent with the discounted kernel concentrating most weight on nearby positions while longer horizons introduce noise from distant, less-relevant trajectory regions. Mixing strength exhibits a similar inverted-U: $\beta = 0.25$ barely departs from local feedback (the temporal correction is attenuated by 75%), while $\beta = 1.0$ over-commits to the window average at the expense of local information, losing the beneficial anchor that the current-token coefficient provides.

H Training Cost and Optimization Dynamics

Figure 10 tracks two key optimization diagnostics during training. Policy entropy and gradient norm trajectories are broadly similar across methods, so the available measurements do not support attributing ACT's gain to greater exploration or a larger overall update scale. TrOPD's slightly higher gradient norm co-occurs with its additional teacher-prefix supervision. Together with the position-selection ablations, these results support attributing the performance difference primarily to how modified coefficients are allocated across token positions.

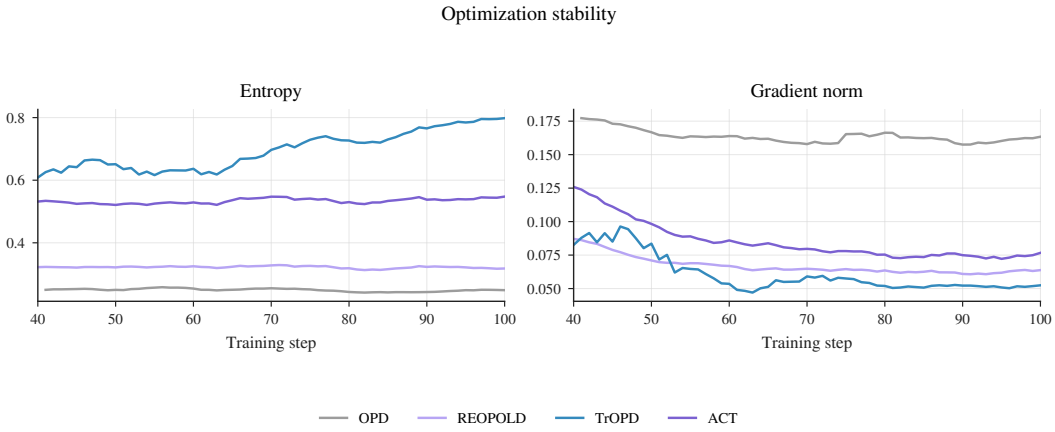

Figure 10: Seven-step smoothed policy entropy and gradient norm during training for OPD, REOPOLD, TrOPD, and ACT.

Figure 11 shows that all methods operate near the 8,096-token training cap with closely matched mean rollout lengths, making response length an unlikely explanation for ACT's performance advantage. Boundary-aware normalization further ensures that the final $H - 1$ positions are renormalized as the valid window shrinks, avoiding artificial attenuation from zero padding.



Figure 11: Mean rollout length during training.

Teacher uncertainty and the retained FKL branch. Figure 12 compares the fraction of teacher-high-entropy positions and the FKL auxiliary loss during training for EOPD, ACT-Directional, and ACT. Their mean high-entropy fractions are 14.9%, 18.6%, and 20.6%, respectively, and remain broadly stable during training, revealing systematic differences in teacher uncertainty along the trajectories visited by each objective. The FKL auxiliary losses remain on the order of 10^{-3}.

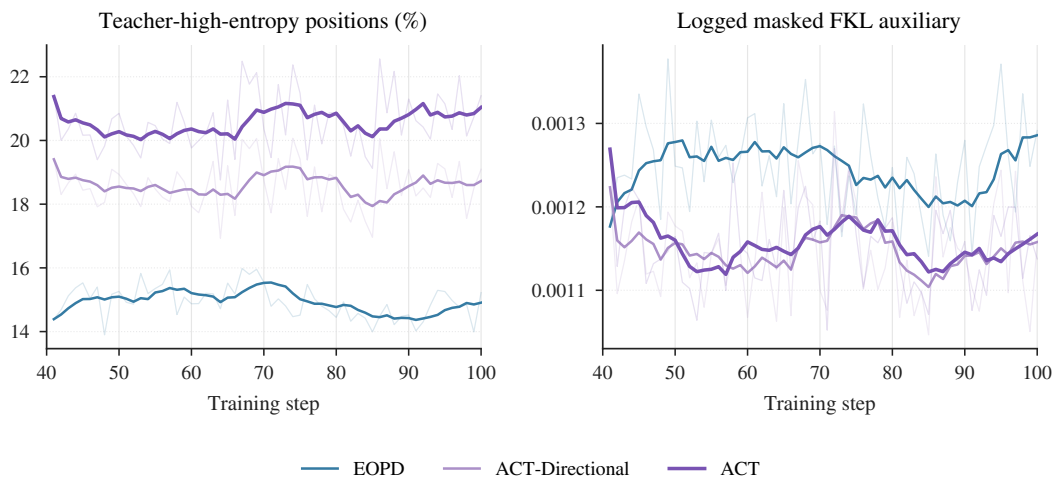

Figure 12: Teacher-high-entropy position ratio (left) and entropy-gated FKL auxiliary loss averaged over valid response tokens (right) during training. Light curves show per-step values; dark curves show seven-step moving averages.

I Inference-Budget Details

Table 8: Math Avg@8 from post-hoc prefix rescoring under common token budgets.

Method	Benchmark	8k	16k	32k	64k
OPD	AIME24	15.00	31.67	36.25	36.25
EOPD	AIME24	17.50	34.17	37.92	37.92
ACT	AIME24	**20.42**	**42.50**	**46.25**	**46.25**
OPD	AIME25	19.17	30.42	31.25	31.25
EOPD	AIME25	**21.67**	29.58	32.92	32.92
ACT	AIME25	20.42	**32.08**	**33.33**	**33.33**
OPD	AMC23	53.12	67.19	71.25	71.25
EOPD	AMC23	57.81	72.50	75.00	75.00
ACT	AMC23	**59.69**	**78.44**	**81.25**	**81.88**

Table 8 shows that ACT is best at all four budgets on AIME24 and AMC23 and at every AIME25 budget except 8k. Its advantage remains after 16k, so the main result is not solely a consequence of longer outputs; the saturation from 32k to 64k also indicates that these math comparisons largely cover the effective generation length.

J Token-Level Insight Visualizations

To provide qualitative support for the mechanism analysis, we extract token-level measurements from eight cross-prompt evaluation cases. These visualizations use the trained ACT model and its corresponding teacher and student log probabilities.

In Figure 13, after approximately matching the current-position D_t, relative position, and teacher entropy, the left panel finds two future windows with $M_t^+ \approx 0$ and 0.88, showing that in these samples local feedback does not determine short-horizon future mismatch. The center panel shows that high mismatch can arise from either sign-consistent accumulation or alternating positive and negative coefficients; observing only net feedback therefore misses the latter. The right panel gives a concrete cancellation case with $M_t = 0.81$ but $|F_t| = 0.06$, consistent with the cancellation-mass interpretation in Equation 5.

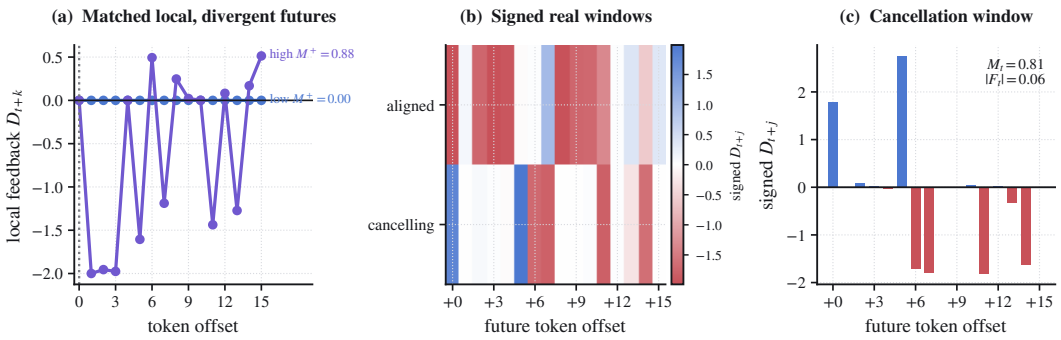

Figure 13: Token-level diagnostics. Left: two measured windows with nearly matched local feedback but sharply different future mismatch. Center: sign-consistent and sign-cancelling patterns in high-mismatch windows. Right: signed token coefficients in a measured cancellation window.

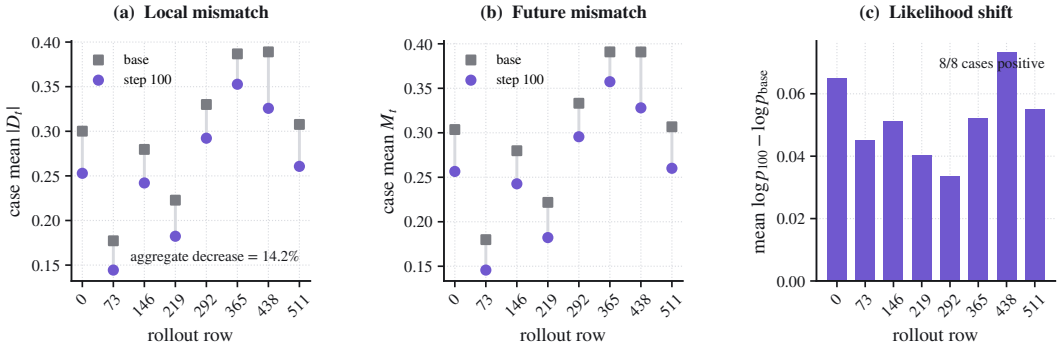

Figure 14: Changes from the base student to the step-100 model on fixed responses. Left: local mismatch by case. Center: future mismatch by case. Right: mean step-100 minus base log probability on the same tokens.

In Figure 14, both local and future mismatch decrease from the base student to step 100 in all eight cases, with a 14.2% decrease in aggregate mean $|D_t|$. Mean log probability on the same fixed responses increases in $8/8$ cases, showing that the trained student assigns greater probability to these measured tokens.

K Additional Diagnostics

Coverage under repeated sampling. Pass@k in Figure 15 complements Avg@8 by measuring whether at least one of k samples solves a problem. ACT achieves the highest Pass@8 on AIME24, AIME25, and AMC23, indicating that selective temporal correction improves not only mean accuracy but also the probability of generating at least one correct solution among eight samples.

L Multi-Domain Data and Full Results

After prompt-length filtering, the multi-domain training set contains 1,193,568 examples: 71.2% math, 20.4% code, and 8.4% science. Figure 16 shows that ACT's advantage is consistent across all five benchmarks in the multi-domain setting. The relative gains are largest on math tasks but extend to code and science, showing that future-mismatch selection applies across different trajectory structures. The multi-domain teacher (Skywork-OR1-7B) differs from the math-domain teacher (Skywork-OR1-Math-7B), yet ACT retains its advantage without hyperparameter changes, suggesting that the temporal mechanism has some robustness in this domain-mixture setting.

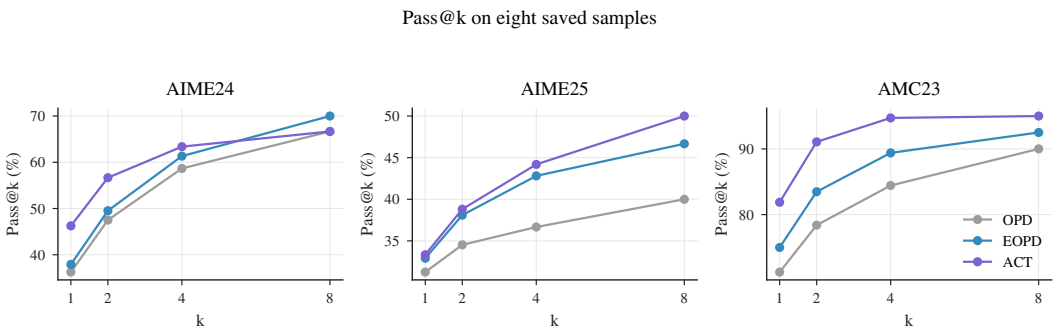

Figure 15: Pass@k computed from eight samples per question.

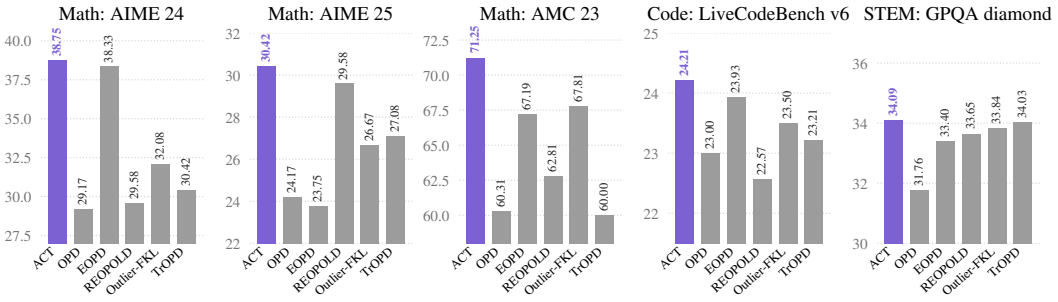

Figure 16: Per-benchmark multi-domain results under the common Avg@8 protocol. ACT is purple and all baselines are gray.

M Cross-Model-Scale Replication

To test whether the gain transfers beyond the main student–teacher pair, we replicate OPD, EOPD, and ACT with a Qwen3-0.6B student and a Qwen3-8B teacher (Yang et al., 2025). Table 9 reports the results after 100 training steps under a fixed 32,768-token Avg@8 evaluation protocol.

Table 9: Qwen3-0.6B → Qwen3-8B replication under the fixed 32K Avg@8 protocol. All values are percentages.

Method	AIME 24	AIME 25	AMC 23	LiveCodeBench v6	GPQA diamond	Avg.
Initial student	12.08	18.75	51.56	16.36	28.09	25.37
Teacher reference	75.42	68.75	95.00	49.57	60.29	69.81
OPD	20.83	25.00	56.56	19.71	22.54	28.93
EOPD	20.00	25.83	60.31	20.93	26.77	30.77
ACT	27.92	26.67	63.12	21.50	28.47	33.54

ACT reaches a Macro Avg@8 of 33.54, improving over OPD by 4.61 points and EOPD by 2.77 points. It also exceeds both baselines on every individual benchmark, rather than obtaining the average gain from a single outlier task, and improves the initial student by 8.17 points. This replication changes the student model family, teacher family, and model scales while preserving the ordering observed in the main study, providing evidence that selective temporal feedback transfers beyond the original 1.5B–7B pair.

N Behavioral Cases

We compare sampled OPD and ACT completions under the same prompt.

N.1 Case 1: Fluent Sentence-Level Self-Correction Loop

Problem AIME 2024 · Answer: 294

Consider length-16 paths from the lower-left to the upper-right corner of an 8×8 grid. How many change direction exactly four times?

OPD Student 2/8 correct

Okay, so I need to count paths that change direction exactly four times.
Wait, perhaps I should count four runs. Wait, no, four changes means five runs. Wait, let me split the right and up moves again ...
For $r = 1$: 35. For $r = 2$: 147. For $r = 3$: 147. For $r = 4$: 35.
Thus, $35 + 147 + 147 + 35 = 364$.
[The trajectory repeatedly reopens the same run-count split before ending at 364.]

ACT Student 7/8 correct

Four direction changes create five alternating positive runs. If the path begins with R, split the eight R-moves into three positive runs and the eight U-moves into two: $\binom{7}{2}\binom{7}{1} = 147$. The same count holds when the path begins with U, so the total is $2 \times 147 = \boxed{294}$.

The OPD completion remains fluent and locally mathematical, but repeated self-correction never stabilizes the meaning of four direction changes and ends in an overcount. ACT more often commits to the five-run representation and carries that state consistently to the final answer.

N.2 Case 2: Self-Dialogue Loop Rewriting the Problem State

Problem AIME 2025 · Answer: 237

Let A be the positive divisors of 2025, and select a subset B of A uniformly at random. If $\Pr[\mathrm{lcm}(B) = 2025] = m/n$, find $m + n$.

OPD Student 0/8 correct

Wait, P is the power set of S, so P is the set of all subsets of S. Wait, but the problem says ``subsets of P.'' Wait, no, P is the power set of S ...
Wait, A is $\{1, 5, 2, 10, 4, 8\}$. So A has 6 elements. Wait, but earlier I thought A had 51 elements ...
[The model keeps posing and answering altered versions of the task.] Extracted answer: 1.

ACT Student 1/8 correct

Since $2025 = 3^4 5^2$, A contains 15 divisors. A subset has LCM 2025 iff it contains a divisor with 3-exponent 4 and one with 5-exponent 2. Inclusion–exclusion gives $2^{15} - 2^{12} - 2^{10} + 2^8 = 27904$ valid subsets. Thus $\Pr = 27904/32768 = 109/128$, and $m + n = \boxed{237}$.

OPD repeatedly restates the task while silently replacing its divisor lattice by unrelated sets and power sets. ACT's successful sample preserves the two prime-exponent coordinates long enough to complete one consistent inclusion–exclusion count.

N.3 Case 3: Periodic Junk-Token Loop at the Decoding Cap

Problem AMC · Answer: 4

A hidden 2×1 rectangle covers two adjacent cells of a 3×3 grid. What is the minimum number of guesses that guarantees hitting it?

OPD Student 0/8 correct · 65k-token cap

```
Okay, so I need the minimum number of squares that must be guessed.
First, let me visualize the 3 × 3 grid ...
points. Wait, points. Wait, points. Wait, points. Wait, points.
Wait, points. Wait, points. Wait, points. Wait, points. Wait ...
```

[The same short fragment repeats until the 65,536-token cap.] `Extracted answer: 13.`

ACT Student 6/8 correct

The unguessed cells must contain no adjacent pair; otherwise that pair could be the hidden rectangle. Hence they form an independent set of the 3×3 grid, whose maximum size is 5. At least $9 - 5 = 4$ cells must therefore be guessed, and a four-cell checkerboard pattern achieves the bound. The answer is $\boxed{4}$.

Unlike the first two cases, this trajectory eventually loses semantic content altogether and collapses into a short periodic fragment. The failure is trajectory-level: locally common tokens remain easy to predict even though the continuation can no longer reach a valid answer.

N.4 Case 4: Correct Reasoning Without an Executable Answer

Problem LiveCodeBench · Expected: executable code

An $n \times n$ cargo deck has one cell per container. Each container weighs w, and total weight cannot exceed `maxWeight`. Return the maximum number of containers.

OPD Student 1/8 correct

```
The deck has n squared cells, so the spatial limit is n*n. The weight
limit gives maxWeight // w containers. Therefore the answer should
be the smaller of these two values. Wait, no. Wait, perhaps ...
No. No. No. No. No. No. No. No. No. No. No. No ...
```

[The rollout never emits a fenced program or class method.] `Extracted code: <empty>.`

ACT Student 8/8 correct

```
class Solution:
  def maxContainers(self, n: int, w: int, maxWeight: int):
    return min(n * n, maxWeight // w)
```

The OPD student reaches the key inequality but fails the task-level contract: no executable program is produced. This separates reasoning quality from completion quality. In these samples, ACT more often carries the valid idea through the required output format.

N.5 Case 5: Answer-Channel Drift in Multiple Choice

Problem GPQA · Answer: C

A synchrocyclotron problem asks for the number of proton revolutions. The response must end with exactly one option letter from A, B, C, or D.

OPD Student 1/8 correct

`Wait, perhaps the problem expects` $N = (U0 * T)/($`something else`$)$`.`
`Wait, perhaps the problem expects` $N = (U0 * T)/($`something else`$)$`.`
`Wait, perhaps the problem expects` $N = (U0 * T)/($`something else`$)$ `...`
[Unresolved formula guesses continue and the option constraint is lost.] `Final extracted token: U.`

ACT Student 4/8 correct

The energy gain accumulates over the two accelerating gaps per revolution, while the changing relativistic frequency keeps the proton synchronized with the alternating voltage. Substitution yields the option corresponding to 3536 revolutions. **ANSWER: C**

This failure occurs after substantial domain reasoning: OPD drifts out of the constrained answer channel and terminates with an invalid symbol. ACT still reasons verbosely, but more often preserves the final answer format required by the task.

N.6 Discussion

These failures span fluent wrong reasoning, self-dialogue that rewrites the task, non-semantic token loops, and output-format drift. They therefore cannot be identified by fluency or single-token plausibility alone. Their common feature is that, once an early decision establishes a bad trajectory, later tokens can remain locally self-consistent. ACT's use of within-response future mismatch to select feedback positions is consistent with addressing this failure pattern. It is not, however, an outcome verifier and cannot guarantee that every sample is correct on difficult problems.

N.7 Token-Level Teacher Support and ACT Selection

To inspect which positions the gate selects, we post-hoc rescore three complete ACT evaluation responses from mathematics, code, and science. Fill color encodes the frozen teacher's log probability on the student-sampled token: gray denotes lower support, purple denotes higher support, and shade intensity denotes its magnitude. A thin red outline marks the top 25% positions selected by ACT using the full-response future mismatch $M_t = (K|D|)_t$.

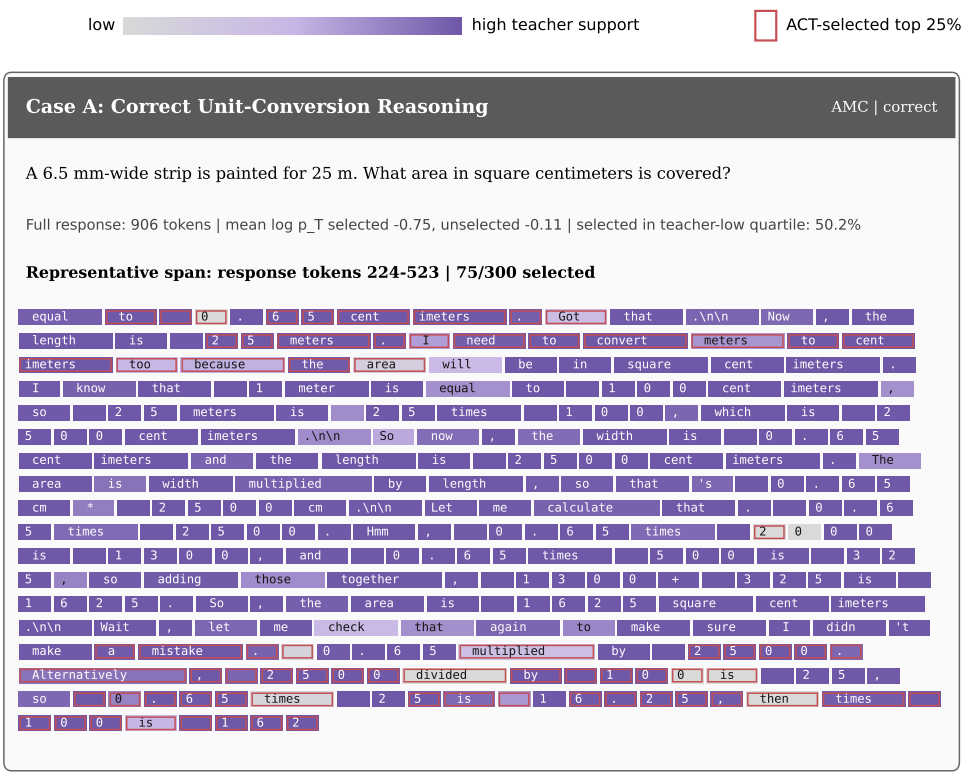

Figure 17: Teacher support and ACT token selection in the AMC mathematics response.



Figure 18: Teacher support and ACT token selection in the LiveCodeBench code response.

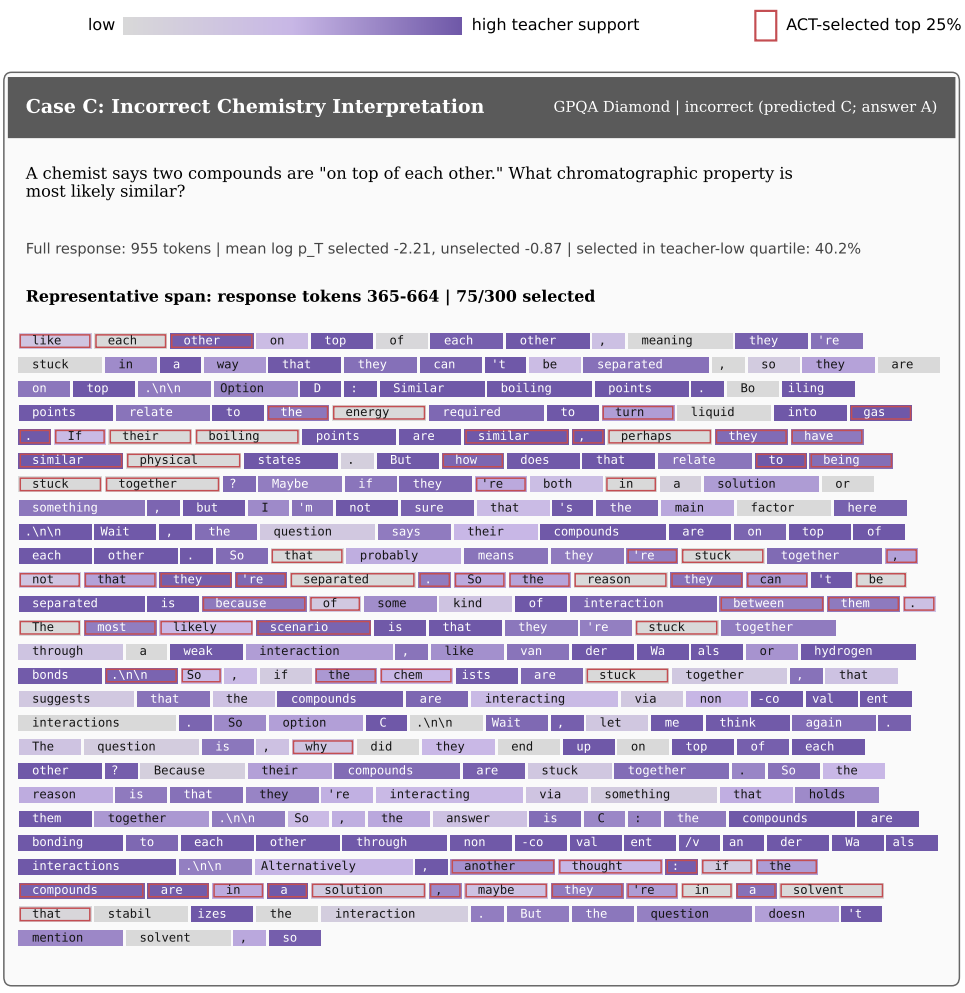

Figure 19: Teacher support and ACT token selection in the incorrect GPQA response.

For the correct mathematics, correct code, and incorrect chemistry cases, mean teacher log probability on selected tokens is $-0.75, -0.82, -2.21$, compared with $-0.11, -0.33, -0.87$ on unselected tokens. The overlap between ACT selection and the lowest teacher-support quartile is 50.2%, 41.4%, and 40.2%, respectively, all above the 25% random baseline, indicating that future-mismatch selection enriches low-teacher-support regions. Yet more than half of selected positions still fall outside the lowest quartile, confirming that ACT is not a low-teacher-probability filter: it selects positions with sustained mismatch over the future window.

O REPRODUCIBILITY

O.1 MODELS

Table 10: Model specifications.

Role	Model	Parameters	Vocab.	Source
Student	DeepSeek-R1-Distill-Qwen-1.5B	1.5B	151,936	DeepSeek
Math teacher	Skywork-OR1-Math-7B	7B	151,936	Skywork
Multi-domain teacher	Skywork-OR1-7B	7B	151,936	Skywork

All models in Table 10 share the same tokenizer, ensuring exact token-level correspondence between teacher and student log probabilities. Both settings retain the same 1.5B student and change only the 7B teacher matched to the training domains.

O.2 TRAINING DATA

Table 11: Training data composition.

Setting	Domain	Prompts	Fraction
Math-domain	Math	849,920	100%
Multi-domain	Math	849,920	71.2%
	Code	243,952	20.4%
	Science	99,696	8.4%
	Total	1,193,568	100%

Table 11 shows that the multi-domain set remains math-heavy (71.2%) while adding 20.4% code and 8.4% science. Prompts exceeding 2,048 tokens are filtered before training.

O.3 HYPERPARAMETER SETTINGS

Table 12 fixes the optimizer, 100-step training schedule, rollout count, length caps, hardware scale, and final evaluation configuration for all methods, so method comparisons do not introduce additional training compute, different optimization settings, or different decoding settings. ACT sensitivity experiments vary only one of ρ, H, β at a time.

O.4 EVALUATION BENCHMARKS

Math benchmarks allow continuation to 65,536 tokens when the initial 32,768-token generation is truncated; LiveCodeBench and GPQA do not use continuation. Table 12 fixes eight samples per question, temperature 0.6, and top-$p = 0.95$, and uses Macro Avg@8 as the summary metric. All reported results use the model obtained after 100 optimization steps. LiveCodeBench uses execution-based judging; all other benchmarks use string-matching answer extraction.

Table 12: Training, method, and final-evaluation hyperparameters used in the main experiments. Method-specific parameters apply only to the corresponding method; all remaining settings are shared.

Hyperparameter	Value
Data and batching	
Training data	Math-domain or multi-domain data; see Table 11
Data order	Shared by all methods
Batch size (prompts per step)	128
Rollouts per prompt	4
Maximum prompt length	2,048 tokens
Maximum response length	8,096 tokens
Policy epochs per rollout batch	1
Optimization	
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
Learning rate	5×10^{-6}
Optimizer steps	100
Loss aggregation	Token mean
Entropy bonus	None
Training framework	VERL/FSDP + vLLM (Sheng et al., 2025; Kwon et al., 2023)
GPUs per run	8
ACT defaults	
Horizon H	32
Discount γ	0.95
Selection ratio ρ	0.25
Mixing strength β	0.5
Shaping threshold τ	2
EOPD auxiliary branch	
Top-k support size	16
Entropy threshold τ_H	0.8
Auxiliary weight α	1.0
Final evaluation	
Samples per question	8
Temperature	0.6
Top-p	0.95
Per-benchmark metric	Avg@8 (mean correctness over 8 samples)
Avg. aggregation	Equal weight across 5 benchmarks
Bootstrap replicates	20,000 (stratified by benchmark)

Table 13: Evaluation benchmarks and inference-length settings.

Benchmark	Type	Questions	Token cap	Continuation
AIME 2024	Math competition	30	32,768	65,536
AIME 2025	Math competition	30	32,768	65,536
AMC 2023 (Evalchemy 40)	Math competition	40	32,768	65,536
LiveCodeBench v6	Code generation	175	32,768	–
GPQA-Diamond	Science QA	198	32,768	–