TEST-TIME RL FOR DIFFUSION LANGUAGE MODELS: GEOMETRY-GUIDED SELECTIVE UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time reinforcement learning (TTRL) offers a route to self-evolution from unlabeled experience, but agreement among sampled answers does not establish which targets should be reinforced. We introduce Geometry-Selective Less-is-More TTRL (Geo.-LiM), a framework for persistent, label-free adaptation of diffusion language models (dLLMs). A frozen policy collects evidence from denoising geometries matched in the number of function evaluations (NFE). Strict pooled-majority support and novelty relative to a deterministic frozen-reference answer jointly determine update eligibility. This rule induces a binary self-reward that assigns positive credit only to trajectories supporting eligible candidates; selected traces update a persistent adapter through positive-only replay with fixed-reference KL regularization. Across three models and five benchmarks, Geo.-LiM improves post-adaptation performance over the unadapted backbone in 27 of 30 model–task–length settings. Compared with our consensus-based TTRL baseline under the same eight-rollout evidence budget, it achieves higher performance in 23 of the 30 settings while admitting fewer prompts and taking fewer optimizer steps in every setting. The corresponding grid totals decrease by 74.2% and 64.4%, respectively. These results support geometry-guided selective reinforcement as a practical approach to persistent dLLM adaptation without target labels.

1 INTRODUCTION

Reinforcement learning (RL) has improved the reasoning capabilities of diffusion language models (dLLMs) (Zhao et al., 2026; Zhu et al., 2026). Yet target problems need not arrive with the labels or verifiable feedback used during post-training. Requiring new supervision before adapting limits the ability to learn directly from these inputs. This motivates *self-evolution from unlabeled experience*: enabling models to construct learning signals from their own responses and improve without access to target labels. We focus on test-time experience, where the problems a model is asked to solve also provide the data from which it can learn. The goal is to turn this unlabeled experience into persistent parameter adaptation, rather than requiring a separately supervised training stage.

Test-time reinforcement learning (TTRL) provides a route to this goal by deriving pseudo-rewards from repeated generations on unlabeled target inputs (Zuo et al., 2026). For dLLMs, prior work has investigated inference-time refinement without parameter updates (Lee et al., 2026; Wan et al., 2026; Bai et al., 2026) and reinforcement learning during post-training (Zhao et al., 2026; Zhu et al., 2026; Wang et al., 2026a). We study the complementary setting of persistent, label-free dLLM TTRL, in which self-generated evidence from the target stream updates a shared adapter retained for evaluation. Here, sampled responses are not merely candidates for the current prediction: they become training targets that modify the policy used across inputs. The central challenge is therefore to determine which targets should receive reinforcement. Repeated agreement provides support for a candidate answer, but does not by itself establish the value of reinforcing that answer.

dLLMs offer a structural source of evidence for this decision. Their block-wise denoising process admits multiple generation factorizations, which expose different conditional contexts (Nie et al., 2026; Ye et al., 2025). By holding the model, prompt, generation length, and number of function evaluations (NFE) fixed, we can vary the denoising geometry to obtain controlled evidence views. We use these views not only as decoding choices, but as a source of support for test-time learning targets. This support serves a different purpose from comparison with a frozen reference: pooled evidence

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.

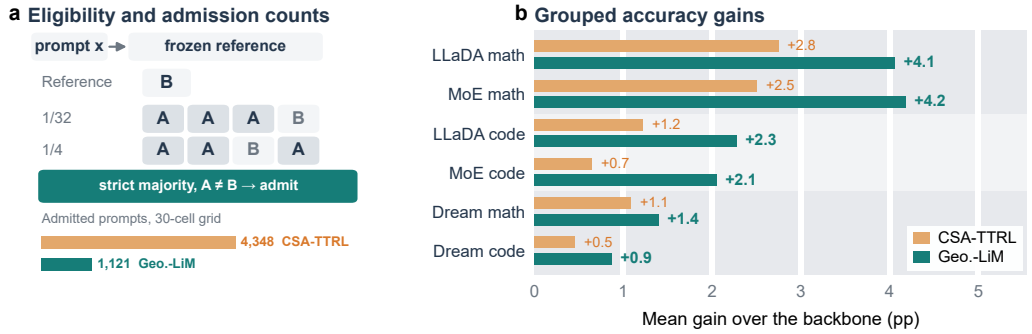


Figure 1: **Selective admission and grouped gains. (a)** One admitted prompt: the candidate has strict pooled-majority support and differs from the frozen reference. The two geometries are one sampled, NFE-matched pair. Bars are 30-cell admission totals (Appendix Table 9). **(b)** Mean backbone-relative gain by architecture–task group (Section 4.2).

indicates whether a candidate is supported across the sampled geometries, whereas reference-relative novelty identifies whether it proposes an answer-level change. We combine these signals to restrict reinforcement to supported alternatives to the reference, without treating novelty as an estimate of correctness or update utility.

Building on this distinction, we introduce **Geometry-Selective Less-is-More TTRL (Geo.-LiM)**. For each test input, a frozen evidence policy produces a deterministic reference answer and eight stochastic trajectories, with four trajectories from each of two reproducibly sampled, NFE-matched geometries. An update is admitted only when the reference and candidate are valid, the candidate receives strict pooled-majority support, and it differs from the reference. This eligibility rule induces a binary self-reward: trajectories supporting an admitted candidate receive positive credit, while all others receive zero credit. One supporting trace per geometry is then replayed through the evolving LoRA-adapted policy under fixed-reference KL regularization. The adapter is retained across the test stream and evaluated afterward through fresh predictions on the same inputs; labels do not guide admission, credit assignment, checkpoint selection, or stopping.

Our contributions are threefold. **(i) Persistent, label-free TTRL for dLLMs.** We formulate and evaluate a setting in which dLLMs learn from their own unlabeled test-time experience, with a frozen evidence policy and a persistent adaptation policy. **(ii) Geometry-guided selective reinforcement.** We introduce an update-eligibility rule that combines pooled support across NFE-matched denoising geometries with frozen-reference novelty, and uses this rule to assign positive credit to selected trajectories. **(iii) Improved adaptation with fewer updates.** Across 30 model–task–length settings spanning three models and five benchmarks, Geo.-LiM improves 27 endpoints over the unadapted backbone and exceeds the matched consensus-based baseline, CSA-TTRL, on 23. It admits fewer prompts and executes fewer optimizer-bearing steps in every setting, reducing the corresponding grid totals by 74.2% and 64.4%, respectively. Figure 1 shows one admitted decision under this rule and the mean backbone-relative gains by architecture–task group. Geometry-source and count-matched controls further examine the roles of evidence geometry and target selection beyond update volume.

2 Related Work

Diffusion language models and test-time geometry. LLaDA, LLaDA-MoE, and Dream instantiate iterative unmasking with dense, sparse, and autoregressively initialized designs (Nie et al., 2026; Zhu et al., 2025; Ye et al., 2025); block diffusion connects left-to-right and arbitrary-order generation (Arriola et al., 2025). Training-time RL methods improve dLLM reasoning with task rewards or verifiers (Zhao et al., 2026; Zhu et al., 2026; Rojas et al., 2026; Pan et al., 2026). At inference time, trajectory refinement, reward-free guidance, energy steering, and trajectory stitching refine a response without retaining parameter changes (Dang et al., 2025; Chen et al., 2025; Li et al., 2026b; Miles et al., 2026); most closely, HEX aggregates predictions across block schedules (Lee et al., 2026), Prism combines hierarchical search with self-verification (Bai et al., 2026), and GeoBlock selects block granularity

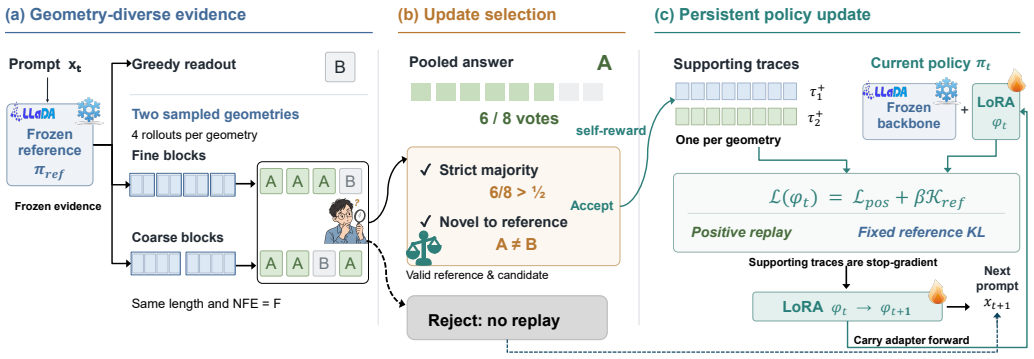


Figure 2: **Geometry-Selective Less-is-More TTRL (Geo.-LiM). (a) Geometry-diverse evidence.** A frozen policy produces a deterministic reference answer and four rollouts from each of two sampled, NFE-matched denoising geometries. **(b) Update selection.** Updates require valid reference and candidate answers, strict pooled-majority support, and reference-relative novelty. **(c) Persistent policy update.** For admitted prompts, one supporting trace per geometry is replayed through the LoRA-adapted policy using a positive-only objective with fixed-reference KL regularization. The adapter is retained across the test stream; rejected prompts contribute no replay gradient.

from dependency geometry (Wan et al., 2026). These establish denoising geometry as an inference variable; we test its value as evidence for persistent updates.

Label-free test-time adaptation. Test-time training exploits structure available at deployment (Sun et al., 2020), while TTRL turns repeated generations into pseudo-rewards (Zuo et al., 2026). Subsequent work improves reward estimation and confidence handling (Du et al., 2026; Li et al., 2026a), draws complementary supervision from analogous inputs or a teacher (Zhang et al., 2026), and characterizes uncertainty in unsupervised RLVR (He et al., 2026)—all on autoregressive backbones. Closest in spirit within dLLMs, Temporal Consistency Reinforcement rewards semantic stability across denoising steps and so also optimizes without labels (Wang et al., 2026a), but its stability axis is the denoising trajectory of a single rollout rather than the block factorization across rollouts, and it optimizes offline on a separate training split rather than adapting on the unlabeled target test inputs.

Selecting what to learn from. Learning from a chosen subset rather than from everything is established in RLVR: critical instances can be identified without labels (Xiao et al., 2026), and filtering medium-consistency samples sharpens advantage estimates (Yu et al., 2026). Those selections operate on a reusable corpus, where a discarded example can be revisited and a bad update diluted by later epochs. In our protocol, admitted test inputs update a shared adapter retained for final evaluation. Selection is therefore a control over both optimization cost and cross-prompt interference.

3 METHOD

Geo.-LiM learns from unlabeled test-time experience by separating evidence construction from policy adaptation. Its central design is to reinforce only candidates that receive geometry-diverse pooled support and differ from a frozen reference answer. Figure 2 summarizes the three components: geometry-diverse evidence, selective update eligibility, and persistent policy adaptation.

Setup. Given a prescribed unlabeled stream $\mathcal{D}_{\text{test}} = (x_1, \ldots, x_T)$, we keep the pretrained parameters θ_0 fixed and adapt only the LoRA parameters ϕ. The frozen evidence policy $\pi_{\text{ref}} = \pi_{\theta_0,\phi_0}$ generates all reference predictions and sampled trajectories, whereas $\pi_t = \pi_{\theta_0,\phi_t}$ is the evolving learner. Fixing the evidence policy prevents preceding adapter updates from changing the distribution used to determine eligibility. For input x_t, the reference answer is $a_{\text{ref}}(x_t) = g(\text{decode}_0(x_t; \pi_{\text{ref}}))$, where decode_0 denotes deterministic canonical decoding and g returns a task-specific answer representation. Malformed, unparsable, or prompt-invalid outputs map to $\varnothing$. For code, equality under g is Python abstract syntax tree comparison, not execution-based functional equivalence.

3.1 GEOMETRY-DIVERSE EVIDENCE AT MATCHED NFE

We construct evidence by varying the dLLM's block-wise denoising factorization. A geometry $r \in \mathcal{G}_0 = \{1/32, 1/16, 1/8, 1/4\}$ specifies a relative block size, with $b_r = rL$ for generation length L. For s_r denoising steps per block and per-trajectory budget F, we match the number of function evaluations (NFE) through

$$s_r = \frac{Fb_r}{L}, \qquad \mathrm{NFE}(r) = \frac{L}{b_r} s_r = F, \tag{1}$$

subject to the implementation's integer rounding. Thus the prompt, reference policy, generation length, and NFE budget are held fixed while the denoising factorization changes. Relative block sizes keep the geometry scale comparable across lengths.

For each input, an unordered pair $\mathcal{G}_t \subset \mathcal{G}_0$, $|\mathcal{G}_t| = 2$, is assigned reproducibly by hashing the global seed and stable prompt identifier into the six possible pairs. From each $r \in \mathcal{G}_t$, the frozen policy generates $R = 4$ trajectories $y_{r,j} \sim \pi_{\mathrm{ref}}(\cdot \mid x_t, r)$, $j = 1, \ldots, R$, giving a fixed evidence budget $B = |\mathcal{G}_t| R = 8$. This allocation preserves repeated sampling within each geometry while collecting evidence under two distinct factorizations.

3.2 FROM POOLED SUPPORT TO UPDATE ELIGIBILITY

Candidate support and reference novelty serve distinct roles: the former aggregates evidence from the sampled factorizations, while the latter excludes targets that confirm the frozen readout. For a valid answer $a \neq \varnothing$, define pooled support and the modal candidate by

$$\begin{aligned} C_t(a) &= \sum_{r \in \mathcal{G}_t} \sum_{j=1}^{R} \mathbf{1}[g(y_{r,j}) = a]\,, \\ a_t^\star &= \arg\max_{a \neq \varnothing} C_t(a). \end{aligned} \tag{2}$$

Invalid outputs contribute no vote. We set $a_t^\star = \varnothing$ for a tie or an absence of valid votes, with $C_t(\varnothing) = 0$ by convention. The update-eligibility rule is

$$\begin{aligned} A_t = \mathbf{1}[a_{\mathrm{ref}}(x_t) \neq \varnothing,\ a_t^\star \neq \varnothing] \\ \times \mathbf{1}[C_t(a_t^\star) > B/2]\, \mathbf{1}[a_t^\star \neq a_{\mathrm{ref}}(x_t)]\,. \end{aligned} \tag{3}$$

With $B = 8$, admission requires at least five votes; invalid rollouts do not reduce this denominator. Because each geometry contributes at most four votes, an admitted candidate necessarily has support from both views, although neither an independent majority in each view nor equal support across views is required. Reference novelty further restricts reinforcement to parser-defined answer-level changes, including unanimous candidates when they differ from the reference. This is an operational eligibility criterion, not an estimator of candidate correctness or parameter-update utility. Conditioned on the input, geometry assignment, and sampling seed, A_t is independent of the current adapter ϕ_t.

3.3 SELF-REWARD AND PERSISTENT POLICY ADAPTATION

Eligibility induces a binary self-reward over the frozen evidence trajectories:

$$\mathcal{R}_{r,j}(x_t) = A_t\, \mathbf{1}[g(y_{r,j}) = a_t^\star]\,. \tag{4}$$

Only trajectories supporting an admitted candidate receive positive credit; all alternatives and rejected prompts receive zero credit. For an admitted prompt, we retain one supporting denoising trace τ_r^+ from each $r \in \mathcal{G}_t$. This selection retains evidence from both geometries without weighting a view in proportion to its number of supporting rollouts.

Let $(s, k) \in \tau_r^+$ index a replayed prediction at denoising step s and token position k, with fixed state $z_{r,s}^+$ and sampled token $y_{r,k}^+$. Write $p_\phi^{r,s,k}$ and $p_{\mathrm{ref}}^{r,s,k}$ for the learner's and frozen reference's categorical token distributions at position k, conditioned on $(x_t, z_{r,s}^+)$. The objective combines positive replay

with fixed-reference KL regularization on the same denoising states:

$$\mathcal{L}_t(\phi) = \frac{1}{|\mathcal{G}_t|} \sum_{r \in \mathcal{G}_t} \frac{1}{|\tau_r^+|} \sum_{(s,k) \in \tau_r^+} \Big[-\log p_\phi^{r,s,k}(y_{r,k}^+) + \beta \, D_{\mathrm{KL}}\Big(p_\phi^{r,s,k} \,\|\, p_{\mathrm{ref}}^{r,s,k} \Big) \Big]. \tag{5}$$

The normalization averages prediction events within each trace and then weights the two geometries equally. Sampled tokens, denoising states, and eligibility decisions are fixed; gradients flow only through the learner's adapter. Because evidence is generated by π_{ref} rather than π_t, this objective is a frozen-reference, reward-weighted replay surrogate, not an on-policy GRPO update.

Rejected prompts contribute no replay gradient. An optimizer step aggregates admitted signals within a synchronized minibatch; if none are admitted, both the adapter and optimizer states remain unchanged. The resulting adapter is retained across the test stream. Every prompt still incurs the reference readout and eight evidence rollouts, so rejection reduces replay and optimization rather than evidence collection. After the stream, the final adapter is evaluated by fresh canonical decoding on the same benchmark inputs, never by reusing adaptation trajectories. Benchmark labels are not used for eligibility, credit assignment, checkpoint selection, or stopping.

4 Experiments

4.1 Experimental Setup

Models and benchmarks. We evaluate LLaDA-8B-Instruct (Nie et al., 2026), LLaDA-MoE-7B-A1B-Instruct (Zhu et al., 2025), and Dream-7B-Instruct (Ye et al., 2025) on MATH500 (500) (Lightman et al., 2024), GSM8K (1,319) (Cobbe et al., 2021), Countdown (256) (Zhao et al., 2026), HumanEval (164) (Chen et al., 2021), and sanitized MBPP (427) (Austin et al., 2021), each at lengths 256 and 512. Countdown uses d1's three-number exact-arithmetic protocol; its prompt-visible checks are shared preprocessing, not an extra Geo.-LiM gate.

Compared methods. Each table leads with the unadapted backbone. Adapted cells report Final (Δ) against that row. **CSA-TTRL** is a matched label-free TTRL baseline: it updates when modes from two fixed schedules agree. **Geo.-LiM** is the full eligibility rule. Both use eight NFE-matched evidence trajectories and differ in eligibility and credit assignment. Labeled-RL rows are published supervised accuracies under a different training regime. They are context for the absolute score, and Δ is never taken against them: Diffu-GRPO/d1 (Zhao et al., 2026), VRPO (Zhu et al., 2026), wd1 (Tang et al., 2026), SAPO (Xie et al., 2026), GDPO (Rojas et al., 2026), and TraceRL (Wang et al., 2026b). Deltas are never taken against those rows.

Adaptation and evaluation. Adaptation uses persistent rank-16 LoRA (Hu et al., 2021), AdamW, fixed-reference KL coefficient $\beta = 0.01$, gradient clipping at 1.0, and evidence temperature 1.0. The canonical schedule uses $F = L/2$ NFE per rollout, and final evaluation is greedy. CSA-TTRL and Geo.-LiM share one learning rate, 10^{-6}, at every endpoint. CSA-TTRL uses absolute blocks $b \in \{16, 64\}$ with four samples per view; Geo.-LiM draws four trajectories from each of two sampled relative geometries. Both therefore use eight evidence rollouts per prompt. Settings with four and sixteen rollouts are sensitivity analyses (Appendix B), not per-endpoint choices. Following TTRL (Zuo et al., 2026), evaluation freshly decodes the same IDs after the stream; adaptation samples are never the reported predictions, and labels are not used for selection, credit, checkpointing, or stopping.

Reporting. Each main-grid endpoint is one seeded adaptation run. Within the run, the seed, prompt order, and stopping rule stay fixed and are recorded with the configuration. On the same T example IDs, let N_{W2C} be the number of wrong-to-correct changes and N_{C2W} the number of correct-to-wrong changes. The paired effect is

$$\Delta_{\mathrm{pair}} = \frac{N_{\mathrm{W2C}} - N_{\mathrm{C2W}}}{T}. \tag{6}$$

K and U count eligible prompts and optimizer-bearing minibatches. Exact-answer tasks use a fixed denominator and code uses pass@1. For the count-matched and retention contrasts we also

Table 1: **LLaDA-family math.** The first row of each block is the unadapted backbone. Italic headers separate published labeled-RL references from label-free test-time adaptation. Adapted cells are Final (Δ, pp) versus the backbone row. Bold/underline rank only CSA/Geo. Avg. Δ is the unweighted mean of backbone-relative endpoint differences, never an average over labeled and label-free regimes.

		Countdown		GSM8K		MATH500		Avg. Δ
Model	Method	256	512	256	512	256	512	(pp)
LLaDA	LLaDA-8B-Instruct	19.5 (+0.0)	16.0 (+0.0)	75.6 (+0.0)	78.2 (+0.0)	32.0 (+0.0)	34.6 (+0.0)	+0.0
	Published labeled-RL references							
	Diffu-GRPO	31.3	37.1	79.8	81.9	34.1	39.0	–
	VRPO	22.3	18.0	80.1	81.5	35.6	34.8	–
	wd1	51.2	46.1	80.8	82.3	34.4	39.0	–
	SAPO	52.0	56.3	80.6	82.1	33.8	38.4	–
	GDPO	64.1	60.2	81.1	82.0	37.0	38.5	–
	TraceRL	50.4	52.6	80.3	82.4	35.6	39.1	–
	Label-free test-time adaptation							
	CSA-TTRL	27.0 (+7.5)	17.2 (+1.2)	**75.9 (+0.3)**	80.4 (+2.2)	33.4 (+1.4)	**38.6 (+4.0)**	+2.8
	Geo.-LiM	**27.3 (+7.8)**	**27.3 (+11.3)**	75.7 (+0.1)	**80.9 (+2.7)**	**33.6 (+1.6)**	35.4 (+0.8)	**+4.1**
LLaDA-MoE	LLaDA-MoE-Instruct	16.4 (+0.0)	27.0 (+0.0)	70.2 (+0.0)	71.0 (+0.0)	30.1 (+0.0)	42.2 (+0.0)	+0.0
	Published labeled-RL references							
	Diffu-GRPO	50.1	48.2	74.2	73.7	38.1	43.4	–
	wd1	56.6	58.7	76.2	76.4	39.8	43.4	–
	SAPO	54.2	54.8	75.2	76.8	38.6	40.4	–
	GDPO	58.1	53.4	74.8	75.2	38.4	41.2	–
	TraceRL	54.2	49.1	75.4	76.1	40.0	44.1	–
	Label-free test-time adaptation							
	CSA-TTRL	17.2 (+0.8)	30.1 (+3.1)	**72.3 (+2.1)**	73.9 (+2.9)	**33.6 (+3.5)**	44.8 (+2.6)	+2.5
	Geo.-LiM	**21.5 (+5.1)**	**33.6 (+6.6)**	71.9 (+1.7)	**76.6 (+5.6)**	33.0 (+2.9)	**45.4 (+3.2)**	**+4.2**

Table 2: **Label-free LLaDA-family code.** The first row is the unadapted backbone. Final pass@1 (Δ, pp) is versus that row. Avg. Δ is the unweighted mean across the four displayed endpoints.

		HumanEval		MBPP		Avg. Δ
Model	Method	256	512	256	512	(pp)
LLaDA	LLaDA-8B-Instruct	25.0 (+0.0)	34.1 (+0.0)	38.9 (+0.0)	40.4 (+0.0)	+0.0
	CSA-TTRL	26.8 (+1.8)	35.4 (+1.3)	38.2 (-0.7)	42.9 (+2.5)	+1.2
	Geo.-LiM	**28.0 (+3.0)**	**36.6 (+2.5)**	**39.6 (+0.7)**	**43.3 (+2.9)**	**+2.3**
LLaDA-MoE	LLaDA-MoE-Instruct	40.2 (+0.0)	48.2 (+0.0)	49.9 (+0.0)	57.1 (+0.0)	+0.0
	CSA-TTRL	40.9 (+0.7)	51.2 (+3.0)	52.5 (+2.6)	53.4 (-3.7)	+0.7
	Geo.-LiM	**41.5 (+1.3)**	**51.8 (+3.6)**	**53.6 (+3.7)**	**56.7 (-0.4)**	**+2.1**

report paired bootstrap intervals over example IDs; those intervals do not measure variation across adaptation seeds.

4.2 Overall Effectiveness

4.2.1 Math benchmarks

Geo.-LiM improves all twelve LLaDA-family math endpoints over the backbone (Table 1). Mean gains are $+4.1\,pp$ on LLaDA and $+4.2\,pp$ on LLaDA-MoE, against CSA-TTRL's $+2.8\,pp$ and $+2.5\,pp$. The largest movements are on Countdown. LLaDA Countdown-512 rises by $+11.3\,pp$, against $+1.2\,pp$ for CSA-TTRL, and the two LLaDA-MoE Countdown lengths rise by $+5.1$ and $+6.6\,pp$. GSM8K and MATH500 move by less, and the advantage over CSA-TTRL is not uniform: CSA-TTRL remains ahead on LLaDA MATH500-512 ($+4.0$ versus $+0.8\,pp$) and on both GSM8K-256 cells. The labeled-RL rows are published accuracies under supervised post-training. Deltas in the table remain relative to the backbone. On LLaDA-MoE MATH500-512, label-free Geo.-LiM reaches 45.4, above every labeled-RL entry in that column, the next being TraceRL at 44.1. On LLaDA-MoE GSM8K-512, 76.6 exceeds Diffu-GRPO, GDPO, TraceRL, and wd1, and sits 0.2 below SAPO.

4.2.2 Code benchmarks

Geo.-LiM has higher final pass@1 than CSA-TTRL in all eight code cells (Table 2), raising mean gains from $+1.2$ to $+2.3\,pp$ on LLaDA and from $+0.7$ to $+2.1\,pp$ on LLaDA-MoE. Seven effects are positive. On LLaDA-MoE MBPP-512, Geo.-LiM is $-0.4\,pp$ and CSA-TTRL is $-3.7\,pp$; on LLaDA MBPP-256 the same contrast is $+0.7$ versus $-0.7\,pp$.

Table 3: **Dream-7B math.** Final (Δ, pp) versus the unadapted backbone. Avg. Δ is the unweighted mean of the six endpoints.

	Countdown		GSM8K		MATH500		Avg. Δ
Method	256	512	256	512	256	512	(pp)
Dream-7B-Instruct	7.8 (+0.0)	13.7 (+0.0)	78.9 (+0.0)	77.6 (+0.0)	38.2 (+0.0)	38.8 (+0.0)	+0.0
CSA-TTRL	**13.7 (+5.9)**	**16.8 (+3.1)**	78.0 (-0.9)	77.8 (+0.2)	36.4 (-1.8)	38.8 (+0.0)	+1.1
Geo.-LiM	11.3 (+3.5)	16.4 (+2.7)	**79.4 (+0.5)**	**79.4 (+1.8)**	**37.8 (-0.4)**	**39.0 (+0.2)**	**+1.4**

Table 4: **Dream-7B code.** Final pass@1 (Δ, pp) versus the unadapted backbone.

	HumanEval		MBPP		Avg. Δ
Method	256	512	256	512	(pp)
Dream-7B-Instruct	34.8 (+0.0)	42.7 (+0.0)	49.7 (+0.0)	48.7 (+0.0)	+0.0
CSA-TTRL	**37.8 (+3.0)**	41.5 (-1.2)	49.9 (+0.2)	48.5 (-0.2)	+0.5
Geo.-LiM	36.6 (+1.8)	**42.7 (+0.0)**	**50.9 (+1.2)**	**49.2 (+0.5)**	**+0.9**

4.2.3 Adaptation on Dream-7B

On Dream-7B the same protocol yields a smaller math gain than on the LLaDA family: $+1.4\,pp$ for Geo.-LiM and $+1.1\,pp$ for CSA-TTRL (Table 3). CSA-TTRL leads on both Countdown lengths. Geo.-LiM leads on both GSM8K lengths, and its MATH500-256 effect is $-0.4\,pp$.

Dream code is smaller still: $+0.9\,pp$ for Geo.-LiM and $+0.5\,pp$ for CSA-TTRL (Table 4). CSA-TTRL leads on HumanEval-256; Geo.-LiM leads on the other three code cells. Across all 30 cells, each a single adaptation run, Geo.-LiM improves 27 endpoints over the backbone and exceeds CSA-TTRL on 23. Several of those margins are below one point.

4.3 Selective Adaptation with Fewer Updates

Across all 30 endpoints, Geo.-LiM admits fewer prompts and takes fewer optimizer-bearing steps than CSA-TTRL. Under the same eight-rollout evidence budget, admitted prompts fall from 4,348 to 1,121 (-74.2%) and optimizer-bearing minibatches from 2,713 to 966 (-64.4%) (Appendix Table 9). The reduction is in replay, backpropagation, and optimizer steps, not in evidence generation. Dream HumanEval-512 is the one empty gate, so its flat score is no adaptation. The two declines are updates: Dream MATH500-256 ($-0.4\,pp$) and LLaDA-MoE MBPP-512 ($-0.4\,pp$). Figure 3 shows where those admissions fall on the three matched streams. Geo.-LiM admits 74–83% fewer prompts on each of them.

Figure 4 partitions admitted proposals after adaptation. Labels are joined only for this audit. A redundant proposal is an admitted update that does not raise accuracy. On LLaDA MATH500-256 that count is 52 of 119 for CSA-TTRL and 52 of 127 for all-vote; on Dream GSM8K-256 it is 425 of 521 and 442 of 526. Geo.-LiM has none on either stream: 0 of 31 and 0 of 87 (Appendix Table 10). Panels (c, d) are the paired corrections and regressions behind the Geo.-LiM endpoint deltas.

Appendix Figure 8 plots the same three streams as paired gain against eligible-prompt count: 119 to 31, 521 to 87, and 12 to 3, with the paired gain rising in each panel. Section 4.4 separates this pattern from the claim that every reduction in K raises accuracy.

On LLaDA MATH500-256 and Dream GSM8K-256, the rolling eligibility median falls from 36.0% to 4.0% and from about 74% to 7.6% (Appendix Figure 6). Conditional on acceptance, the reverse-KL proxy is smaller on the LLaDA streams and slightly larger on Dream GSM8K, so the accuracy gaps do not require uniformly smaller per-update drift.

4.4 What Drives the Gains?

The main grid compares complete methods. The three streams named in Section 4.1 then separate three questions: whether the geometry source matters, whether fewer updates alone explain the gains,

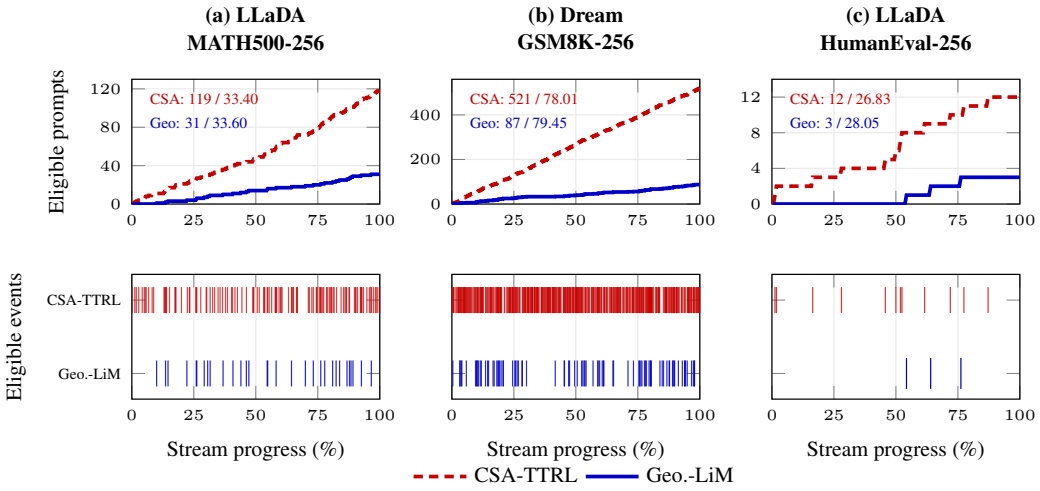


Figure 3: **Matched update dynamics.** CSA-TTRL and Geo.-LiM use the same prompt streams and eight-rollout evidence budget. Top: cumulative admitted-prompt counts K, with annotations reporting K/final score (%). Bottom: admission positions along each stream. Geo.-LiM admits 74–83% fewer prompts on these three streams.

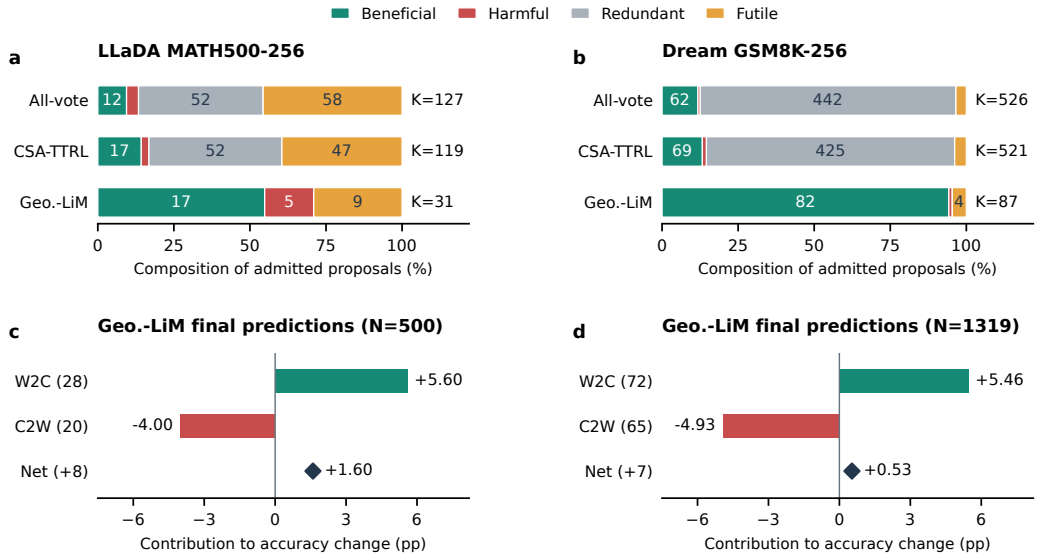


Figure 4: **Broad gates admit many redundant proposals. (a, b)** Post-hoc partition of admitted proposals after benchmark labels are joined. Redundant means the admitted update does not raise accuracy: 52/119 and 425/521 for CSA-TTRL, against 0/31 and 0/87 for Geo.-LiM. **(c, d)** Geo.-LiM final-prediction corrections and regressions on the same two streams.

and which selector components contribute. We vary only the stated factor. When geometry changes, eligibility and update counts are allowed to change with it.

Geometry source. Table 5 keeps the reference-novelty gate, positive-only replay, and eight-rollout budget fixed, and changes only how the evidence geometries are built. The sampled pair has the largest gain on all three streams. Single-G reverses the HumanEval effect; Fixed-2G admits no prompt on that stream.

Table 5: **Geometry-source ablation.** The reference-novelty gate, positive-only replay, and eight-rollout budget are fixed. Resulting eligibility and optimizer-bearing minibatches U (in parentheses) are allowed to change. Cells are backbone-relative gain in pp. The sampled pair is the main configuration; $U = 0$ denotes no adaptation.

Endpoint	Sampled pair	Single-G	Fixed-2G
Dream GSM8K-256	**+0.5** (78)	+0.2 (87)	+0.0 (71)
LLaDA MATH500-256	**+1.6** (29)	+1.2 (20)	+1.2 (24)
LLaDA HumanEval-256	**+3.0** (3)	−0.6 (6)	+0.0 (0)

Fewer updates versus which updates. The rightmost columns of Table 6 match both of Geo.-LiM's counts K and U. Matching the counts does not reproduce the geometry rule, and no column leads on every stream.

Selector components. The left block of the same table varies the rule rather than matching counts. All-vote and Band keep most prompts; Band+N adds reference novelty and cuts K sharply; Geo.-S changes the replay sign; Geo.-LiM is strict pooled support, novelty, and positive-only replay. These columns are not count-matched to one another.

Table 6: **Selector components and count-matched controls.** Cells show backbone-relative gain in pp and (K/U). The left block changes the selector: Band uses 5–7$/8$ agreement, Band+N adds reference novelty, and Geo.-S uses signed replay. The right block, Unif.-K and Conf.-K, matches both Geo.-LiM counts. CSA matches the eight-rollout evidence budget but not K/U. Avg. Δ averages only these three streams.

	Component configurations					Baseline	Matched K/U	
Endpoint	All-vote	Band	Band+N	Geo.-S	Geo.-LiM	CSA	Unif.-K	Conf.-K
LLaDA MATH500-256	+0.2 (127/83)	+1.8 (115/81)	+2.4 (23/20)	+2.0 (30/28)	+1.6 (31/29)	+1.4 (119/80)	+0.8 (31/29)	+3.2 (31/29)
Dream GSM8K-256	-0.9 (526/282)	-0.2 (524/287)	+0.1 (73/67)	-0.3 (82/74)	+0.5 (87/78)	-0.9 (521/287)	+0.6 (87/78)	-1.4 (87/78)
LLaDA HumanEval-256	+0.0 (15/12)	+0.0 (11/8)	-0.6 (2/2)	+1.2 (3/3)	+3.0 (3/3)	+1.8 (12/10)	+0.6 (3/3)	-0.6 (3/3)
Avg. Δ	-0.2	+0.5	+0.6	+1.0	**+1.7**	+0.8	+0.7	+0.4

On LLaDA HumanEval-256, All-vote admits 15 prompts and finishes at $+0.0\,pp$, while Geo.-LiM admits 3 and finishes at $+3.0\,pp$ (Table 6). Matching that count is not enough: Uniform-K and Confidence-K use the same K/U and finish at $+0.6$ and $-0.6\,pp$. The other direction also occurs. On LLaDA MATH500-256, Confidence-K reaches $+3.2\,pp$ against Geo.-LiM's $+1.6\,pp$. A smaller K and the geometry rule are separable, and neither wins every stream.

The same streams also separate corrections from regressions. On Dream GSM8K-256, crediting every voted prompt yields 79 wrong-to-correct changes and 91 correct-to-wrong changes ($-0.9\,pp$); Geo.-LiM yields 72 and 65 ($+0.5\,pp$). On LLaDA HumanEval-256, the three admitted prompts correspond to 7 corrections and 2 regressions. The paired gain is this net. On LLaDA MATH500-256 the same rule yields 28 corrections and 20 regressions. The remaining rows are in Appendix Table 13. Rollout-budget and 4×2 allocation checks, which change more than one design choice at a time, are reported in Appendices B and D.

5 Conclusion

Geo.-LiM uses denoising geometry to determine which self-generated targets receive persistent test-time updates. Across the evaluated grid, it improves 27/30 endpoints over the backbone and exceeds CSA-TTRL on 23/30, with fewer admitted prompts and optimizer-bearing steps in every cell. The practical contribution is an explicit update-eligibility rule that combines cross-geometry support with reference-relative novelty under a fixed evidence budget. Selection controls and prediction ledgers expose complementary dimensions of adaptation: which targets are learned from and how the resulting parameter changes affect other predictions. Reducing collateral regressions while retaining the gains of selective adaptation remains an important direction.

AI USE STATEMENT

Generative AI tools assisted literature discovery, code implementation and inspection, experiment automation, auditing of recorded results, figure preparation, and manuscript drafting and editing. The authors manually verified all AI-assisted claims, citations, code, experimental records, and figures. All reported experiments were executed and validated by the authors, who take responsibility for the complete content of this submission.

ETHICS STATEMENT

This work uses publicly available model checkpoints and standard reasoning and code benchmarks; it involves no human participants or private data. Persistent test-time adaptation can amplify erroneous self-generated signals, so we report negative transfers and design rejection as a no-update event. Deployment in high-stakes settings would require task-specific oversight beyond the benchmarks studied here.

REPRODUCIBILITY STATEMENT

Section 3 specifies the admission rule and optimization objective, Section 4 defines the matched evaluation protocol, and Appendix A provides the online procedure and endpoint configurations. The supplementary artifact contains per-endpoint manifests, reference and final predictions, adapter checks, update counts, and W2C/C2W ledgers sufficient to reconstruct the reported paired metrics.

REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Jinbin Bai, Yixuan Li, Yuchen Zhu, Yi Xin, Qingyu Shi, Aosong Feng, Xiaohong Liu, Molei Tao, Jianru Xue, Xiangtai Li, and Ming-Hsuan Yang. Prism: Efficient test-time scaling via hierarchical search and self-verification for discrete diffusion language models. *arXiv preprint arXiv:2602.01842*, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Tianlang Chen, Minkai Xu, Jure Leskovec, and Stefano Ermon. Rfg: Test-time scaling for diffusion large language model reasoning with reward-free guidance. *arXiv preprint arXiv:2509.25604*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Meihua Dang, Jiaqi Han, Minkai Xu, Kai Xu, Akash Srivastava, and Stefano Ermon. Inference-time scaling of diffusion language models via trajectory refinement. *arXiv preprint arXiv:2507.08390*, 2025.

Bodong Du, Xuanqi Huang, and Xiaomeng Li. Distribution-aware reward estimation for test-time reinforcement learning. *arXiv preprint arXiv:2601.21804*, 2026.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, Xiusi Chen, Youbang Sun, Xingtai Lv, Xuekai Zhu, Li Sheng, Ran Li, Huan-ang Gao, Yuchen Zhang, Bowen Zhou, Zhiyuan Liu, and Ning Ding. How far can unsupervised rlvr scale llm training? In *International Conference on Learning Representations*, volume 2026, pp. 14823–14865, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jihoon Lee, Hoyeon Moon, Kevin Zhai, Arun Chithanar, Anit Kumar Sahu, Soummya Kar, Chul Lee, Souradip Chakraborty, and Amrit Bedi. Test-time scaling in diffusion llms via hidden semi-autoregressive experts. In *International Conference on Learning Representations*, volume 2026, pp. 46880–46899, 2026.

Jiahui Li, Jianfeng Shan, Wenpei Chen, Shunyu Wu, Jian Lou, Wenjie Feng, Dan Li, and See-Kiong Ng. Exploiting verification-generation gap: Test-time reinforcement learning with confidence-conditioned verification. *arXiv preprint arXiv:2606.03608*, 2026a.

Xiuyu Li, Jinkai Zhang, Mingyang Yi, Yu Li, Longqiang Wang, Yue Wang, and Ju Fan. Ets: Energy-guided test-time scaling for training-free rl alignment. *arXiv preprint arXiv:2601.21484*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Roy Miles, Aysim Toker, Andreea-Maria Oncescu, Songcen Xu, Jiankang Deng, and Ismail Elezi. Test-time scaling with diffusion language models via reward-guided stitching. *arXiv preprint arXiv:2602.22871*, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Zheyu Fu, Liancheng Fang, Minghua He, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. d-treerpo: Towards more reliable policy optimization for diffusion language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21801–21822, 2026.

Kevin Rojas, Jiahe Lin, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, Molei Tao, and Wei Deng. Improving reasoning for diffusion language models via group diffusion policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 1938–1973, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Xiaohang Tang, Rares Dolga, Sangwoong Yoon, and Ilija Bogunovic. wd1: Weighted policy optimization for reasoning in diffusion language models. In *International Conference on Learning Representations*, volume 2026, pp. 150357–150385, 2026.

Lipeng Wan, Junjie Ma, Jianhui Gu, Zeyang Liu, Xuyang Lu, and Xuguang Lan. Geoblock: Inferring block granularity from dependency geometry in diffusion language models. *arXiv preprint arXiv:2603.26675*, 2026.

Wen Wang, Bozhen Fang, Chenchen Jing, Yongliang Shen, Yangyi Shen, Qiuyu Wang, Hao Ouyang, Hao Chen, and Chunhua Shen. Time is a feature: Exploiting temporal dynamics in diffusion language models. In *International Conference on Learning Representations*, volume 2026, pp. 72909–72935, 2026a.

Yinjie Wang, Ling Yang, Bowen Li, Ye Tian, Ke Shen, and Mengdi Wang. Revolutionizing reinforcement learning framework for diffusion large language models. In *International Conference on Learning Representations*, volume 2026, pp. 101476–101501, 2026b.

Zhiping Xiao, Yusheng Zhao, Qixin Zhang, Jiaye Xie, Wanjia Zhao, Weizhi Zhang, Xiao Luo, Philip S. Yu, and Ming Zhang. Sample lottery: Unsupervised discovery of critical instances for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 8919–8949, 2026.

Shaoan Xie, Lingjing Kong, Xiangchen Song, Xinshuai Dong, Guangyi Chen, Eric P. Xing, and Kun Zhang. Advancing reasoning in diffusion language models with denoising process rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42703–42720, 2026.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Yongcan Yu, Lingxiao He, Jian Liang, Kuangpu Guo, Meng Wang, Qianlong Xie, Xingxing Wang, and Ran He. Understanding and mitigating spurious signal amplification in test-time reinforcement learning for math reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37424–37436, 2026.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Xuan Li, Xiao Feng, Jiangchao Yao, Bo Han, et al. Co-rewarding: Stable self-supervised rl for eliciting reasoning in large language models. In *International Conference on Learning Representations*, volume 2026, pp. 82520–82558, 2026.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:56729–56762, 2026.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, et al. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*, 2025.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11425–11460, 2026.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

Appendix A specifies implementation and reproducibility. Appendices B–D report supplementary results, selection controls, and an alternative geometry protocol. Appendix E analyzes prediction changes and adaptation dynamics; Appendix F discusses what these comparisons do and do not establish.

APPENDIX CONTENTS

A Online Procedure and Configuration **15**

A.1 Reproducibility contract 15

B Additional Main-Grid Results **16**

B.1 Rollout-budget sensitivity 16

C Update Counts and Selection Controls **17**

C.1 Admitted prompts and optimizer steps 17

C.2 Exact selector and count-matched results 18

C.3 Post-hoc selector oracle audit 18

C.4 Component configuration comparisons 18

D Alternative Four-Geometry Protocol **20**

E Prediction Changes and Adaptation Diagnostics **21**

E.1 Eligibility and policy-drift diagnostics 21

E.2 Selector dynamics on a sparse code stream 21

E.3 Representative transition ledger 22

E.4 Endpoint-level efficiency scatter 23

E.5 Eligible-prompt gradient diagnostics 23

E.6 Cross-task capability retention 23

F Interpretation and Scope of Evidence **24**

A Online Procedure and Configuration

Executable counterpart of Section 3: the adaptation procedure, the geometry/eligibility matrix, optimization constants, and the checks a scored endpoint must pass.

Sequential adaptation. For each x_t: (1) temporarily restore the initial adapter and decode a_{ref}; (2) under that same fixed selection anchor, deterministically select a geometry pair and sample four matched-NFE trajectories per view; (3) apply the task-common parser and prompt-visible validity rules to the readout and trajectories, reject ties/failures, and count pooled support; (4) apply the strict-majority and reference-novelty gates; (5) restore the persistent current adapter and, if eligible, replay one supporting trace per view; and (6) take one synchronized clipped LoRA step if any rank contributes a signal, then carry the adapter forward. The final score uses a separate deterministic decode. Table 7 lists the geometry set, repeat count, eligibility rule, and credit assignment for every variant compared in the paper.

Table 7: Geometry and eligibility configurations. Every listed main configuration contains eight trajectories; b is absolute and r is length-normalized. The Geo.-LiM pair matches the evidence budget and samples prompt-indexed pairs reproducibly over the stream.

Variant	Geometry set	Repeats	Eligibility	Credit
Direct Vote	one b	8	pooled strict vote	supporting mode
CSA-TTRL	$b \in \{16, 64\}$	4/view	cross-schedule agreement	signed replay
Fixed-schedule selectors	$b \in \{16, 64\}$	4/view	All-vote / Band / Band+N	branch-specific
Geo.-LiM Pair	2 of 4 pre-specified r	4/view	strict vote + novelty	positive-only

Optimization and artifacts. Main runs use AdamW with zero weight decay, LoRA rank/scale 16/32 and dropout 0, fixed-reference KL coefficient $\beta = 0.01$, gradient cap 1.0, and four synchronous GPUs. Evidence is sampled at temperature 1.0; the independent reference/final readouts use temperature 0. The canonical block schedule has $b = 16$ and eight reveal steps per block, hence $F = L/2$ NFE per trajectory: 128 at $L = 256$ and 256 at $L = 512$. CSA-TTRL and Geo.-LiM use the same learning rate, 10^{-6}, at every endpoint. Within each adaptation run, the seed, prompt order, and stopping rule remain fixed; intermediate test labels are not used for checkpoint selection. The released configurations record this shared learning rate with the seed, prompt order, and evaluator for every endpoint.

The anonymous artifact uses stable endpoint IDs and records model family, task, length, method, metric, reference/final values, paired differences, status, and the shared protocol. It excludes storage paths, scheduler job names, and internal run aliases. Geometry sampling and reference/final readouts are reproducible functions of the pre-specified seed and stable example ID. Training is seeded but not claimed to be bitwise identical across accelerator types or CUDA/FlashAttention kernels; hardware and software versions are therefore part of the artifact protocol.

A.1 Reproducibility contract

Every scored endpoint must contain: (i) a run manifest with model family, prompt profile, method ID, relative geometry, seed, learning rate, generation length, NFE, and evaluator; (ii) complete reference and final predictions with the same unique IDs; (iii) a saved nonempty adapter and training summary; (iv) a paired evaluation whose responses match the source files; and (v) W2C/C2W counts reproducible from per-example correctness. Training-stream or online accuracy is never substituted for the independent final decode. These checks prevent partial streams, best-checkpoint selection, or mismatched evaluators from entering the main result tables.

B ADDITIONAL MAIN-GRID RESULTS

These tables supplement the Dream results in Section 4.2 and report rollout-budget sensitivity. Update counts and selection controls appear in Appendix C; stream-level diagnostics appear in Appendix E.

Dream math and code are main-text Tables 3 and 4.

B.1 ROLLOUT-BUDGET SENSITIVITY

Table 8 records rollout-budget controls, complementing the geometry-source ablation in main-text Table 5.

Table 8: **Rollout-budget sensitivity.** Backbone and Final scores (adaptation effect in parentheses, pp). B is the number of evidence rollouts per prompt, split across two views.

		CSA-TTRL			Geo.-LiM		
Task	Base	$B = 4$	$B = 8$	$B = 16$	$B = 4$	$B = 8$	$B = 16$
GSM8K-256	75.6	75.8 (+0.2)	75.9 (+0.3)	76.0 (+0.4)	76.2 (+0.6)	75.7 (+0.1)	76.3 (+0.7)
MATH500-256	32.0	34.4 (+2.4)	33.4 (+1.4)	33.0 (+1.0)	33.4 (+1.4)	33.6 (+1.6)	34.0 (+2.0)
HumanEval-256	25.0	25.6 (+0.6)	26.8 (+1.8)	25.6 (+0.6)	26.2 (+1.2)	28.0 (+3.0)	25.0 (+0.0)

C Update Counts and Selection Controls

The complete K/U ledger records update counts across the 30 endpoints. Count-matched controls, post-hoc proposal audits, and component contrasts then distinguish update quantity, target composition, and the factors varied by each comparison.

C.1 Admitted prompts and optimizer steps

Table 9 reports, for every cell of the grid, how many prompts each method admitted (K) and how many distinct optimizer-bearing minibatches it executed (U). Both methods draw the same eight NFE-matched rollouts per prompt, so these counts are the part of the cost that the admission rule controls. Because four distributed ranks can contribute several admitted prompts to one synchronized minibatch, U may be smaller than K.

Table 9: **Admitted prompts and optimizer-bearing minibatches on the complete 30-cell grid.** N is the stream length, K the admitted prompts, and U the distinct optimizer-bearing minibatches; Δ gives the reduction from CSA-TTRL to Geo.-LiM. Geo.-LiM is lower on both counts in 30 of 30 cells. Totals are sums of raw counts rather than averages of the per-cell percentages.

				Admitted prompts K			Optimizer steps U		
Model	Benchmark	L	N	CSA	Geo.	Δ (%)	CSA	Geo.	Δ (%)
LLaDA-8B	Countdown	256	256	87	40	54.0	52	30	42.3
	Countdown	512	256	67	53	20.9	46	38	17.4
	GSM8K	256	1,319	380	110	71.1	245	100	59.2
	GSM8K	512	1,319	344	90	73.8	234	79	66.2
	MATH500	256	500	119	31	73.9	80	29	63.8
	MATH500	512	500	118	32	72.9	85	27	68.2
	HumanEval	256	164	12	3	75.0	10	3	70.0
	HumanEval	512	164	12	5	58.3	11	5	54.5
	MBPP	256	427	61	12	80.3	52	11	78.8
	MBPP	512	427	66	15	77.3	48	14	70.8
	Subtotal		5,332	1,266	391	**69.1**	863	336	**61.1**
LLaDA-MoE	Countdown	256	256	102	20	80.4	55	18	67.3
	Countdown	512	256	134	83	38.1	60	51	15.0
	GSM8K	256	1,319	427	120	71.9	247	101	59.1
	GSM8K	512	1,319	475	94	80.2	276	84	69.6
	MATH500	256	500	146	36	75.3	92	31	66.3
	MATH500	512	500	141	33	76.6	92	30	67.4
	HumanEval	256	164	21	3	85.7	17	3	82.4
	HumanEval	512	164	21	6	71.4	20	6	70.0
	MBPP	256	427	52	18	65.4	40	18	55.0
	MBPP	512	427	55	15	72.7	48	15	68.8
	Subtotal		5,332	1,574	428	**72.8**	947	357	**62.3**
Dream-7B	Countdown	256	256	33	25	24.2	26	19	26.9
	Countdown	512	256	49	25	49.0	36	22	38.9
	GSM8K	256	1,319	521	87	83.3	287	78	72.8
	GSM8K	512	1,319	540	99	81.7	282	92	67.4
	MATH500	256	500	112	19	83.0	79	17	78.5
	MATH500	512	500	116	22	81.0	78	21	73.1
	HumanEval	256	164	6	1	83.3	6	1	83.3
	HumanEval	512	164	7	0	100.0	7	0	100.0
	MBPP	256	427	63	19	69.8	52	18	65.4
	MBPP	512	427	61	5	91.8	50	5	90.0
	Subtotal		5,332	1,508	302	**80.0**	903	273	**69.8**
All 30 endpoints			15,996	4,348	1,121	**74.2**	2,713	966	**64.4**

Admission rates K/N span 3.7–52.3% for CSA-TTRL and 0.0–32.4% for Geo.-LiM. The single 100% reduction is Dream HumanEval-512, where the gate admits no prompt at all: that endpoint's neutral result is a no-adaptation outcome, so only two of the 30 cells reflect an update that made the model worse.

C.2 EXACT SELECTOR AND COUNT-MATCHED RESULTS

Main-text Table 6 reports the complete component configurations and count-matched controls. Figure 5 visualizes these configurations and controls; the following audits examine proposal composition and configuration differences.

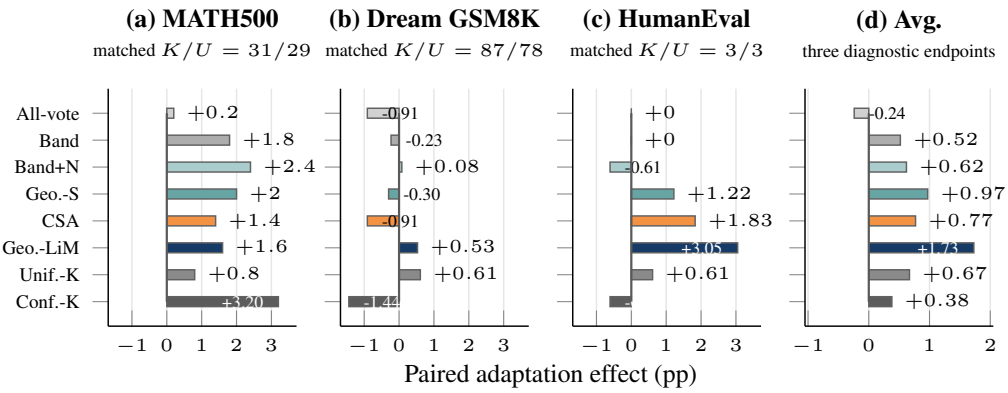


Figure 5: **Selector configurations and count-matched controls.** Bars show paired accuracy or pass@1 adaptation effect. The subtitles report the pre-specified Geo.-LiM count (K/U); Uniform-K and Confidence-K exactly match both counts, stream, positive replay, and rollout budget. CSA-TTRL matches the eight-rollout evidence budget but not K/U. Panel (d) is a descriptive average over only these three diagnostic endpoints.

C.3 POST-HOC SELECTOR ORACLE AUDIT

Table 10: **Post-hoc selector oracle audit.** Eligible pseudo-labels are joined with benchmark ground truth only after adaptation. The four count columns partition the eligible set. "Post-hoc pseudo-label acc." is the fraction of selected proposals whose pseudo-label matches that ground truth; it is a selector diagnostic, not the final benchmark accuracy. Redundant counts admitted updates that do not raise accuracy. "Net" is beneficial minus harmful and is distinct from final W2C−C2W.

Endpoint	Selector	Eligible	Post-hoc acc. (%)	Beneficial	Harmful	Redundant/Futile	Net
LLaDA MATH500-256	All-vote	127	50.4	12	5	52/58	7
	Band	115	55.7	9	3	55/48	6
	Band+N	23	39.1	9	2	0/12	7
	CSA-TTRL	119	58.0	17	3	52/47	14
	Geo.-LiM	31	54.8	17	5	0/9	**12**
Dream GSM8K-256	All-vote	526	95.8	62	4	442/18	58
	Band	524	95.6	64	6	437/17	58
	Band+N	73	87.7	64	6	0/3	58
	CSA-TTRL	521	94.8	69	7	425/20	62
	Geo.-LiM	87	94.3	82	1	0/4	**81**

Table 10 reports this audit for the representative streams. The machine-readable audit for the 30 main cells together with 16 diagnostic/control endpoints additionally records paired bootstrap intervals, optimizer-bearing minibatches, and all 5,080 selected rows. Labels are joined only after completed adaptation and are never used as an online selection or model-selection signal.

C.4 COMPONENT CONFIGURATION COMPARISONS

Table 11 reorganizes the existing paired runs by the design choices represented. All-vote→Band changes the reliability rule, and Band→Band+N adds reference-relative novelty. The Geo signed and positive configurations differ in credit assignment and recorded K/U, so their scores are configuration

comparisons. Band+N→Geo jointly changes the evidence distribution and reliability rule, and is not an isolated geometry-only contrast.

Table 11: **Existing component matrix (paired Δ, pp).** All-vote/Band/ Band+N uses fixed $4 + 4$ schedules; Band denotes the pre-specified 5–7 agreement range. Geo uses a sampled pair with strict pooled majority. The positive-replay configuration has larger gains than the signed configuration on Dream GSM8K and LLaDA HumanEval, but not LLaDA MATH.

Endpoint	All-vote	Band	Band+N	Geo signed	Geo positive
LLaDA MATH500-256	+0.2	+1.8	**+2.4**	+2.0	+1.6
Dream GSM8K-256	-0.9	-0.2	+0.1	-0.3	**+0.5**
LLaDA HumanEval-256	+0.0	+0.0	-0.6	+1.2	**+3.0**

D ALTERNATIVE FOUR-GEOMETRY PROTOCOL

Table 12 compares the main protocol with an alternative that allocates the same eight NFE-matched rollouts across all four pre-specified geometries. The main 2×4 design uses four repeats per sampled view and requires reference novelty. The alternative 4×2 design uses two repeats per view and omits novelty; both use pooled strict majority and positive-only replay. This comparison therefore changes geometry breadth, within-view repetition, and eligibility together.

The alternative achieves larger gains on LLaDA MATH500-256, LLaDA HumanEval-256, and Dream GSM8K-256. It also incurs a larger loss on Dream MATH500-256 and produces no net gain on Dream HumanEval-256, where the main protocol improves. These results identify a competitive, task-dependent alternative and do not establish that either allocation is uniformly preferable. The main design retains within-view replication and explicitly excludes reference-confirming targets; its use does not imply that it maximizes accuracy at every endpoint. Isolating the two choices requires comparing both allocations with novelty enabled and disabled, rather than attributing this joint contrast to either factor alone.

Table 12: **Eight-rollout protocol comparison.** Both protocols use NFE-matched rollouts and positive-only replay. The main 2×4 protocol includes reference novelty; the alternative 4×2 protocol does not. Final scores and backbone-relative gains (pp) are shown for all five reported alternative-protocol endpoints, including negative and neutral outcomes.

		Main 2×4 + novelty		Alternative 4×2	
Endpoint	Base	Final	Δ	Final	Δ
LLaDA MATH500-256	32.0	33.6	+1.6	34.8	+2.8
LLaDA HumanEval-256	25.0	28.0	+3.0	28.7	+3.7
Dream GSM8K-256	78.9	79.4	+0.5	81.6	+2.7
Dream MATH500-256	38.2	37.8	-0.4	35.8	-2.4
Dream HumanEval-256	34.8	36.6	+1.8	34.8	+0.0

E PREDICTION CHANGES AND ADAPTATION DIAGNOSTICS

Proposal quality and final prediction changes describe complementary aspects of adaptation. Transition ledgers record corrections and regressions on the target benchmark; cross-task retention evaluates the adapted policy on a held-out task.

Figure 3 in the main text shows cumulative admissions and their positions on the three matched streams.

E.1 ELIGIBILITY AND POLICY-DRIFT DIAGNOSTICS

Building on the event dynamics in Figure 3, this appendix separates gate-eligibility density from the policy-drift signals conditional on an accepted update.

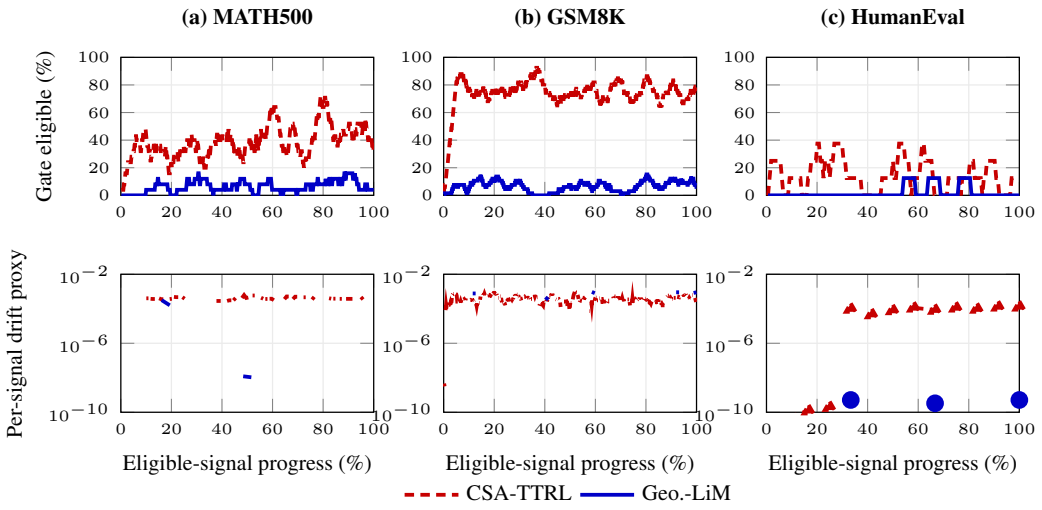


Figure 6: **Stricter gates, task-dependent drift.** Columns are the three matched streams. Top: rolling fraction of prompts passing each method's evidence gate; lower is stricter. Bottom: rolling median absolute reverse-KL proxy on eligible traces. Red dashed curves are CSA-TTRL and blue solid curves are Geo.-LiM. HumanEval uses markers because only 12 and 3 prompt signals are eligible.

Raw policy losses are not overlaid: CSA-TTRL uses a signed objective whereas Geo.-LiM is positive-only, so their loss scales have different meanings.

On LLaDA MATH500-256 and Dream GSM8K-256, the rolling-window median of this gate-eligible fraction drops from 36.0% to 4.0% and from about 74% to 7.6%, respectively. Those medians summarize local eligibility along the stream. They are distinct from the stream-level admission rates K/N on the same endpoints, which fall from $119/500$ (23.8%) to $31/500$ (6.2%) and from $521/1{,}319$ (39.5%) to $87/1{,}319$ (6.6%). Under both summaries, Geo.-LiM demands substantially rarer agreement. Conditional on an accepted trace, its reverse-KL proxy is much smaller on the two LLaDA streams but slightly larger on Dream GSM8K. Thus the consistent effect is *selective eligibility*; uniformly smaller per-update motion is neither observed nor required for the accuracy gains.

E.2 SELECTOR DYNAMICS ON A SPARSE CODE STREAM

Figure 7 expands the component configurations in Table 6 for LLaDA HumanEval-256. All-vote and Band admit 15 and 11 prompts with neutral endpoint effects; Band+N admits two with a negative effect, while Geo.-LiM admits three with a positive effect. The configurations vary multiple factors, so these event patterns do not isolate the effect of novelty alone.

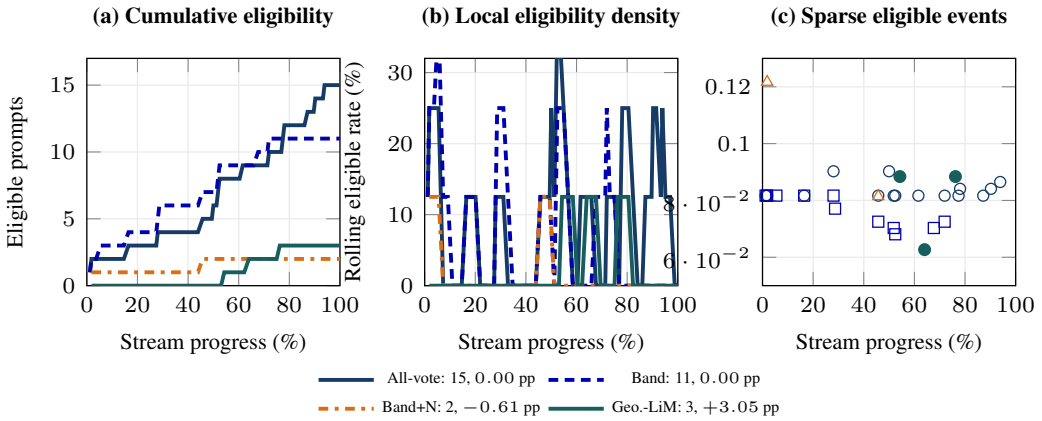


Figure 7: **HumanEval selector dynamics.** The same LLaDA HumanEval-256 stream is shown for All-vote, Band, Band+N, and Geo.-LiM. Left: cumulative eligible prompts. Middle: local eligibility density. Right: gradient norm only at eligible events. Legend entries give admitted-prompt counts and final gains; gradient norms are shown only at eligible events and are not accuracy trajectories.

E.3 Representative transition ledger

Table 13 lists eligible prompts alongside paired wrong-to-correct and correct-to-wrong counts for the representative complete endpoints. The paired gains reflect the net balance of corrections and regressions across evaluated inputs.

Table 13: Eligible prompts and paired transitions for representative complete endpoints. W2C/C2W audit the direction and magnitude of change; each diagnostic Δ is computed from the corresponding paired transition counts.

Model	Task	Method	Eligible	W2C	C2W	Δ
LLaDA	MATH500-256	all	127	23	22	+0.2
LLaDA	MATH500-256	band-5–7	115	25	16	+1.8
LLaDA	MATH500-256	band+novelty	23	27	15	+2.4
LLaDA	MATH500-256	Geo.-LiM	31	28	20	+1.6
Dream	GSM8K-256	all	526	79	91	−0.9
Dream	GSM8K-256	band-5–7	524	83	86	−0.2
Dream	GSM8K-256	band+novelty	73	69	68	+0.1
Dream	GSM8K-256	Geo.-LiM	87	72	65	+0.5
LLaDA-MoE	Countdown-256	Geo.-LiM	20	38	25	+5.1
LLaDA	HumanEval-256	Geo.-LiM	3	7	2	+3.0

E.4 Endpoint-level efficiency scatter

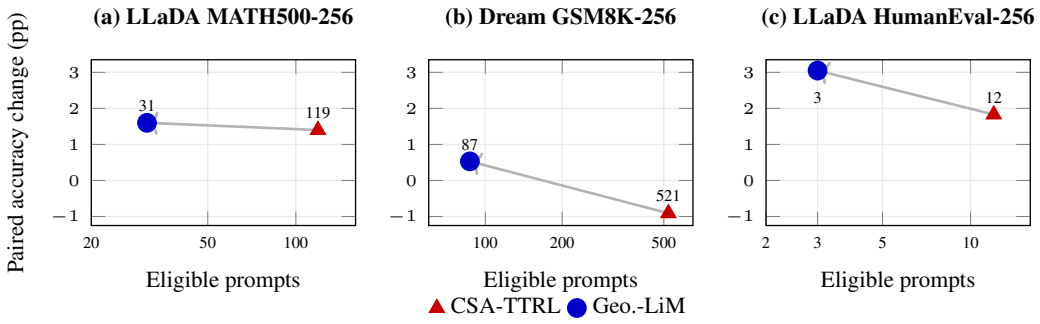


Figure 8: **Fewer admitted prompts, larger paired gain.** Each panel is one matched stream. The arrow runs from CSA-TTRL to Geo.-LiM. Eligible prompts fall from 119 to 31, from 521 to 87, and from 12 to 3, and the paired gain rises in each panel.

E.5 Eligible-prompt gradient diagnostics

Figure 9 records gradient magnitude at the accepted events themselves, which is what makes the sparsity of the Geo.-LiM update stream visible; we report it as a diagnostic of update frequency, not as a performance metric.

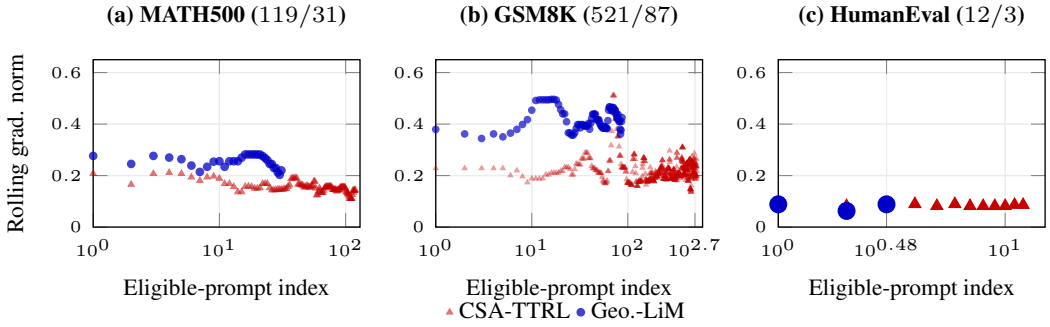


Figure 9: Gradient diagnostics for eligible prompts in the matched streams in Figure 3. Each marker is the rolling median gradient norm over at most the ten most recent eligible prompt signals; no line interpolation is used. Titles report the CSA-TTRL/Geo.-LiM marker counts. In particular, the HumanEval panel contains all 12 red and three blue accepted events rather than implying a continuous three-point training curve.

E.6 Cross-task capability retention

Table 14: **Pre-specified cross-task retention ledger.** Source-task adapters are evaluated without further updates on a full held-out target set. All three pre-specified directions are shown. Geo.–CSA is the paired target accuracy difference between Geo.-LiM and CSA-TTRL; it measures relative retention, not source-task improvement.

Model	Source → target	Base	CSA-TTRL	Geo.-LiM	Geo.–CSA (pp)
LLaDA-8B	MATH500 → GSM8K	75.6	76.2	75.9	−0.3
Dream-7B	GSM8K → MATH500	38.2	37.6	37.0	−0.6
LLaDA-8B	HumanEval → MBPP	38.9	38.9	39.1	+0.2

F INTERPRETATION AND SCOPE OF EVIDENCE

Denoising geometry as a nuisance variable. A block geometry changes how a dLLM exposes conditional structure but should not change which final answer is useful. Agreement among repeats from one geometry can reflect a schedule-specific basin; pooling a seeded pair estimates answer support under a controlled change in factorization. The main 2×4 estimator uses a prompt-indexed pair and spends the remaining budget on within-view repeats. Hashing assigns each of the 15,996 main-grid prompts to one of the six unordered pairs, so the estimator mixes factorization across $\mathcal{G}_0$ rather than reusing one pair for the whole stream.

Reliability, learning value, and overwrite risk. Geometry consensus supplies support for a target, while reference novelty identifies an answer-level change. These signals jointly determine eligibility; neither measures the utility of the resulting parameter update. Positive-only replay avoids explicitly suppressing alternative outputs, but each accepted update changes the shared adapter used across prompts. Endpoint accuracy, eligible counts, and W2C/C2W therefore describe distinct aspects of adaptation.

What is and is not identified. The component configurations vary both update identity and count. CSA-TTRL controls rollout NFE, while the completed Uniform-K and Confidence-K groups also match eligible prompts and optimizer-bearing minibatches. Their heterogeneous results support geometry-selective novelty as a useful practical gate on several endpoints, but not universal ranking superiority over every equally sparse subset. Likewise, supervised results optimize against task rewards on a separate distribution and serve only as an absolute reference.