TRACESTENCIL: SIGNED TEMPORAL NEIGHBORHOODS FOR TOKEN-LEVEL POLICY OPTIMIZATION IN RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards supplies sequence-level feedback while updating token-level probabilities, leaving the choice of which neighboring positions should influence each update largely implicit. We introduce TRACESTENCIL, a token-level policy-optimization objective that represents this choice as a signed temporal neighborhood. Over the selected past and/or future offsets, TRACESTENCIL aggregates likelihood ratios in separately clipped directional branches and applies their detached product to the current-token score, making direction, radius, boundary exposure, and gradient flow explicit. We derive exact structural relations among these quantities and identify the conditions under which the objective reduces to a forward-only trace estimator. Controlled multi-run experiments on mathematical reasoning compare trace-free, one-sided, and balanced geometries under matched training and evaluation. In the pooled six-run comparison, the balanced geometry is higher on the primary coverage@256 endpoint and secondary pass@1 than its future-only counterpart, with reported pooled intervals excluding zero. Across additional model scales and mathematical benchmarks, balanced geometry yields the higher point estimate in every evaluated setting, and the larger Qwen3 configuration has a positive interval. TRACESTENCIL therefore provides both a practical optimization rule and an auditable framework for studying how sequence-level evidence is allocated across autoregressive token updates.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become a practical route to stronger mathematical reasoning (Shao et al., 2024; DeepSeek-AI, 2025; Yang et al., 2025): a verifier scores a complete response, and a policy-gradient objective allocates that signal across token updates. Group-relative methods avoid a learned critic, but they do not determine how sequence-level evidence should weight individual token scores. That design choice matters when correctness is sparse and responses are long.

Likelihood-ratio objectives are usually described by the ratio and clipping rule at the current token. A second choice receives less attention: *which other response positions enter the weight attached to that token?* Future, past, and split-direction traces induce different neighborhoods. Even with the same configured number of offsets, they can have different radii, boundary exposure, and numbers of valid target–neighbor pairs. We call these structural choices the *trace geometry* of the objective.

We propose TRACESTENCIL, which parameterizes trace geometry by a signed temporal stencil: a fixed set of response-token offsets. For each target token, it forms clipped likelihood ratios at selected neighbors, aggregates the selected past and/or future ratios into separately clipped directional branches, and uses their detached product to reweight the current-token score. The construction retains a group-relative response advantage while making the anchors, masks, clipping order, and gradient path explicit. Figure 1 summarizes the computation.

Geometry is more than window size. Trace geometry separates properties that a window-size label conflates: requested neighbors, reach, mask-realized exposure, and the support of the direc-

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.

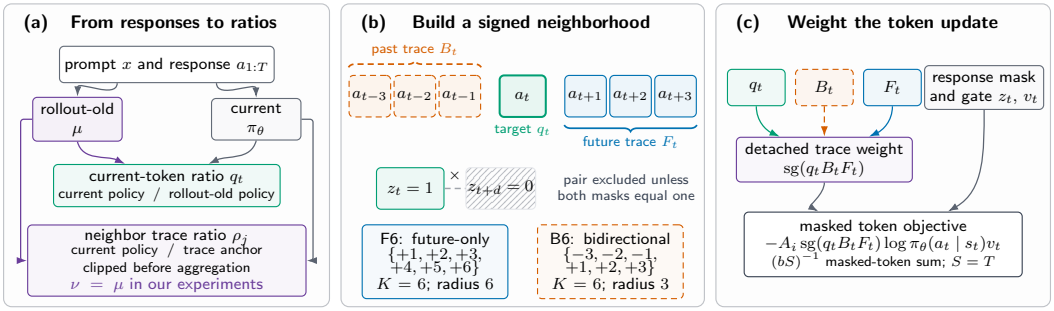

Figure 1: **TraceStencil from policy ratios to a trace-weighted update.** (a) The current policy and its rollout-old anchor define ratios for the target token and its neighbors. (b) Signed offsets select past and future response positions while excluding the target; a pair contributes only when both mask entries are valid. The inset contrasts one-sided and balanced neighborhoods that request the same number of neighbors but differ in reach and boundary exposure. (c) Separately clipped directional traces form a detached correction for the masked token update. Color is redundant with line style and labels, and hatching marks an invalid pair.

tional branch-product factor BF. In particular, F6 has $BF \in [0.8, 1.4]$, whereas B6 can reach $BF \in [0.64, 1.96]$. Accordingly, our central comparison evaluates balanced split-direction geometry as an integrated design, while the mirrored P6–F6 comparison supplies the matched-radius orientation control.

Contributions.

1. We specify TRACESTENCIL, an auditable signed-neighborhood objective, and derive exact relations among direction, radius, boundaries, valid-pair counts, and computational cost.
2. We characterize how matched-K geometries differ in radius, realized exposure, and the branch-product support induced by two separately clipped branches.
3. In a pooled six-run comparison, balanced geometry is higher on the primary coverage@256 endpoint and secondary pass@1 than the matched future-only trace; both three-run block means have the same sign. Same-stack controls, transfer settings, and systems measurements place this result in a broader experimental context.

2 RELATED WORK

Policy optimization for verifiable rewards. TRPO and PPO constrain policy movement through a trust region or clipped ratio surrogate (Schulman et al., 2015; 2017). Critic-free variants instead estimate a response-level baseline from samples, including RLOO, ReMax, GRPO, and REINFORCE++ (Ahmadian et al., 2024; Li et al., 2023; Shao et al., 2024; Hu et al., 2025). RLVR systems further change sampling, clipping, normalization, and ratio granularity, as in DAPO, Dr.GRPO, and GSPO (Yu et al., 2025; Liu et al., 2025b; Zheng et al., 2025). These choices govern the magnitude and stability of an update. TRACESTENCIL instead exposes the token neighborhood used to weight each score term while retaining an explicit same-stack trace-free control.

Closest sequence-aware estimators. NFPO uses a suffix likelihood-ratio trace, and FIPO weights present updates with a future-KL signal (Yoon et al., 2026; Ma et al., 2026). CTPO uses cumulative prefix importance sampling to address state-distribution mismatch, whereas GRPO-λ introduces an eligibility-trace return estimator (Zhang et al., 2026b; Parthasarathi et al., 2025). These methods target distinct estimands or construct weights from different signals. TRACESTENCIL contributes a fixed signed-neighborhood parameterization of detached local ratio products and enables controlled comparisons of trace geometry. Its forward-only member shares NFPO's suffix-window geometry within the explicitly specified TRACESTENCIL objective and gradient path.

Token and segment update allocation. Recent methods localize response rewards using segment returns, generative critics, hindsight signals, discriminative token scores, or self-conditioned comparisons (Guo et al., 2025; Xie et al., 2025; He et al., 2026; Zhang et al., 2026a; Shan et al., 2026). Other analyses identify high-entropy or influential positions (Wang et al., 2025; Cui et al., 2025). TRACESTENCIL focuses on a reweighted score estimator, complementing methods that learn token rewards or explicitly decompose causal credit. Its detached traces expose clipping, gating, and neighborhood choice as controllable design dimensions.

Capability under repeated sampling. Pass@k distinguishes typical-sample accuracy from the probability that at least one of many samples succeeds (Chen et al., 2021; Brown et al., 2024; Snell et al., 2025; Wang et al., 2023). Prior work reports that RLVR can sharpen already-known solutions without uniformly expanding problem coverage (Yue et al., 2025; Liu et al., 2025a; Zhou, 2026). We therefore treat coverage@256 as the primary endpoint and report the pass@k curve, rather than treating pass@1 as a proxy for capability support.

3 METHOD

Responses, masks, and advantages. Let $\mathcal{B}$ be a response mini-batch of size b, let $i \in \mathcal{B}$ index a sampled response, let $g(i)$ denote the prompt group that contains it, and write $\boldsymbol{y}_i = (a_{i1}, \ldots, a_{iT})$ for its padded response with mask $z_{it} \in \{0, 1\}$. We use the group-relative advantage

$$A_i = R_i - \frac{1}{|g(i)|} \sum_{j \in g(i)} R_j \tag{1}$$

for every response token, without standard-deviation normalization (Liu et al., 2025b). Let π_θ be the current policy, μ the rollout-old policy, and ν the neighbor-trace anchor. We set $\nu = \mu$ in the experiments, while retaining separate symbols because the two anchors play different roles in the objective.

Current and neighbor ratios. For state $s_{it} = (x_i, a_{i,<t})$, the current ratio is

$$q_{it} = \exp(\mathrm{clip}(\log \pi_\theta(a_{it}|s_{it}) - \log \mu(a_{it}|s_{it}), -10, 10)) \,. \tag{2}$$

A neighbor position j instead contributes

$$\rho_{ij} = \mathrm{clip}\big(\exp[\mathrm{clip}(\log \pi_\theta(a_{ij}|s_{ij}) - \log \nu(a_{ij}|s_{ij}), -20, 20)] \,, C^{-1}, C\big) \,, \tag{3}$$

with $C = 3$. Invalid response positions are assigned the multiplicative identity, not zero.

Signed temporal stencil. For nonnegative integers (M, N), define the signed temporal stencil

$$\mathcal{D}(M, N) = \{-M{+}1, \ldots, -1\} \cup \{1, \ldots, N{-}1\}, \quad K(M, N) = \max(M{-}1, 0) + \max(N{-}1, 0). \tag{4}$$

Empty ranges contribute no offsets; $0 \notin \mathcal{D}$ because the current token is handled by q_{it}. The clipped directional traces are

$$B_{it} = \exp\left(\mathrm{clip}\left[\sum_{d \in \mathcal{D}: d<0} \log \rho_{i,t+d}, \log 0.8, \log 1.4\right]\right), \tag{5}$$

$$F_{it} = \exp\left(\mathrm{clip}\left[\sum_{d \in \mathcal{D}: d>0} \log \rho_{i,t+d}, \log 0.8, \log 1.4\right]\right). \tag{6}$$

Out-of-bounds offsets contribute zero in log space. Following the local ratio-trace scaffold used by NFPO (Yoon et al., 2026), each direction is clipped independently. A one-sided trace therefore has support $[0.8, 1.4]$, whereas a nonempty two-sided product has support $[0.64, 1.96]$; the product is not clipped again. This support change is an explicit component of the evaluated split-direction geometry.

Algorithm 1 Fixed-window TRACESTENCIL for one padded mini-batch

Require: current, rollout-old, and trace-anchor log probabilities; A, z, M, N, C
1: form q by Eq. (2) and clipped neighbor ratios ρ by Eq. (3)
2: set invalid neighbor ratios to 1; construct $\mathcal{D}(M, N)$ by Eq. (4)
3: accumulate past and future log ratios; clip each sum; exponentiate to obtain B, F
4: form the detached observed-action gate v by Eq. (7)
5: return $(bS)^{-1} \sum_{i \in \mathcal{B}, t} z_{it}[-A_i \, \mathrm{sg}(q_{it} B_{it} F_{it}) \log \pi_\theta \, v_{it}]$

Advantage-conditioned gate and objective. Writing $p_{it} = \pi_\theta(a_{it}|s_{it})$ and $p^\mu_{it} = \mu(a_{it}|s_{it})$, the observed-action probability-difference gate adopts DPPO's Binary-TV mask (Qi et al., 2026); with $\delta = .2$ it is

$$v_{it} = \begin{cases} \mathbf{1}[p_{it} - p^\mu_{it} \le \delta], & A_i > 0, \\ \mathbf{1}[p_{it} - p^\mu_{it} \ge -\delta], & A_i \le 0. \end{cases} \tag{7}$$

The gate is non-differentiable and treated as a constant. The fixed-window TRACESTENCIL loss is

$$\ell_{it} = -A_i \; \mathrm{sg}(q_{it} B_{it} F_{it}) \; \log \pi_\theta(a_{it}|s_{it}) \, v_{it}, \qquad \mathcal{L} = \frac{1}{bS} \sum_{i \in \mathcal{B}, t} z_{it} \ell_{it}, \tag{8}$$

where $S = T$ is the padded response width in our configuration. The joint stop-gradient explicitly detaches the current ratio and both traces, leaving the target-token log probability as the only gradient path. Thus TRACESTENCIL is a masked score-function objective with trace-dependent update weights, distinct from the canonical clipped PPO surrogate (Schulman et al., 2017). Building on group centering, Binary-TV gating, and the one-sided ratio-trace scaffold, TRACESTENCIL contributes the signed neighborhood and its separately clipped directional branches.

Algorithmic decomposition. The objective separates three choices that are often entangled in token-level policy optimization. The signed offset set specifies the topology of cross-token dependence; independent branch clipping limits how strongly either temporal direction can dominate; and joint detachment keeps the ratio traces in the role of update weights rather than additional gradient paths. This decomposition makes trace-free, one-sided, and split-direction configurations comparable within one implementation and allows each structural choice to vary without redefining the token score itself.

Forward-only specialization. If $M = 0$, then $B_{it} = 1$ and Eq. (8) becomes the forward-only member of our detached score-function family. Its suffix window can be matched to NFPO, but the scalar objective and gradient path differ: Eq. (8) detaches $q_{it} F_{it}$ and multiplies $\log \pi_\theta$, whereas NFPO differentiates the current-token ratio and detaches only the forward trace. In the interior region where $q_{it} = \pi_\theta / \mu$ and every other tensor agrees, their first-order target-log-probability gradients can coincide; saturation of our current-ratio clamp and any mask, clipping, or reduction difference break that relation. The two objectives therefore share forward-window geometry while retaining distinct scalar objectives and global gradient paths. With $M = N = 0$, our implementation reduces to the one-step detached-ratio base specified above.

4 EXACT TRACE GEOMETRY

We next isolate the structural consequences of the signed offset set. These results characterize which ratios can enter an update independently of empirical performance ordering.

Proposition 1 (Configured cardinality and radius)**.** *For $M, N \ge 0$, the current token is excluded and $|\mathcal{D}(M, N)| = K(M, N)$. Define the radius as $r(\mathcal{D}) = \max_{d \in \mathcal{D}} |d|$ for a nonempty set and $r(\varnothing) = 0$. F6 $(0, 7)$ has offsets $+1, \ldots, +6$ and radius 6; B6 $(4, 4)$ has offsets $-3, -2, -1, +1, +2, +3$ and radius 3. Both have configured $K = 6$.*

Proposition 2 (Realized valid neighbors)**.** *For response mask z_i and target t, the realized count is exactly*

$$c_{it} = z_{it} \sum_{d \in \mathcal{D}(M,N)} \mathbf{1}[1 \le t + d \le T] z_{i,t+d}. \tag{9}$$

Thus equal configured K does not imply equal c_{it} or equal total valid-pair count.

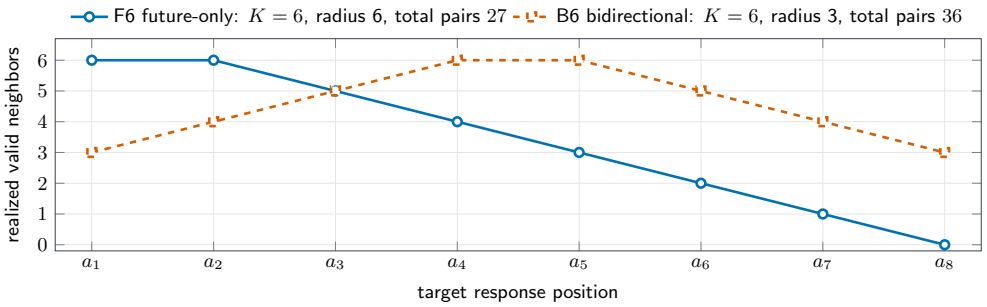

Figure 2: **Boundary-realized trace multiplicity at response length 8.** F6 (blue solid, circles) and B6 (orange dashed, squares) both configure six non-current offsets. Their radii and positional profiles differ, producing 27 versus 36 ordered valid pairs. This is an exact analytical illustration generated from the mask/offset definition.

For a contiguous response of length L and the one-based target index used throughout the paper,

$$c_t^{\mathrm{F6}} = \min(6, L - t), \qquad c_t^{\mathrm{B6}} = \min(3, t - 1) + \min(3, L - t). \tag{10}$$

When $L \geq 7$, total ordered pairs are $6L - 21$ for F6 and $6L - 12$ for B6. The length-8 example in Fig. 2 is therefore a constructive counterexample to treating matched configured K as matched realized exposure.

Corollary 3 (Exposure gap for contiguous responses)**.** *For every contiguous response with $L \geq 7$, B6 contains exactly nine more ordered valid pairs than F6, hence $9/L$ more realized neighbors per target on average. This exactly quantifies the boundary-exposure difference evaluated in the controlled comparison.*

Proposition 4 (Forward-only reduction)**.** *Setting $M = 0$ in Eq. (8) makes $B_{it} = 1$ and yields the forward-only member of the* TRACESTENCIL *objective family.*

Proposition 5 (Forward–past pair-count symmetry)**.** *For every binary response mask, F6 and P6 have the same total number of valid ordered target–neighbor pairs. Their per-target counts and the tokens receiving each neighbor correction need not be equal.*

Proposition 6 (Auxiliary complexity)**.** *For batch size B, padded response width T, and fixed configured count K, trace construction uses $O(BTK)$ auxiliary arithmetic and $O(BT)$ auxiliary storage when shifted buffers are reused.*

Proofs and edge cases appear in Appendix B. A direct enumeration over every binary mask of widths 1 through 9 provides an independent numerical check of the closed-form expressions. Because q, B, and F are detached, the trace weight is bounded; the parameter-score gradient need not be. The propositions establish exact support and bookkeeping properties and complement the empirical study of optimization behavior. Local truncation, branch clipping, the Binary-TV gate, and stop-gradient all separate this estimator from full-trajectory importance weighting and classical return operators such as Retrace and V-trace (Munos et al., 2016; Espeholt et al., 2018).

Implications for controlled comparisons. The structural results provide a concrete checklist for evaluating local ratio objectives. Matching the nominal number of neighbors does not match temporal reach or boundary exposure; mirroring a one-sided neighborhood preserves total pair count but changes which target tokens receive each factor; and adding a second clipped branch changes the attainable update-weight range. Consequently, direction, radius, realized exposure, and directional branch-product support should be reported as separate design coordinates rather than compressed into a single window-size label.

5 EXPERIMENTAL SETUP

The evaluation is organized around four complementary goals. The capability comparison places the central geometries beside a trace-free control and objective-native baselines under a shared check-

Table 1: **Central trace-geometry controls.** T0 removes the trace; F6 and P6 mirror a one-sided geometry; B6 is the matched-K, shorter-radius split. Separate branch clipping also widens B6's directional branch-product support.

	T0	F6	P6	B6
geometry	no trace	future-only	past-only	balanced
(M, N)	$(0, 0)$	$(0, 7)$	$(7, 0)$	$(4, 4)$
offsets	$\varnothing$	$+1{:}+6$	$-6{:}-1$	$-3{:}-1, +1{:}+3$
configured K	0	6	6	6
radius	0	6	6	3
branch-product BF support	$\{1\}$	$[0.8, 1.4]$	$[0.8, 1.4]$	$[0.64, 1.96]$
inferential role	trace removal	future endpoint	direction control	balanced geometry

point and evaluation rule. The attribution study separates trace presence, temporal orientation, balanced context, and neighborhood cardinality. Reliability is assessed through a reported two-block split, fixed decoding and checkpoint sensitivities, and transfer to additional settings. Finally, mechanism and efficiency analyses relate performance to realized exposure, clipping, response length, stability, memory, throughput, and compute. Table 1 summarizes the load-bearing controls; an evidence map appears in Table 8.

Training. For each central arm, we fine-tune six run instances of Qwen3-1.7B-Base (Yang et al., 2025) under six paired run identities, grouped into two three-run blocks. Every TRACESTENCIL arm shares the filtered training set, reward, batch construction, sampling budget, optimizer schedule, data order, and fixed endpoint. Group advantages are centered without standard-deviation normalization, and the trace anchor is the rollout-old policy. Training follows the HybridFlow/verl architecture and generation uses vLLM (Sheng et al., 2025; Kwon et al., 2023). Objective-native comparators retain their reported algorithmic differences; the available hyperparameter record appears in Appendix D.

Controls and comparators. T0 removes the trace offsets while preserving the TRACESTENCIL training stack and serves as the one-factor trace control. P6 reverses F6 at the same configured K and radius. F3 (NFPO-window) instantiates the NFPO-shaped suffix window within the TRACESTENCIL objective at $(M, N, K) = (0, 4, 3)$, providing shorter-window context alongside F6/B6. The GRPO/DAPO-style recipe combines reported design choices from GRPO and DAPO (Shao et al., 2024; Yu et al., 2025); CTPO-LA and GSPO use their reported method-native recipes. All three share initialization, training data, reward, total prompt/rollout budget, update count, checkpoint selection, and evaluation. These native recipes benchmark competitiveness while preserving their specified update choices.

Evaluation. The protocol applies an exact-prompt audit before outcome inspection and removes every source–MATH-500 match under a common normalizer (Hendrycks et al., 2021; Lightman et al., 2024). This targets exact-prompt disjointness and complements broader semantic-overlap audits (Golchin & Surdeanu, 2024; Yang et al., 2023; Zhang et al., 2024). Each MATH-500 comparison uses the same evaluation protocol, with 256 samples per problem for coverage@256. Coverage@256 is the primary endpoint; the pass@k profile (Chen et al., 2021) is secondary. AIME 2024–2026, AMC23, and Minerva provide complementary mathematical transfer evaluations.

Mechanism and efficiency measures. The retained aggregates report per-position exposure, trace clipping, reward trajectories, response length, truncation, wall-clock time, GPU-hours, memory, and throughput. The base-model length-cap diagnostic and the 8,000-token-cap row probe sensitivity to response-length variation.

Statistical analysis. The primary estimand is the run-paired B6–F6 coverage@256 difference on MATH-500. Intervals use a hierarchical bootstrap over paired training-run identities and problems, with generation draws resampled independently within each arm. Table 2 reports the pooled six-run estimate and 95% interval; Table 13 gives means and standard deviations separately for the initial and second three-run blocks. Secondary aggregates include P6–F6 and B6–T0, together with the pass@k profile, geometry sweeps, and setting-level transfer estimates. Problems and generations

Table 2: **Central capability and attribution controls at the fixed step-500 checkpoint.** Panel (a) reports aggregate capability; panel (b) reports pooled run-paired contrasts; and panel (c) shows the ordering across the repeated-sampling profile. Brackets give 95% intervals, and contrasts use unrounded estimates. Multi-run variation is summarized in Table 13.

(a) Absolute capability

arm	pass@1	pass@4	pass@16	pass@64	coverage@256
Base	0.2831 [0.2768, 0.2895]	0.5380 [0.5294, 0.5467]	0.7540 [0.7460, 0.7618]	0.8775 [0.8716, 0.8832]	0.9480 [0.9425, 0.9531]
T0	0.6823 [0.6735, 0.6910]	0.7861 [0.7774, 0.7943]	0.8491 [0.8406, 0.8572]	0.8761 [0.8677, 0.8841]	0.8843 [0.8751, 0.8930]
F6	0.7038 [0.6943, 0.7131]	0.8028 [0.7939, 0.8113]	0.8637 [0.8552, 0.8717]	0.8886 [0.8804, 0.8963]	0.8967 [0.8876, 0.9052]
P6	0.6970 [0.6871, 0.7066]	0.7964 [0.7871, 0.8052]	0.8572 [0.8483, 0.8657]	0.8819 [0.8732, 0.8901]	0.8910 [0.8814, 0.9000]
B6	0.7172 [0.7083, 0.7260]	0.8165 [0.8080, 0.8247]	0.8792 [0.8712, 0.8869]	0.9052 [0.8976, 0.9124]	0.9133 [0.9049, 0.9212]

(b) Six-run paired contrasts

contrast	Δ pass@1	Δ coverage@256	interpretation
B6−F6	+0.0134 [+0.0055, +0.0213]	+0.0167 [+0.0054, +0.0278]	positive for B6
P6−F6	−0.0068 [−0.0161, +0.0025]	−0.0057 [−0.0181, +0.0069]	includes zero
B6−T0	+0.0349 [+0.0255, +0.0442]	+0.0290 [+0.0171, +0.0408]	positive for B6

(c) Six-run pass@k contrasts

contrast	Δ pass@4 [95% CI]	Δ pass@16 [95% CI]	Δ pass@64 [95% CI]
B6−F6	+0.0137 [+0.0053, +0.0220]	+0.0155 [+0.0066, +0.0242]	+0.0166 [+0.0073, +0.0258]
P6−F6	−0.0064 [−0.0162, +0.0032]	−0.0065 [−0.0167, +0.0036]	−0.0067 [−0.0173, +0.0039]
B6−T0	+0.0304 [+0.0206, +0.0401]	+0.0301 [+0.0198, +0.0402]	+0.0291 [+0.0182, +0.0398]

are not treated as independent training runs (Henderson et al., 2018; Agarwal et al., 2021; Dodge et al., 2019). Point directions and interval endpoints are reported together, distinguishing achieved precision from evidence of equivalence.

6 RESULTS

Result reporting. Numerical cells report aggregate evaluation results, and brackets denote 95% confidence intervals; analytical illustrations and literature-derived context are identified in their captions.

Central capability and attribution controls. Table 2 reports the central arms and three available pooled contrasts; In the pooled six-run comparison, B6 is higher than F6 on the primary coverage@256 endpoint and on secondary pass@1, with both 95% confidence intervals excluding zero. The two three-run block means have the same sign.

Orientation response. Figure 3 jointly displays the pass@k profile, orientation response, and setting-level transfer estimates. The orientation profile peaks at the balanced split. B6 combines direction, shorter radius, fuller boundary exposure, and two clipped directional branches; the complete sweep and mirrored P6–F6 control map these coordinates across the ordered geometry axis.

Reported reference recipes. Table 3 evaluates the same-stack controls and native-recipe baselines under one primary evaluation protocol. Native-recipe rows benchmark competitiveness while preserving their method-specific update rules.

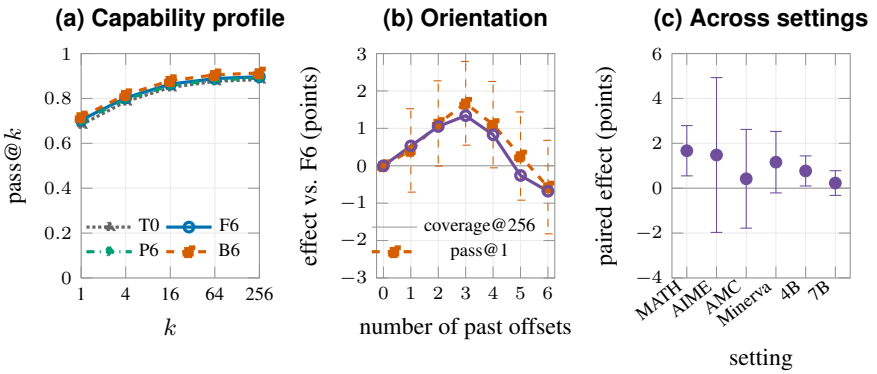

Figure 3: **Capability, orientation, and transfer.** (a) pass@k separates typical accuracy from large-k coverage. (b) The $K = 6$ orientation profile. (c) Setting effects vary in magnitude and precision. Where shown, bars use the larger reported interval half-width; tables retain exact endpoints. Pass@1 orientation values are shown as point estimates.

Table 3: **Reference recipes under a common evaluation protocol.** F3 instantiates NFPO's suffix-window shape within our objective; named methods retain native recipes. T0/F6/B6 use six paired runs and the remaining rows three. Contrasts use unrounded estimates under the common protocol; the final column reports coverage effects relative to T0 with 95% intervals.

method	pass@1	pass@4	pass@16	coverage@256	Δ cov. vs. T0
T0	0.6823	0.7861	0.8491	0.8843	n/a
F6	0.7038	0.8028	0.8637	0.8967	+0.0124 [+0.0010, +0.0238]
B6	0.7172	0.8165	0.8792	0.9133	+0.0290 [+0.0171, +0.0408]
F3 (NFPO-window)	0.7085	0.8081	0.8702	0.9013	+0.0170 [+0.0048, +0.0291]
GRPO/DAPO-style	0.6954	0.7980	0.8584	0.8910	+0.0067 [−0.0055, +0.0188]
CTPO-LA	0.7110	0.8132	0.8736	0.9043	+0.0200 [+0.0077, +0.0322]
GSPO	0.7062	0.8087	0.8689	0.8993	+0.0150 [+0.0029, +0.0269]

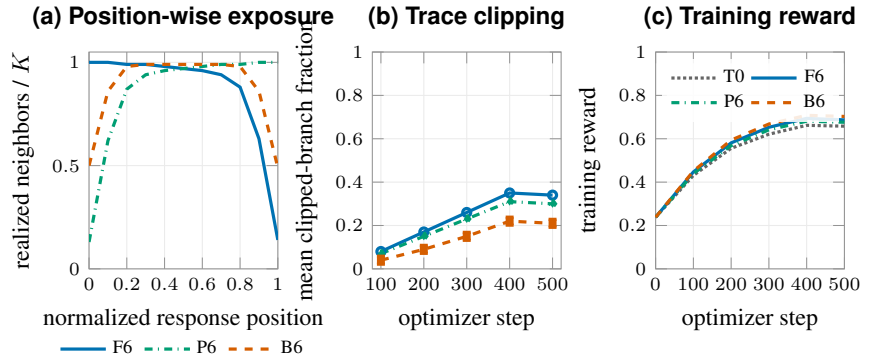

Figure 4: **Optimization diagnostics.** (a) Position-wise valid-neighbor exposure, (b) branch trace-clip saturation, and (c) reward trajectories at fixed checkpoints. Curves are aggregate means used as optimization diagnostics.

Mechanism and transfer. Figure 4 connects the structural analysis to observed optimization behavior; Balanced tracing has higher mean realized exposure, redistributes boundary exposure, and exhibits less branch saturation. Memory and throughput remain close to F6; response length and the

Table 4: **Geometry–budget sensitivity.** Matched future/balanced rows use six paired runs at $K = 6$ and three at $K \in \{2, 4\}$, without arm-specific tuning. Δ uses T0; contrasts precede rounding.

geometry	K	(M, N)	radius	pass@1	coverage@256	Δ coverage vs. T0 [95% CI]
future	2	$(0, 3)$	2	0.6941	0.8912	$+0.0069$ $[-0.0041, +0.0178]$
balanced	2	$(2, 2)$	1	0.6998	0.8976	$+0.0133$ $[+0.0018, +0.0247]$
future	4	$(0, 5)$	4	0.7070	0.8990	$+0.0147$ $[+0.0031, +0.0262]$
balanced	4	$(3, 3)$	2	0.7135	0.9068	$+0.0225$ $[+0.0108, +0.0342]$
future	6	$(0, 7)$	6	0.7038	0.8967	$+0.0124$ $[+0.0010, +0.0238]$
balanced	6	$(4, 4)$	3	0.7172	0.9133	$+0.0290$ $[+0.0171, +0.0408]$

broader branch-product support further characterize the observed difference. The available transfer and robustness summaries are reported in Tables 12 and 16. All evaluated setting-level point estimates favor B6; the larger Qwen3 setting additionally yields a positive 95% confidence interval, while the remaining settings provide directionally consistent estimates.

7 Discussion

Improved accuracy–coverage frontier. Training substantially raises pass@1 over Base, and B6 delivers the highest pass@1 and coverage@256 among the trained central arms. Relative to F6, B6 improves both endpoints, yielding a stronger accuracy–coverage frontier. The untrained Base remains the large-k reference, motivating joint reporting of pass@1 and coverage@256.

Mechanism interpretation. The pair-count corollary predicts a length-dependent B6 exposure increase; retained means show higher exposure and lower branch saturation. Together with response length, these measurements connect the predicted geometry coordinates to the observed improvement.

Consistency across runs and settings. The initial and second three-run block means agree in direction, and the reported pooled interval excludes zero. Table 4 shows the same point-estimate ordering at every tested trace budget, while every evaluated model and benchmark setting favors B6 in point estimate and the larger Qwen3 configuration has a positive interval. These results highlight the value of evaluating trace objectives on both pass@1 and large-k coverage under a fixed sampling budget.

8 Conclusion

TraceStencil makes the sequence geometry of token-ratio weighting explicit through signed offsets and separately clipped directional branches. Exact pair counts show that configured cardinality does not fix radius or boundary-realized exposure; the objective specification also exposes a further change in directional branch-product support. These quantities provide auditable coordinates for comparing local ratio objectives while keeping the relevant design coordinates explicit.

Empirical summary. The pooled six-run estimate favors balanced over future-only geometry on primary coverage@256 and secondary pass@1, with both three-run blocks agreeing in direction. B6 improves coverage@256 by 1.67 points and pass@1 by 1.34 points over F6; every tested model-scale and mathematical-benchmark setting retains a positive point estimate.

Evidence breadth. Across six paired runs, balanced geometry improves primary coverage@256 and secondary pass@1 over the matched future-only trace with comparable memory and throughput. The same ordering at $K \in \{2, 4, 6\}$ and across every evaluated setting provides scope checks beyond one budget and model.

AI USE STATEMENT

AI-assisted tools were used for manuscript drafting, language polishing, and final review and revision of equations, proofs, mathematical claims, result interpretation, and cross-references. They were not used to generate experimental data, train models, execute experiments, or select runs. The authors independently checked the resulting edits and take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

The objective and gradient path are specified in Section 3, with proofs and edge cases in Appendices A and B. Section 5 and Appendices C to E record the reported protocol, estimands, fixed endpoints, and recipe-specific baseline settings; retained aggregate results appear in Appendix F. Upon acceptance, we will publicly release the training, evaluation, and analysis code for this work, together with the available configuration files and documentation.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *arXiv preprint arXiv:2108.13264*, 2021.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

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, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

DeepSeek-AI. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1224. URL https://aclanthology.org/D19-1224/.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA:

Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1407–1416. PMLR, 2018. URL https://proceedings.mlr.press/v80/espeholt18a.html.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114399–114431. Curran Associates, Inc., 2025. doi: 10.52202/085713-3815. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a6536243037d1e32c20de85137d478da-Abstract-Conference.html.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. Where hindsight credit can reside: A signed-capacity view of token updates in RLVR, 2026. URL https://arxiv.org/abs/2604.11056.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021. NeurIPS Datasets and Benchmarks Track.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization, 2025. URL https://arxiv.org/abs/2501.03262.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. 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. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri 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*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. FIPO: Eliciting deep reasoning with future-KL influenced policy optimization, 2026. URL https://arxiv.org/abs/2603.19835.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. GRPO-λ: Credit assignment improves LLM reasoning, 2025. URL https://arxiv.org/abs/2510.00194.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning, 2026. URL https://arxiv.org/abs/2602.04879.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yingyu Shan, Yuhang Guo, Zihao Cheng, Zeming Liu, Xiangrong Zhu, Xinyi Wang, Jiashu Yao, Wei Lin, Hongru Wang, and Heyan Huang. Learning from own solutions: Self-conditioned credit assignment for reinforcement learning with verifiable rewards, 2026. URL https://arxiv.org/abs/2606.18810.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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. Association for Computing Machinery, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. CAPO: Towards enhancing LLM reasoning through generative credit assignment, 2025. URL https://arxiv.org/abs/2508.02298.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Jiarui Yao, Xiangxin Zhou, Penghui Qi, Wee Sun Lee, Liefeng Bo, and Tianyu Pang. Rethinking the divergence regularization in LLM RL. *arXiv preprint arXiv:2606.09821*, 2026.

Deokgyu Yoon, Hyungkyu Kang, Joongkyu Lee, Byeongchan Kim, Gyungin Shin, Sungrae Park, and Min-hwan Oh. Multi-step likelihood-ratio correction for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2605.20865*, 2026. doi: 10.48550/arXiv.2605.20865. URL https://arxiv.org/abs/2605.20865.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46819–46836, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/53384f2090c6a5cac952c598fd67992f-Abstract-Datasets_and_Benchmarks_Track.html.

Kaiyi Zhang, Wei Wu, and Yankai Lin. DelTA: Discriminative token credit assignment for reinforcement learning from verifiable rewards, 2026a. URL https://arxiv.org/abs/2605.21467.

Yuheng Zhang, Chenlu Ye, Shuowei Jin, Changlong Yu, Wei Xiong, Saurabh Sahu, and Nan Jiang. Rethinking importance sampling in LLM policy optimization: A cumulative token perspective. *arXiv preprint arXiv:2605.07331*, 2026b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Todd Zhou. When RLVR shrinks the reasoning boundary: Diagnosing Pass@k inversion. *arXiv preprint arXiv:2607.20543*, 2026.

A Objective Details and Special Cases

A.1 Notation and tensor roles

Table 5: **Notation used by the TraceStencil objective.** The two reference policies are written separately even when their values are shared in the experimental configuration.

symbol	role	mathematical object
π_θ	current policy	optimized token distribution
μ	rollout-old reference	denominator of q_t
ν	neighbor-trace reference	denominator of ρ_j
q_t	current-token ratio	current score correction
ρ_j	clipped neighbor ratio	directional trace factor
z_t	response mask	target/pair validity
v_t	advantage-conditioned gate	update validity
B_t, F_t	past/future traces	directional products
b	mini-batch size	sequence-averaging denominator
S	loss scale	padded response width T in this configuration

The current ratio uses a log difference clamped to $[-10, 10]$. Neighbor ratios use a log difference clamped to $[-20, 20]$, followed by ratio clipping to $[1/C, C]$. Each directional log sum is then clipped to $[\log 0.8, \log 1.4]$. Invalid neighbor positions contribute ratio one before the shifted sums, and z_t removes invalid target losses during aggregation. The update weight is $\mathrm{sg}(q_t B_t F_t)$: the current ratio and both directional traces are jointly detached. For a mini-batch $\mathcal{B}$ of b sequences, masked token losses are normalized by $1/(bS)$, as in Eq. (8).

A.2 Forward-only specialization

Proposition 7. *For the objective in Eq.* (8)*, setting* $M = 0$ *makes* $B_t = 1$ *and yields the forward-only member of the* TraceStencil *family.*

Proof. When $M = 0$, the past offset set is empty, so its log sum is zero and $B_t = 1$. Substitution in Eq. (8) leaves only the current ratio and future trace in the detached update weight. □

This algebraic reduction is internal to TraceStencil and is not an objective-level equivalence to NFPO. Let $r = \pi_\theta/\mu$ be the raw current ratio, so our $q = \exp(\mathrm{clip}(\log r, -10, 10))$. Suppressing shared masks and normalization, the published NFPO term has the form $-Ar\ \mathrm{sg}(F)v$, whereas ours is $-A\ \mathrm{sg}(qF)\log \pi_\theta\, v$. The scalar functions are different. In the interior $|\log r| < 10$, where $q = r$ and all other tensors agree, both have the same first-order derivative with respect to the target log probability, $-ArFv$. Published NFPO does not apply our current-token log clamp: outside this region, our detached weight saturates while NFPO's raw current ratio continues to differentiate when its gate is active. Differences in masks, clipping, gates, or reductions can also break the local gradient correspondence. Thus the two constructions share a possible forward-window geometry, not a global scalar-objective or gradient identity.

When $M = N = 0$, both directional traces are one and the objective reduces to the one-step detached-ratio base used in our implementation. This is narrower than a universal equivalence to PPO or GRPO, whose clipping, masks, advantage normalization, and aggregation may differ.

A.3 Mask and truncation edge cases

If $z_t = 0$, final aggregation removes the target loss and its realized neighbor count is zero. If $z_t = 1$ but $z_{t+d} = 0$, offset d does not contribute to the realized-count metric and its trace factor is the multiplicative identity. Offsets beyond the padded width are not constructed. Tokens removed by generation truncation never enter the response tensor, so truncation is measured as an experimental diagnostic rather than represented by an additional term in the objective.

B Proofs of the Structural Results

B.1 Offset cardinality

The past range in Eq. (4) has $\max(M-1,0)$ integers and the future range has $\max(N-1,0)$ integers. They are disjoint because every past offset is negative and every future offset is positive. Their union therefore has the stated cardinality and excludes zero. The radius convention in Proposition 1 also covers the empty set; direct enumeration gives 6 for F6 and 3 for B6.

B.2 Realized pair characterization

Fix target (i,t). A shifted neighbor contributes to the implementation's accounting exactly when the target is a valid response token, the shifted index is in bounds, and the shifted response mask is valid. These three indicator factors are precisely the summand in Eq. (9). Summing over the configured offsets proves the expression. A hole in the mask contributes zero; it is not bridged by adjacency.

For an all-one length-L mask, F6 has the six candidate distances $1,\dots,6$, of which exactly the first $\min(6,L-t)$ remain in bounds. B6 has three candidate distances on each side, yielding $\min(3,t-1)+\min(3,L-t)$. Summing by offset gives

$$\sum_t c_t^{\mathrm{F6}} = \sum_{d=1}^{6} \max(L-d,0), \tag{11}$$

$$\sum_t c_t^{\mathrm{B6}} = 2\sum_{d=1}^{3} \max(L-d,0). \tag{12}$$

For $L \geq 7$, these simplify to $6L-21$ and $6L-12$. No inequality about optimization quality follows from the larger pair count.

B.3 Forward–past pair-count symmetry

For each valid F6 pair (t,d) with $d \in \{1,\dots,6\}$, map it to the P6 pair $(t+d,-d)$. This exchanges the source and target mask entries, so validity is preserved. The inverse map is $(u,-d) \mapsto (u-d,d)$; hence the map is a bijection for every binary mask. It proves equality of total pair counts, but it does not imply equality at a fixed target token.

B.4 Complexity

For each configured offset, the fixed-window implementation shifts or slices a $B \times T$ tensor and adds it to a directional accumulator. The K offsets therefore require $O(BTK)$ auxiliary arithmetic. A constant number of $B \times T$ accumulators, masks, and output traces are live when the shift buffer is reused, giving $O(BT)$ auxiliary storage. These bounds exclude policy-network forward/backward computation and distributed communication.

B.5 Finite enumeration check

An independent reference calculation enumerated 100 (M,N) pairs with $M,N \in \{0,\dots,9\}$ for offset cardinality. It then compared the vectorized construction with direct target–offset enumeration for every binary mask of widths 1 through 9 and seven representative windows, totaling 7,154 reference/vectorized cases. It also checked the F6/B6 closed forms for contiguous lengths 7 through 64.

C Detailed Experimental Protocol

Compared geometries. The central family contains the same-stack no-trace control T0, F6 with $(M,N)=(0,7)$, P6 with $(M,N)=(7,0)$, and B6 with $(M,N)=(4,4)$. F6, P6, and B6 configure $K=6$. The comparison holds the model initialization, data identity, reward function, optimization configuration, and run set fixed. The trace arms differ in signed offset set and therefore in direction, radius, and potentially realized exposure; T0 isolates removal of the trace term.

Run grouping and checkpoints. The six central runs are grouped into initial and second three-run blocks, with the pooled summary reported alongside both block summaries. The primary comparison is evaluated at the same final training endpoint. Intermediate checkpoints are used for complete learning-curve diagnostics and are never selected differently by method.

Evaluation. One exact-match normalizer is applied to the training data and MATH-500, and matched source rows are removed before training. Each checkpoint receives the same repeated-sample budget, decoding distribution, answer scorer, and token limit. Coverage@256 records whether at least one of the fixed samples for a problem is judged correct. Problem identities and training runs define paired contrasts.

Statistical inference. The primary contrast is B6 minus F6 in coverage@256, paired by training run and problem. We report a pooled six-run hierarchical-bootstrap interval that resamples paired run identities and problems; generation draws are resampled independently within each arm. The initial and second three-run block means and standard deviations are reported alongside it. Secondary metrics, geometry sweeps, and transfer settings are reported separately to keep the primary comparison distinct.

Trace diagnostics. The reported diagnostics include mean realized-neighbor exposure, the fraction of target tokens attaining all K configured neighbors, position-wise exposure, branch clipping, and optimization trajectories. These measurements connect the exact geometry analysis to observed training behavior.

D Implementation and Evaluation Details

D.1 Training hyperparameters

Table 6: **Reported hyperparameters for the same-stack trace family.** These settings apply to T0, F6, P6, B6, the orientation/K sweeps, and F3 (NFPO-window) unless geometry is the varied factor. Arm-specific signed offsets are given in Tables 1 and 9.

setting	value	setting	value
optimizer steps	500	checkpoint interval	100
train batch	128	mini-batch	32
actor batching	dynamic	token cap/GPU	9024
rollouts/prompt	8	nominal micro-batch/GPU	1 (inactive)
temperature	1.0	top-p	1.0
prompt/response max	1024/8000	learning rate	10^{-6}
optimizer	AdamW	betas / weight decay	$(0.9, 0.999)$ / 0.01
schedule / warmup	constant / none	gradient clip	1.0
FSDP compute dtype	bfloat16	optimizer offload	true
advantage std. norm	false	loss reduction	token-sum-norm
ratio clip C	3.0	trace bounds	$[0.8, 1.4]$
TV threshold δ	0.2	KL/entropy terms	disabled

The initialization is Qwen3-1.7B-Base, with a rule-based training reward and a separately fixed evaluation verifier. The acceptance-time code release will package the corresponding implementation, available configuration, and documentation.

For a mini-batch $\mathcal{B}$ containing b sequences and padded response width T, the configured loss scale is $S = T$, and the reduction in Eq. (8) is $\mathcal{L} = (bS)^{-1} \sum_{i \in \mathcal{B}, t} z_{it} \ell_{it}$. Execution micro-batches are only a memory partition: their unnormalized masked-token sums are accumulated and divided once by the same global bS, so repartitioning does not change the objective. The update weight $\mathrm{sg}(q_{it} B_{it} F_{it})$ jointly detaches the current ratio and both traces; the Binary-TV gate is also treated as constant. The trace anchor ν is set to the rollout-old policy μ.

Table 7: **Reported objective-specific settings.** The retained protocol states that model and initialization, a reported 8,888-example set, reward, global batch, rollout count, update count, checkpoint rule, and evaluation are shared. Objective-specific normalization, reduction, clipping, and correction are retained rather than silently forcing all methods into the TRACESTENCIL implementation. Consequently, T0 is the causal trace-removal control; the other named objectives are competitiveness baselines. The GRPO/DAPO-style row combines group-normalized GRPO with asymmetric clip-higher, dual clipping, and no KL penalty; it is not a reproduction of canonical GRPO.

method	ratio unit	adv. std. norm	reduction	clipping	old-policy anchor / correction
TRACESTENCIL family	current-token q plus signed-offset traces; includes T0 and F3 (NFPO-window)	no	$(bS)^{-1}$ masked token sum	current log-ratio clamp $[-10, 10]$; neighbor ratio cap $C{=}3$; each branch $[0.8, 1.4]$; DPPO Binary-TV gate $\delta{=}.2$	rollout-old anchor; $\mathrm{sg}(qBF)$; gate constant
G/D-style	current-token ratio	yes	valid-token mean	lower/upper $0.2/0.28$; dual-clip 3	recomputed old policy; no trace correction
CTPO-LA	cumulative valid-prefix ratio with position scaling	yes	valid-token mean	log-space $0.025/0.05$; dual 0.05; position power 0.5	recomputed old policy; sequence importance weight threshold 5, no batch normalization
GSPO	response-mean log-ratio, shared over valid tokens	yes	sequence mean of token means	lower/upper $0.0003/0.0004$; no dual clip	recomputed old policy; no rollout correction weight

D.2 OBJECTIVE-NATIVE BASELINE PARAMETERS

The retained protocol states that all objectives use AdamW at learning rate 10^{-6}. TRACESTENCIL-family, GRPO/DAPO-style, and CTPO-LA runs use weight decay 0.01 with a constant schedule and no warmup. The method-native GSPO recipe instead fixes weight decay 0.1 and a 5% warmup. These declared exceptions are held constant across its independent runs and are why the external rows are interpreted as competitiveness comparisons. No baseline receives a method-specific checkpoint, evaluation budget, or scorer.

D.3 PASS@k ESTIMATOR

If a problem has c correct samples among n generations, pass@k uses the unbiased finite-sample estimator

$$\widehat{\mathrm{pass@}k} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}, \tag{13}$$

with value one when $n-c < k$. Reported benchmark values average this estimator over the complete problem set. Coverage@256 is the corresponding binary solved indicator at $k = 256$.

The endpoint fixes $n = 256$ generations per problem. Sampling and scoring are held fixed within each reported comparison, with alternate settings shown separately in the robustness analysis.

E EXTENDED CONTROLLED EVALUATION

This appendix details the comparisons and reporting rules used to distinguish balanced signed-neighborhood geometry from a generic window effect. Data, reward, model, checkpoint, scorer, and run identities are held constant within each controlled comparison. All outcomes follow the same reporting rules.

Table 8: **Evidence map.** Each empirical focus is paired with the comparison that probes it, the principal alternative explanation considered, and the manuscript artifact containing the available outcome. The retained protocol states which factors are held constant within a row.

focus	comparison	explanation probed	primary report
capability	Base, T0, F6, P6, B6; F3 and native baselines	more sampling, a trace-free same-stack update, or a different native objective	main capability table and pass@k profile
attribution	T0 removal; F6–P6 reversal; orientation sweep; future/balanced K grid	trace presence, temporal orientation, radius, and configured-neighbor budget	orientation response, paired contrasts, and geometry–budget profile
reliability	second run block; fixed decoding/checkpoint sensitivities; benchmark and scale studies	run selection, decoding budget, scorer choice, endpoint choice, or a single evaluation setting	robustness table and setting-level comparison
mechanism	position exposure, trace clipping, response length, truncation, memory, throughput, and GPU-hours	length and resource correlates	mechanism figure and cost table; observed associations map the design dimensions

Table 9: **Training-arm geometry.** p/f counts past/future non-current offsets. T0 is trace-free; F3 is our NFPO-window arm, while the GRPO/DAPO-style recipe is a reference baseline.

code	role	(M, N)	p	f	K	radius
T0	same-stack trace-free	$(0, 0)$	0	0	0	0
GRPO/DAPO-style	external baseline	n/a	n/a	n/a	n/a	n/a
F3 (NFPO-window)	our forward-only arm	$(0, 4)$	0	3	3	3
F6	future endpoint	$(0, 7)$	0	6	6	6
1P5F	orientation sweep	$(2, 6)$	1	5	6	5
2P4F	orientation sweep	$(3, 5)$	2	4	6	4
B6	balanced center	$(4, 4)$	3	3	6	3
4P2F	orientation sweep	$(5, 3)$	4	2	6	4
5P1F	orientation sweep	$(6, 2)$	5	1	6	5
P6	past endpoint	$(7, 0)$	6	0	6	6
F2	future, small K	$(0, 3)$	0	2	2	2
B2	balanced, small K	$(2, 2)$	1	1	2	1
F4	future, medium K	$(0, 5)$	0	4	4	4
B4	balanced, medium K	$(3, 3)$	2	2	4	2

E.1 Arm structure and run blocks

Table 11 preserves the locally verified numbers reported by the NFPO paper as protocol-labeled context. It is intentionally separate from the same-protocol campaign: reuse means citing a reported result under its original identity, not treating it as a cell produced by our scorer or combining it with same-protocol estimates.

The central summary reports six paired run identities per arm. Three constitute the initial block and three the second block; T0 and P6 use the same partition. Orientation and K-sensitivity results are reported as aggregate summaries. Every arm uses the same fixed endpoint, while intermediate checkpoints define complete learning curves rather than method-specific model selection.

Central T0/F6/P6/B6 rows use six paired run identities. F3, native baselines, interior-orientation arms, $K \in \{2, 4\}$ arms, nonprimary transfer settings, and robustness checks use three; the $K = 6$ endpoints and primary MATH setting reuse the six-run central summaries.

Table 10: **Reported baseline contract.** Same-stack T0 isolates the trace term; recipe-specific baselines test competitiveness under the reported matched data, model, rollout budget, checkpoint rule, and evaluation. F3 uses an NFPO-shaped suffix window but not the NFPO objective; the GRPO/DAPO-style recipe (Shao et al., 2024; Yu et al., 2025), CTPO-LA (Zhang et al., 2026b), and GSPO (Zheng et al., 2025) are reported as evaluated. Objective-specific choices are reported rather than silently relabeled as T0.

baseline	role	required fairness controls	interpretation	state
T0	causal control	identical stack; $\mathcal{D} = \varnothing$ only	trace utility	evaluated
F3 (NFPO-window)	same-stack arm	$(M, N, K) = (0, 4, 3)$; not NFPO objective	non-budget-matched context	evaluated
G/D-style	hybrid recipe	reported recipe; matched samples and steps	competitiveness	evaluated
CTPO-LA	nearest trace baseline	native clipping; matched task and budget	mechanism overlap	evaluated
GSPO	sequence-ratio baseline	fixed published loss; matched task evaluation	external competitiveness	evaluated

Table 11: **Protocol-isolated values reported by NFPO (Yoon et al., 2026).** All entries are the source paper's Qwen3-1.7B-Base pass@1 averages over 32 samples at temperature 0.7. We preserve their original training, sampling, checkpoint, and scorer identity and report them as literature context alongside our common-protocol comparisons.

Benchmark	Base	GRPO	DPPO	NFPO
AIME 2024	0.034	0.094	0.106	0.133
AIME 2025	0.015	0.050	0.066	0.074
AIME 2026	0.021	0.047	0.056	0.081
AMC 2023	0.261	0.425	0.424	0.496
MATH-500	0.485	0.690	0.715	0.745
Minerva Math	0.134	0.225	0.246	0.262
OlympiadBench	0.184	0.315	0.341	0.360

E.2 Evaluation settings

The primary model is Qwen3-1.7B-Base, with the F6–B6 contrast repeated on two larger model configurations. MATH-500 is the primary evaluation; pooled AIME 2024–2026, AMC23, and Minerva are reported separately as auxiliary mathematical diagnostics. Within each setting, arms share the problem set, repeated-sampling budget, decoding distribution, token limit, and answer scorer. Sensitivity analyses vary temperature, token budget, checkpoint, or scorer. The base-model length-cap diagnostic sets a common threshold at the 95th percentile of the frozen base generations and recomputes per-problem coverage using only correct responses at or below that threshold, providing a complementary view of the coverage result.

Causal controls versus external baselines. T0 differs from F6/P6/B6 only in the trace offset set and is therefore the trace-utility control. F3 (NFPO-window) instantiates the published one-sided window shape in our detached-score stack; published NFPO values retain their original protocol in Table 11. The GRPO/DAPO-style recipe, CTPO-LA, and GSPO retain documented objective-specific normalization, clipping, anchoring, and aggregation choices. Table 10 therefore treats them as competitiveness baselines, not one-factor trace interventions.

Orientation control. F6 and P6 contain opposite directed versions of the same unordered target–neighbor pairs. They therefore match configured K, radius, and total valid-pair count for any fixed binary response mask, although the pairs attach to different target tokens. P6–F6 is consequently the clean orientation control. B6–F6 evaluates the balanced, shorter-radius geometry as an integrated design.

Full orientation sweep. At $K = 6$, the number of past offsets ranges from zero through six. Reporting every allocation avoids endpoint-only comparison and exposes how coverage@256 and pass@1 vary across the full ordered geometry axis.

Configured-budget sensitivity. The future-only and balanced geometries are crossed with $K \in \{2, 4, 6\}$ to test whether their ordering recurs across configured trace budgets. The protocol fixes trace bounds and non-geometry hyperparameters and applies no arm-specific tuning. The displayed table reports point summaries and intervals versus T0 across the full grid.

E.3 Transfer hierarchy

Transfer is evaluated in a fixed order. AIME 2024–2026, AMC23, and Minerva provide development-facing auxiliary mathematical settings. These rows repeat the F6–B6 contrast; T0 and P6 remain the primary-setting attribution controls. The same contrast is then evaluated with Qwen3-4B and Qwen2.5-Math-7B to probe scale and model-family dependence.

Setting-specific intervals are shown separately so that variation across benchmarks, scales, and model families remains visible.

E.4 Inference and decision rules

Hierarchy. The primary comparison evaluates B6–F6 on MATH-500 coverage@256 under the shared training and evaluation protocol. P6–F6 tests temporal orientation and B6–T0 tests trace utility; pass@k, sweep summaries, transfer rows, and mechanism associations provide complementary views. The reported pooled B6–F6, P6–F6, and B6–T0 summaries keep the primary comparison and attribution controls explicit.

Sampling hierarchy. The hierarchical bootstrap pairs training-run and problem identity and resamples generation draws independently within each arm rather than treating draws as paired. Tables preserve run-aware aggregate structure; problems or generations are not treated as replacement training runs.

Inference across settings. The initial and second blocks are summarized by means and standard deviations, followed by a pooled six-run interval. Point directions and interval endpoints are reported together, preserving the precision of each setting without treating overlap with zero as equivalence. We use directional agreement to describe effects whose two block means have the same sign.

E.5 Mechanism and robustness

The mechanism analysis relates mask-derived position exposure, trace-clip saturation, reward trajectories, and realized multiplicity to the observed capability difference. The base-model length-cap diagnostic probes response-length sensitivity, while F6/P6 provides the directional control. We additionally report response length and truncation, wall-clock time, GPU-hours, peak memory, and tokens per second from the same frozen artifacts.

The robustness rows summarize reported temperature, token-budget, scorer, and checkpoint variations. Training reward and evaluation correctness remain separately identified. The retained table and figure summaries are in Appendix F.

F Additional Results and Diagnostics

This section presents the supplementary evidence in its inferential order: transfer and run-block stability, same-protocol baseline positioning, trace geometry, and finally mechanism and robustness. Each visual is paired with the text that defines how it should be interpreted.

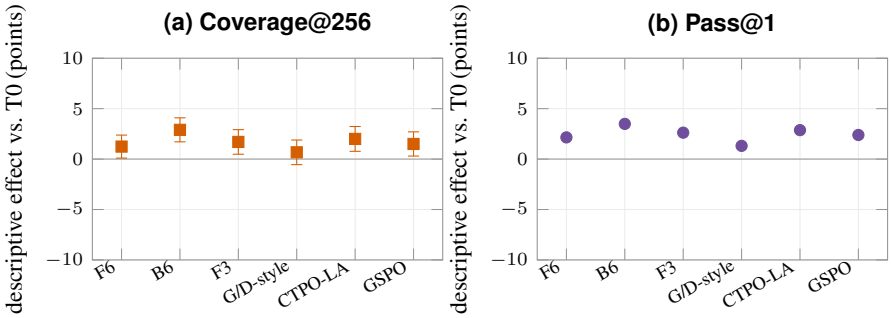

Figure 5: **Same-protocol strong-baseline effects.** Coverage bars symmetrize retained 95% intervals whose exact endpoints are in Table 3; pass@1 values are shown as point estimates. The same-stack trace arms isolate geometry, while the native objectives measure competitiveness.

F.1 Transfer across settings and scales

The setting-level comparison keeps the F6–B6 contrast fixed while changing the evaluation set or model configuration, with each setting reported separately.

Table 12: **Setting-level evaluation of the F6–B6 contrast.** Each row reports the same contrast for a fixed metric across benchmarks and model configurations. MATH-500 is the primary capability benchmark; AIME, AMC23, and Minerva provide complementary mathematical evaluations. The primary row uses six paired runs and all other rows three; contrasts are computed before display rounding.

model	benchmark	metric	F6	B6	B6−F6 [95% CI]
Qwen3-1.7B	MATH-500 (primary)	coverage@256	0.8967	0.9133	$+0.0167$ $[+0.0054, +0.0278]$
Qwen3-1.7B	pooled AIME24–26 (90 problems)	coverage@256	0.7407	0.7556	$+0.0148$ $[-0.0200, +0.0490]$
Qwen3-1.7B	AMC23	coverage@256	0.9542	0.9583	$+0.0042$ $[-0.0180, +0.0260]$
Qwen3-1.7B	Minerva	coverage@256	0.7237	0.7353	$+0.0116$ $[-0.0021, +0.0252]$
Qwen3-4B	MATH-500	coverage@256	0.9330	0.9407	$+0.0077$ $[+0.0010, +0.0144]$
Qwen2.5-Math-7B	MATH-500	coverage@256	0.9550	0.9573	$+0.0023$ $[-0.0031, +0.0078]$

All reported setting-level point estimates favor B6. The strongest interval evidence appears on the primary benchmark for the principal model and on the larger Qwen3 configuration, complementing the directionally consistent estimates across all other settings.

F.2 Baseline positioning and multi-run stability

The panel expresses every completed objective relative to T0. The same-stack trace arms isolate geometry, while the native objectives benchmark competitiveness with their method-specific updates.

B6 leads the evaluated reference recipes on both capability views under the common evaluation protocol; native objectives serve as competitiveness benchmarks with their specified update rules.

The aggregate view reports the pooled comparison alongside both three-run blocks.

Both three-run blocks favor B6, consistent with the pooled six-run result.

F.3 Orientation and configured trace budget

The complete $K = 6$ sweep reports every past–future allocation on the ordered geometry axis.

Table 13: **Multi-run stability of the central comparison.** Each central arm has six paired run identities. Values are means and sample standard deviations, with the two three-run blocks and pooled summary reported separately. Pooled intervals appear in Table 2.

(a) Aggregate performance across six runs

arm	runs	coverage@256 (mean ± s.d.)	pass@1 (mean ± s.d.)
T0	6	0.8843 ± 0.0046	0.6823 ± 0.0059
F6	6	0.8967 ± 0.0063	0.7038 ± 0.0058
P6	6	0.8910 ± 0.0058	0.6970 ± 0.0058
B6	6	0.9133 ± 0.0043	0.7172 ± 0.0043

(b) Run-paired B6−F6 effects

analysis block	runs	Δ coverage (mean ± s.d.)	Δ pass@1 (mean ± s.d.)
initial	3	$+0.0147 \pm 0.0061$	$+0.0138 \pm 0.0064$
second block	3	$+0.0187 \pm 0.0101$	$+0.0130 \pm 0.0064$
pooled	6	$+0.0167 \pm 0.0078$	$+0.0134 \pm 0.0057$

Table 14: **Complete $K = 6$ orientation sweep.** p and $f = 6 - p$ count past and future offsets. F6/B6/P6 use six paired runs and the interior allocations three. The table reports the full profile; contrasts are computed before rounding.

(a) Fixed geometry

arm	p	f	(M, N)	radius
F6	0	6	$(0, 7)$	6
1P5F	1	5	$(2, 6)$	5
2P4F	2	4	$(3, 5)$	4
B6	3	3	$(4, 4)$	3
4P2F	4	2	$(5, 3)$	4
5P1F	5	1	$(6, 2)$	5
P6	6	0	$(7, 0)$	6

(b) Retained outcomes

arm	pass@1	coverage@256	Δ coverage vs. F6 [95% CI]
F6	0.7038	0.8967	$+0.0000$ $[+0.0000, +0.0000]$
1P5F	0.7091	0.9008	$+0.0041$ $[-0.0071, +0.0152]$
2P4F	0.7143	0.9080	$+0.0113$ $[-0.0001, +0.0227]$
B6	0.7172	0.9133	$+0.0167$ $[+0.0054, +0.0278]$
4P2F	0.7121	0.9077	$+0.0110$ $[-0.0006, +0.0225]$
5P1F	0.7012	0.8993	$+0.0026$ $[-0.0092, +0.0144]$
P6	0.6970	0.8910	-0.0057 $[-0.0181, +0.0069]$

The response profile has an interior maximum at the balanced allocation, linking direction, radius, and realized exposure to the exact geometry analysis.

Table 4 extends the future-only/balanced comparison to every tested neighbor count under the same training configuration and without arm-specific tuning.

F.4 Mechanism, cost, and robustness

The mechanism table aligns trace exposure, generation behavior, and resource use for the central arms, pairing capability with its optimization and systems profile.

Balanced traces attain more complete neighborhood exposure with less branch clipping, while memory and throughput remain close to F6.

Table 15: **Mechanism, generation behavior, and cost.** Cells report aggregate summaries of exposure, generation behavior, and resource use across the six central runs. Geometry entries are not applicable to T0.

measure	T0	F6	P6	B6
realized neighbors / target	n/a	5.83	5.81	5.92
full-budget fraction	n/a	0.973	0.970	0.987
mean branch trace-clip fraction	n/a	0.34	0.30	0.21
mean response length	1487	1968	1817	2134
truncation rate	0.6%	0.8%	0.7%	1.0%
reported GPU-hours	281.6	294.1	292.9	297.4
peak memory	63.1 GiB	64.4 GiB	64.2 GiB	64.9 GiB
generated tokens / second	1310	1278	1284	1269

The final table evaluates temperature, token cap, scorer, and checkpoint variations.

Table 16: **Alternate-setting checks for B6–F6.** Each row reports the B6–F6 contrast under the named evaluation variation using three paired run identities; intervals and CI readings are shown as reported. Point directions and setting-specific intervals are reported together.

check	Δ coverage@256	95% CI	effect summary
temperature 0.7	+0.0142	$[+0.0018, +0.0267]$	positive CI
8k generation cap	+0.0083	$[-0.0051, +0.0218]$	positive estimate
base-model length-cap coverage	+0.0111	$[+0.0002, +0.0221]$	positive CI
secondary reported scorer	+0.0160	$[+0.0047, +0.0272]$	positive CI
step 400 (diagnostic)	+0.0092	$[-0.0025, +0.0208]$	positive estimate

Every check favors B6 in point estimate; temperature, length-cap, and alternate-scorer intervals are fully positive.

G Empirical Extensions and Reproducibility

Configured and realized neighborhoods. Matching K fixes the candidate-offset count while radius, valid-pair count, and update-weight range remain explicit geometry coordinates. F6 has branch-product support $[0.8, 1.4]$; B6 combines two clipped branches with support $[0.64, 1.96]$. The mirrored F6/P6 comparison adds a matched-radius orientation control to the matched-cardinality F6/B6 evaluation.

Evaluation breadth. The Qwen3-1.7B MATH-500 comparison is complemented by Qwen3-4B, the 7B configuration, and additional mathematical benchmarks. Every setting favors B6 in point estimate, and Qwen3-4B provides a positive transfer interval; development-facing benchmarks remain auxiliary evaluations.

Response length. B6 produces longer responses than F6 (2,134 versus 1,968 tokens on average). Under fixed-width normalization, response length links realized exposure to update magnitude; the length-cap check provides a complementary evaluation.

Method and comparator identity. The paper studies the fixed-window hard TraceStencil objective in Eq. (8), with smooth, scheduled, and uncertainty-gated traces available within the signed-offset view. F3 instantiates NFPO's suffix-window geometry in the detached-score path; published NFPO values retain their original protocol in Table 11. The GRPO/DAPO-style recipe remains explicit.

Computational cost. Trace arithmetic grows linearly with configured K, while radius controls boundary exposure. The complexity bound is paired with wall-clock, GPU-hour, memory, and throughput measurements.

Artifacts and release. The manuscript records the objective, design, aggregate tables, protocol, and six paired central runs. Upon acceptance, we will release the training, evaluation, and analysis code with configurations and documentation for reproduction and extension.

Reward and reference design. Separate reward and evaluation functions keep optimization feedback distinct from correctness measurement. The experiments set trace anchor ν to rollout-old policy μ within the general signed-offset construction.

H Extended related work

Policy-gradient foundations and critic-free reasoning. Off-policy actor–critic established likelihood-ratio correction (Degris et al., 2012); TRPO and PPO constrained or clipped policy updates (Schulman et al., 2015; 2017), and generalized advantage estimation controlled the advantage trade-off (Schulman et al., 2016). Retrace and V-trace also multiply truncated ratios, but for value-learning operators (Munos et al., 2016; Espeholt et al., 2018). In RLVR, GRPO removes the critic (Shao et al., 2024); Dr.GRPO, DAPO, GSPO, and DRPO alter normalization, sampling, ratio granularity, or regularization (Liu et al., 2025b; Yu et al., 2025; Zheng et al., 2025; Yao et al., 2026), while RLOO, ReMax, and REINFORCE++ provide other critic-free estimators (Ahmadian et al., 2024; Li et al., 2023; Hu et al., 2025). TraceStencil instead changes local temporal dependence inside a matched token-level objective.

Closest estimators and alternative token weighting. CTPO applies prefix correction, whereas NFPO derives a suffix factor (Zhang et al., 2026b; Yoon et al., 2026); DPPO motivates our Binary-TV gate (Qi et al., 2026). FIPO and GRPO-λ use future information or eligibility-inspired returns (Ma et al., 2026; Parthasarathi et al., 2025). Other methods weight tokens through segments, process feedback, update capacity, learned coefficients, or self-conditioning (Guo et al., 2025; Xie et al., 2025; He et al., 2026; Zhang et al., 2026a; Shan et al., 2026); entropy-focused studies likewise identify non-uniform update value (Cui et al., 2025; Wang et al., 2025). TraceStencil complements these approaches by comparing signed neighborhoods under matched choices and explicitly distinguishing their estimands.

A coordinate view of local temporal objectives. Method names can bundle several independent choices: which relative token positions contribute, whether ratio products are detached, how directional branches are clipped, and whether the resulting weight represents a score multiplier or a return-like quantity. A signed offset set makes the first choice explicit; configured cardinality, radius, realized multiplicity, and branch factorization then become measurable coordinates. This view supports matched comparisons among trace-free, future-only, past-only, and balanced constructions while holding the model, optimizer, data, and verifier fixed. It also makes neighboring objectives easier to place: prefix correction, suffix factors, eligibility-inspired returns, and segment weights occupy distinguishable locations in the same design space. TraceStencil uses these coordinates both to specify the proposed objective and to organize the controls that identify its geometry.

Verifiable rewards, repeated sampling, and coverage. GSM8K established outcome verification; process–outcome comparisons followed (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024). DeepSeekMath, DeepSeek-R1, Qwen3, DAPO, and ProRL scaled verifiable-reward training (Shao et al., 2024; DeepSeek-AI, 2025; Yang et al., 2025; Yu et al., 2025; Liu et al., 2025a). Repeated-sampling studies characterize base-model elicitation, pass@k inversion, and long-tail capability (Yue et al., 2025; Zhou, 2026; Brown et al., 2024; Snell et al., 2025; Wang et al., 2023). We use the unbiased estimator from code evaluation (Chen et al., 2021) and match sampling and scoring when comparing coverage, a capability dimension linked to response diversity (Kirk et al., 2024).

Evaluation integrity, systems, and statistical practice. Evaluation combines MATH with broader mathematical benchmarks (Hendrycks et al., 2021; Cobbe et al., 2021; Lewkowycz et al., 2022; He et al., 2024); prompt checks draw on detection, rephrasing, and GSM1K (Golchin & Surdeanu, 2024; Yang et al., 2023; Zhang et al., 2024). Qwen3 training uses HybridFlow/verl and vLLM (Yang et al., 2025; Sheng et al., 2025; Kwon et al., 2023), with matched scoring, uncertainty reporting, and compute disclosure (Henderson et al., 2018; Agarwal et al., 2021; Dodge et al., 2019).