Let Credit Follow Computation: Architecture-Aware Credit Transport for Large Language Model Reinforcement Learning

Qifan Shi Zhaolu Kang Chenghua Zhu

August 2026

Abstract

Credit assignment in large-language-model reinforcement learning (LLM RL) can be separated into three objects: *evidence* about success, a *transport operator* that converts this evidence into token-level advantages, and an *update geometry* that turns advantages into policy changes. Recent work has greatly improved evidence, sampling, and update geometry, but the transport operator is usually architecture-agnostic. Fixed-discount GAE applies a stationary geometric kernel along token time; group-relative methods broadcast an outcome statistic across an entire response. Neither operator represents the trajectory-specific computation used by the Transformer policy itself.

We introduce *computation-conditioned credit transport* (CCT), a general framework in which a detached statistic of the behavior policy's internal computation parameterizes the causal kernel that transports downstream value through a rollout. Our concrete algorithm, CompPO, maps native attention concentration to a bounded per-token retention gate, uses the gate in both the one-step bootstrap and a path-dependent generalized-advantage trace (Comp-GAE), and co-designs a transport-aligned critic (TAC) that reuses the actor's hidden states and routing information without a second same-scale Transformer. The task reward and clipped PPO policy objective remain unchanged; a constant gate recovers fixed-coefficient GAE.

Across five Qwen3-4B seeds, CompPO reaches 61.4% final held-out development accuracy (95% CI $[60.8, 62.0]$), versus 53.8% $[52.9, 54.7]$ for a separately tuned GRPO baseline. A controlled 2×2 gate-by-critic experiment shows that neither Comp-GAE with a standard critic (55.2%) nor TAC with a mean-matched fixed gate (56.4%) reproduces the full result; the seed-matched final interaction is +2.4 points $[1.9, 2.9]$. Gate shuffling and position-only controls show that trajectory-specific alignment matters beyond a favorable marginal schedule. In a matched PPO stress grid, CompPO is stable in 10/12 runs versus 3/12 for PPO. Final-checkpoint frozen evaluation improves over GRPO by 4.3 and 3.9 greedy pass@1 macro points on Qwen3-4B and Llama-3.1-8B-Instruct, respectively. These results establish policy-internal computation as a useful estimator variable and open a broader research program on architecture-aware credit kernels, traces, and critics.

Keywords: large language models, reinforcement learning, credit assignment, actor–critic, attention, generalized advantage estimation

1 Introduction

Reinforcement learning with verifiable or preference-derived rewards has become a primary route to stronger large-language-model (LLM) reasoning and decision making. The algorithmic landscape has advanced quickly: PPO supplies bootstrapped token-level advantages (Schulman et al., 2017, 2015); RLOO and GRPO replace the critic with response-level comparison (Ahmadian et al., 2024; Shao et al., 2024); DAPO, Dr. GRPO, and GSPO improve clipping, normalization, sampling, and importance-ratio geometry (Yu et al., 2025; Liu et al., 2025b; Zheng et al., 2025); process-reward and rollout-based methods generate denser evidence about intermediate decisions (Lightman et al., 2023; Uesato et al., 2022; Cui et al., 2025; Kazemnejad et al., 2024). Yet one part of the learning stack remains comparatively implicit: *the operator that transports delayed evidence through the generated trajectory.*

This omission matters because a Transformer policy is not a homogeneous token chain. At each position it performs a content-dependent computation over the prefix: heads retrieve different events, representations compose them, and the active dependency pattern changes across trajectories. A late mathematical conclusion may use an early substitution; a tool action may depend on an observation many turns earlier; a code token may be constrained by a distant declaration. The policy is therefore architecture-rich and history-structured. Its credit estimator is usually not. Fixed-discount GAE uses one continuation coefficient at every transition; group-relative methods assign a response-level statistic to all valid tokens in the response. We call this discrepancy the **architecture–credit mismatch**.

To make the mismatch precise, we factor a token-level policy-gradient pipeline into three conceptually distinct objects:

1. **credit evidence**: what indicates that a decision sequence was good or bad—terminal verifiers, process

rewards, judges, counterfactual continuations, or critic residuals;

2. **credit transport**: the causal operator that moves that evidence across positions or states to produce token-level advantages; and
3. **update geometry**: importance ratios, clipping, KL constraints, entropy terms, and normalization that convert advantages into a policy update.

This factorization clarifies why many powerful LLM-RL advances are complementary rather than interchangeable. Better verifiers change evidence. GRPO changes the baseline and sampling statistic. GSPO changes update geometry. A process reward changes where evidence enters. None of these choices uniquely determines how downstream evidence should be transported through the policy's own computation.

For standard GAE, the transport can be written as a stationary upper-triangular kernel. If $\boldsymbol{\delta}$ is the vector of TD residuals, then

$$\mathbf{A} = \mathbf{K}^{\gamma,\lambda}\boldsymbol{\delta}, \qquad K^{\gamma,\lambda}_{t,u} = \mathbf{1}[u \geq t](\lambda\gamma)^{u-t}. \tag{1}$$

The kernel depends only on temporal separation. A canonical group-relative estimator instead produces $\mathbf{A}^{(i)} = b_i\mathbf{1}$ within response i: a rank-one broadcast geometry with no within-response transport structure. Both are coherent estimators; both discard the realized internal computation of the policy.

Our thesis is:

> *The policy already computes a trajectory-specific dependency structure; credit transport should be allowed to condition on that computation.*

We introduce **computation-conditioned credit transport** (CCT). Let $\Phi_{\theta^-}(h_t)$ be a statistic extracted from the behavior policy's internal computation at prefix h_t, and let a bounded map produce a retention gate κ_t. The resulting path kernel is

$$K^{\kappa,\lambda}_{t,u} = \mathbf{1}[u \geq t]\lambda^{u-t}\prod_{j=t}^{u-1}\kappa_j. \tag{2}$$

Unlike Equation (1), it is nonstationary and trajectory-specific: the sensitivity of A_t to a later residual δ_u is the realized product of gates along the intervening computation. The gate is computed by the old policy, detached, and stored with the rollout. Thus CCT changes the estimator-side transport kernel, not the environment reward or task objective.

We instantiate the framework as **Computation-Conditioned Policy Optimization** (CompPO).[1] CompPO uses native attention concentration as an inexpensive scalar summary of whether the current computation is dominated by a sparse subset of history or supported diffusely. The gate enters a one-step computation-conditioned bootstrap and a path-dependent generalized-advantage trace, Comp-GAE. A transport-aligned critic (TAC) pools value-relevant history from the actor's existing hidden states and attention and uses the same gate to couple local and routed-history value estimates. The outer clipped PPO objective, verifier, rollout format, and reward are unchanged.

The method is deliberately more than an attention-weighted loss. Attention does not directly multiply the policy gradient. It parameterizes the bootstrap coefficient, every downstream residual's trace weight, and the critic state geometry used to estimate those targets. Nor do we claim that attention is causal token attribution. Concentration is a policy-native structural proxy, and its usefulness must be established through controlled alignment tests.

The empirical evidence is organized to test the main alternative explanations directly. First, GRPO and CompPO are screened over the same 3×3 actor-learning-rate/KL grid and confirmed with five seeds. CompPO reaches 61.4% final held-out development accuracy, compared with 53.8% for GRPO, and every CompPO seed exceeds every GRPO seed at both best and final endpoints. Second, a five-seed 2×2 experiment crosses dynamic versus mean-matched fixed transport with standard versus aligned critics. Dynamic transport with the standard critic reaches 55.2% final accuracy; the aligned critic with a fixed gate reaches 56.4%; full CompPO reaches 61.4%, with a positive final interaction of +2.4 points $[1.9, 2.9]$. The standard critic has negative explained variance and weak TD-target rank correlation, whereas the aligned critic tracks the induced targets substantially better. Third, global gate shuffling and a frozen position-only schedule reduce performance, showing that the result is not explained by marginal coefficient scale or position alone. Fourth, a matched PPO stress grid supports a bounded robustness claim: CompPO is stable in 10/12 runs versus 3/12 for PPO. Finally, final-checkpoint frozen evaluation improves over GRPO by 4.3 and 3.9 greedy macro points on Qwen3-4B and Llama-3.1-8B-Instruct.

The deeper implication is that LLM RL need not remain architecture-agnostic. The dominant critic-free trend may reflect a contingent mismatch between conventional value representations and long-horizon targets, rather than an intrinsic impossibility of actor–critic learning for language models. More broadly, attention, retrieval, memory access, MoE routing, tool dependencies, or cross-modal information flow could parameterize future Bellman-style operators, traces, critics, and trust regions.

Our contributions are:

- **A factorization and problem abstraction.** We separate credit evidence, credit transport, and update geometry, identify the architecture–credit mismatch, and

[1] The anonymous submission used the working name NMPO and denoted the gate by α_t. We rename the framework and algorithm to foreground the more general contribution and to avoid conflating estimator-side retention with task discounting. The implemented mechanism and objective experimental results are unchanged.

characterize fixed-discount GAE and group-relative broadcast as two architecture-agnostic transport geometries.

- **A general framework.** CCT turns detached policy-internal computation into a path-dependent causal transport kernel. We derive its matrix form, sensitivity interpretation, reduction to fixed GAE, contraction, and bounded-trace properties.
- **A concrete algorithm.** CompPO couples an attention-derived retention gate, Comp-GAE, and a transport-aligned actor-feature critic while preserving the external reward and clipped PPO update.
- **Mechanism-level evidence.** Matched tuning, five-seed factorials, critic diagnostics, shuffling, position-only schedules, a PPO stress grid, a stronger group baseline, and two-backbone frozen evaluation isolate what the mechanism does and where the evidence stops.
- **A research agenda.** The work elevates policy-internal computation to a first-class RL estimator variable and motivates architecture-aware credit transport beyond the scalar attention instantiation studied here.

2 Credit Assignment as Evidence, Transport, and Update

2.1 Trajectory setup

Let x be a prompt and $y_{1:T}$ a response sampled from $\pi_\theta(y_t \mid h_t)$, where $h_t = (x, y_{<t})$. Let r_t denote externally specified reward evidence at position t; the main experiments use a sparse terminal verifier. A policy-gradient method ultimately constructs token coefficients A_t and applies them through an update rule such as PPO. We write this process abstractly as

$$\underbrace{\mathbf{e}}_{\text{evidence}} \xrightarrow{\mathbf{K}} \underbrace{\mathbf{A}}_{\text{transported credit}} \xrightarrow{\mathcal{U}} \underbrace{\Delta\theta}_{\text{policy update}}, \qquad \mathbf{A} = \mathbf{K}\mathbf{e}. \tag{3}$$

Here $\mathbf{e}$ may contain rewards, TD residuals, process scores, counterfactual estimates, or other local evidence. $\mathbf{K}$ is a causal transport operator, and $\mathcal{U}$ contains likelihood ratios, clipping, KL, entropy, and normalization. The factorization is analytic rather than exclusive: an algorithm may jointly modify all three terms. Its purpose is to identify which object a contribution changes.

2.2 Fixed-discount GAE is stationary temporal transport

With a value function V_ψ, standard GAE uses

$$\delta_t^\gamma = r_t + \gamma V_\psi(h_{t+1}) m_{t+1} - V_\psi(h_t) m_t, \tag{4}$$

$$A_t^{\gamma,\lambda} = \delta_t^\gamma + \gamma\lambda A_{t+1}^{\gamma,\lambda} m_{t+1}, \tag{5}$$

where m_t masks invalid or padded positions. Ignoring masks for notation and stacking residuals gives

$$\mathbf{A} = (\mathbf{I} - \lambda\gamma\mathbf{S})^{-1}\boldsymbol{\delta}, \tag{6}$$

where $(\mathbf{S}\mathbf{v})_t = v_{t+1}$. Therefore

$$K_{t,u}^{\gamma,\lambda} = \mathbf{1}[u \geq t](\lambda\gamma)^{u-t}. \tag{7}$$

The transport kernel is stationary and Toeplitz: it depends only on token distance. This is not equivalent to saying that GAE has no local signal; TD bootstrapping can make δ_t informative even with sparse reward. The structural limitation is narrower: the continuation geometry contains no variable representing how the policy computed the realized action.

2.3 Group-relative credit is response-level broadcast

For n responses to the same prompt, a canonical group-relative baseline is

$$b_i = \frac{R_i - \operatorname{mean}_j R_j}{\operatorname{std}_j R_j + \varepsilon}, \qquad \mathbf{A}^{(i)} = b_i \mathbf{1}_{T_i}. \tag{8}$$

Within a response, this is a rank-one broadcast geometry: every valid token receives the same outcome-derived coefficient. DAPO, Dr. GRPO, and GSPO materially improve clipping, normalization, sampling, or sequence-ratio behavior (Yu et al., 2025; Liu et al., 2025b; Zheng et al., 2025), but do not by themselves insert a trajectory-specific Transformer-computation variable into a Bellman/GAE transport kernel.

When a sampled group has identical rewards, the standardized group evidence may vanish; when rewards differ, the estimator still does not distinguish tokens that were pivotal from routine tokens. These properties help explain why fine-grained credit has become a major LLM-RL research direction (Zhang, 2026).

2.4 Architecture–credit mismatch

A causal Transformer generates each action through a realized internal computation graph. Two responses can have identical lengths, rewards, and group statistics yet route information through different historical tokens, heads, memories, or experts. An estimator that assigns them the same transport kernel discards this difference.

Definition 1 (Architecture–credit mismatch)**.** *A credit estimator has an architecture–credit mismatch when its transport operator is invariant to changes in the policy's realized internal computation that may alter which history is used to produce the optimized actions.*

This definition does not imply that internal computation is a faithful causal explanation. It identifies an unused information channel. Whether that channel improves learning is an empirical question requiring controls for critic capacity, position, marginal gate distribution, and training engineering.

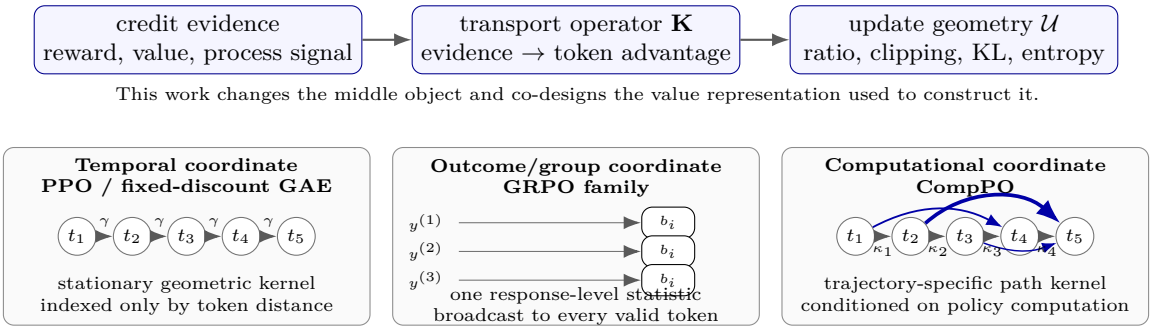


Figure 1: Credit assignment factorized into evidence, transport, and update geometry (top), with three transport coordinates used in LLM RL (bottom). Fixed GAE uses a stationary temporal kernel; group-relative methods broadcast response-level evidence; CompPO lets behavior-policy computation parameterize the causal transport path.

2.5 Three coordinates, not three mutually exclusive algorithms

Figure 1 visualizes temporal, group/outcome, and computational coordinates. They are not mutually exclusive sources of evidence. A process reward can be transported temporally or computationally; a group baseline can be combined with a token trace; a sequence-level importance ratio can consume an architecture-aware advantage. The contribution of this paper is to make the transport coordinate explicit and optimizable rather than treating fixed time or response broadcast as unavoidable defaults.

3 Computation-Conditioned Credit Transport

3.1 Policy-internal computation as an estimator variable

Let θ^- denote the behavior policy that generated the rollout. Its forward pass exposes a computation object $\mathcal{G}_{\theta^-}(h_t)$, such as attention, retrieval scores, memory access, expert routing, or cross-modal flow. A statistic Φ and bounded map g produce

$$z_t = \Phi(\mathcal{G}_{\theta^-}(h_t)), \qquad \kappa_t = g(z_t), \qquad 0 \le \kappa_t \le \bar{\kappa} < 1. \tag{9}$$

The gate is stored with the rollout and detached. The same realized κ_t is used throughout the optimization epochs for that batch. This behavior-policy conditioning is central: CCT does not differentiate through a policy-dependent task discount and does not redefine the environment's utility. It constructs a rollout-conditioned credit estimator.

3.2 Computation-conditioned bootstrap and trace

For a value function over a computation-aware state representation s_t^{C}, define

$$\delta_t^{\kappa} = r_t + \kappa_t V_\psi(s_{t+1}^{\mathrm{C}}) m_{t+1} - V_\psi(s_t^{\mathrm{C}}) m_t, \tag{10}$$

$$A_t^{\kappa,\lambda} = \delta_t^{\kappa} + \lambda \kappa_t A_{t+1}^{\kappa,\lambda} m_{t+1}. \tag{11}$$

We call Equation (11) Comp-GAE. The gate has two linked effects: it changes the one-step bootstrap target and the retention of all later residuals across the current transition.

With $\mathbf{D}_\kappa = \mathrm{diag}(\kappa_1, \dots, \kappa_T)$ and shift matrix $\mathbf{S}$, the valid-token recursion is

$$\begin{aligned} &\mathbf{A} = \boldsymbol{\delta} + \lambda \mathbf{D}_\kappa \mathbf{S} \mathbf{A}, \\ &\mathbf{A} = \mathbf{K}^{\kappa,\lambda} \boldsymbol{\delta}, \qquad \mathbf{K}^{\kappa,\lambda} = (\mathbf{I} - \lambda \mathbf{D}_\kappa \mathbf{S})^{-1}. \end{aligned} \tag{12}$$

Because $\mathbf{S}$ is nilpotent on a finite response, the inverse is a finite causal series. Its entries are

$$K_{t,u}^{\kappa,\lambda} = \mathbf{1}[u \ge t] \lambda^{u-t} \prod_{j=t}^{u-1} \kappa_j. \tag{13}$$

Thus the exact local sensitivity is

$$\frac{\partial A_t}{\partial \delta_u} = K_{t,u}^{\kappa,\lambda}. \tag{14}$$

This equation is the core conceptual object. Standard GAE uses one stationary kernel for all rollouts. CCT uses a policy-computation-conditioned causal kernel whose paths differ across positions and trajectories.

3.3 Latent continuation interpretation

An estimator-side latent variable $c_t \in \{0, 1\}$ can be used for intuition. If

$$\Pr(c_t = 1 \mid \mathcal{G}_{\theta^-}(h_t)) = \kappa_t, \tag{15}$$

then marginalizing c_t yields the one-step rule

$$(\mathcal{T}_\kappa V)(h_t) = \mathbb{E}[r_t + \kappa_t V(h_{t+1}) \mid h_t]. \tag{16}$$

This does not claim that the environment terminates with probability $1-\kappa_t$. It says that the estimator retains an expected fraction κ_t of downstream value across the current computational transition.

Transition-dependent discounting is well established as a way to specify task horizon or temporal abstraction (White, 2017; Harutyunyan et al., 2019); state-dependent discounting can also be learned with additional consistency controls (Wang et al., 2026). CCT differs in role and coupling: the external task remains fixed, the gate is read from behavior-policy computation, and the same coordinate is carried through the trace and value representation. We therefore use "retention gate" rather than treating κ_t as a new normative task discount.

3.4 Outer update and scope

The actor uses the unchanged clipped surrogate

$$\mathcal{L}_{\text{clip}}(\theta;\theta^-) = \mathbb{E}_t\big[\min\big(\rho_t(\theta)\widehat{A}_t, \operatorname{clip}(\rho_t(\theta),1-\epsilon,1+\epsilon)\widehat{A}_t\big)\big]. \tag{17}$$

where $\rho_t = \pi_\theta(y_t \mid h_t)/\pi_{\theta^-}(y_t \mid h_t)$ and $\widehat{A}_t$ is the normalized Comp-GAE estimate. Reward computation, rollout sampling, and KL regularization are unchanged.

This scope avoids two overclaims. First, CCT is not an attention-weighted policy loss: κ_t acts inside the bootstrap and trace before the policy objective is formed. Second, we do not claim an unbiased gradient theorem for a current-policy-dependent discounted-return objective. The implemented object is a detached, rollout-conditioned surrogate. Gate staleness across multiple update epochs is a real approximation analyzed as a limitation.

3.5 The general design space

The scalar gate studied here is the smallest member of a larger family. A computation-conditioned transport design specifies:

1. an internal computation object $\mathcal{G}$ and summary Φ;
2. a bounded transport map g or a graph-valued kernel;
3. a local evidence vector, such as TD residuals, process rewards, tree values, or counterfactual scores;
4. a transport-aligned baseline or critic; and
5. an outer update geometry, potentially including computation-conditioned clipping or KL.

Attention concentration is one deployable instantiation, not a claim of optimality. Future variants can use full attention-flow graphs, retrieval provenance, tool-observation edges, MoE expert assignments, recurrent memory gates, or modality-specific routing.

4 CompPO: An Attention-Routed Instantiation

4.1 Attention concentration as routing sharpness

For each valid response token t, let $\mathcal{I}_t$ be the set of valid historical positions exposed to the selected causal-attention layer and let $n_t = |\mathcal{I}_t|$. We aggregate final-layer GQA attention across query heads and renormalize over $\mathcal{I}_t$ to obtain $a_{ti} \geq 0$ with $\sum_{i\in\mathcal{I}_t} a_{ti} = 1$. The routing-sharpness statistic is the Herfindahl concentration

$$H_t = \sum_{i\in\mathcal{I}_t} a_{ti}^2, \qquad H_t \in [1/n_t, 1]. \tag{18}$$

Uniform attention gives $H_t = 1/n_t$ and a point mass gives $H_t = 1$. Because the raw lower bound changes with available history, we normalize it as

$$c_t = \frac{\log(n_t H_t)}{\log n_t} \in [0,1], \tag{19}$$

for $n_t > 1$, with $c_t = 1/2$ for the first valid position. This transformation maps the two position-dependent extremes exactly to zero and one while preserving ordering. It is not a learned position schedule.

The gate is

$$\kappa_t = \kappa_{\text{lo}} + (\kappa_{\text{hi}} - \kappa_{\text{lo}})\sigma(\tau(c_t - 1/2)), \tag{20}$$

with configured envelope parameters $(\kappa_{\text{lo}}, \kappa_{\text{hi}}, \tau) = (0.1, 0.9, 4)$ in all main experiments. Because the sigmoid has finite temperature, the attained range is the strict subinterval $[g(0), g(1)]$. Larger c_t means that the current policy computation is dominated by fewer historical inputs and therefore retains more downstream value across this transition.

The semantic claim is intentionally limited. H_t does not identify which token causally earned the reward. It summarizes the *shape* of upstream support for the current decision. The question answered by κ_t is "how much downstream value should cross this transition under the chosen transport model?", not "which past token is a causal explanation?" The shuffle and position-only controls in Section 7.3 test whether the realized signal carries information beyond a generic position-dependent gate.

4.2 Computation-aligned critic

The value estimator must predict targets generated by Equation (10) while remaining inexpensive enough for 16K-token rollouts. Rather than train a second Transformer, we reuse actor activations and add a small critic with fewer than 0.5% new parameters.

Cross-layer actor feature fusion. For a selected set of L' upper layers,

$$\bar{h}_t = \sum_{\ell=1}^{L'} \omega_\ell h_t^{(\ell)}, \qquad \omega = \text{softmax}(\eta), \tag{21}$$

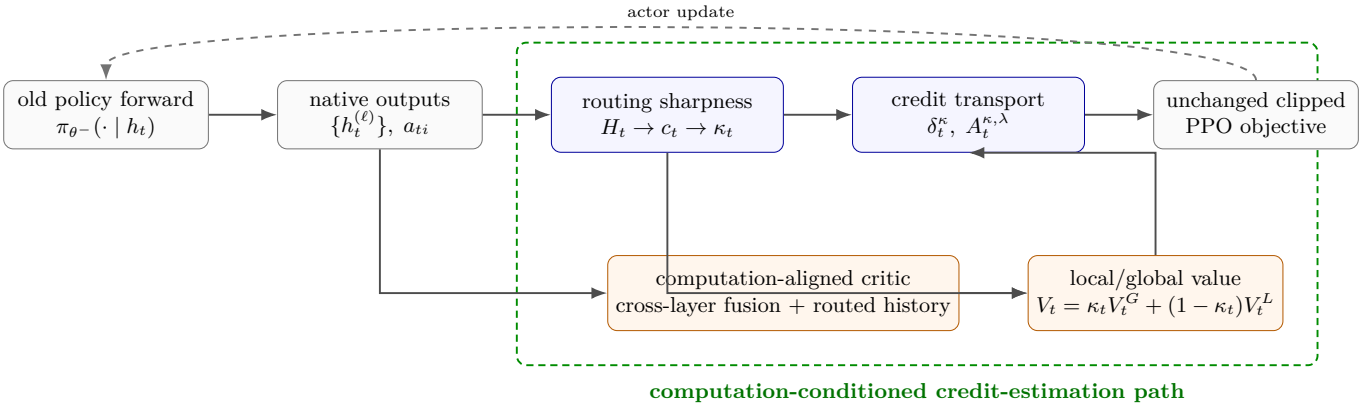


Figure 2: CompPO changes the credit-estimation path while preserving the task reward and outer clipped policy update. A native policy-computation statistic yields a detached gate κ_t. The same gate controls one-step bootstrap, eligibility-trace continuation, and the local/global geometry of a critic that reuses actor features.

so the critic can combine representations at different abstraction levels.

Value-refined history routing. Native attention was trained for next-token computation, not scalar value prediction. We therefore modulate each edge with a learned bilinear relevance gate,

$$\beta_{ti} = \frac{\langle W_{\beta,q}\bar{h}_t, W_{\beta,k}\bar{h}_i\rangle}{\sqrt{d}}, \tag{22}$$

$$\widetilde{a}_{ti} = a_{ti}\,\sigma(\beta_{ti}). \tag{23}$$

A second nonnegative token-relevance score

$$m_{ti} = \mathrm{ReLU}\left(\frac{\langle W_{r,q}\bar{h}_t, W_{r,k}\bar{h}_i\rangle}{\sqrt{d}}\right) \tag{24}$$

filters superficial routing correlations. The earlier version called this module "TokenMI"; because no explicit mutual-information bound is optimized, the present paper uses the operationally accurate name *token-relevance gate*. The final causal pooling score and routed-history representation are

$$s_{ti} = \widetilde{a}_{ti} m_{ti} \mathbf{1}[i \in \mathcal{I}_t], \tag{25}$$

$$h_t^G = \sum_{i\in\mathcal{I}_t} \frac{s_{ti}}{\sum_{j\in\mathcal{I}_t} s_{tj} + \varepsilon}\bar{h}_i. \tag{26}$$

Gate-aligned local/global value. Two lightweight heads produce a local value $V^L(\bar{h}_t)$ and a routed-history value $V^G(h_t^G)$. We combine them using the same transition regime that generated the bootstrap target:

$$V_\psi(s_t^{\mathrm{C}}) = \mathrm{clip}\big(\kappa_t V^G(h_t^G) + (1-\kappa_t)V^L(\bar{h}_t), -1, 1\big)\,. \tag{27}$$

This is a co-design hypothesis, not a universal theorem about all critics. Under concentrated routing, the value estimate relies more on explicitly pooled history; under diffuse routing, it relies more on the local actor state. The factorial evidence in Section 7.2 tests whether this coupling matters empirically while holding the stability stack fixed.

4.3 Actor and critic objectives

The actor uses Equation (17) with normalized CompGAE advantages. The critic regresses to the detached λ-return

$$\widehat{G}_t^{\kappa,\lambda} = A_t^{\kappa,\lambda} + V_\psi(s_t^{\mathrm{C}})m_t \tag{28}$$

with clipped value loss. The stored attention-derived gate is detached and cannot be changed by the critic loss within the current rollout batch. The critic consumes actor features from the same forward pass; all reported variants use the same feature-sharing convention so that the controlled contrasts isolate only the gate and critic architecture. The exact backbone-gradient routing is an implementation fact that must be copied from the released code and is flagged in the accompanying author-verification checklist.

4.4 Stability and systems implementation

History-conditioned bootstrapping introduces the familiar actor–critic risks of early value error and policy–critic co-divergence. The reported system uses three controls:

1. a fixed value-domain clamp $[-1, 1]$, matched to the bounded reward scale;
2. zero-output initialization and a brief critic warm-up, so early advantages are not dominated by arbitrary bootstrap predictions; and
3. a closed-loop phase controller that adjusts actor learning rate, clip range, value-loss clip, and critic multiplier only when jointly monitored stability statistics cross predefined thresholds.

The core 2×2 experiment holds all three controls identical across cells, so they cannot explain the gate-by-critic interaction.

The implementation avoids materializing the full $T \times T$ attention matrix: only the final layer exposes chunked attention; GQA heads are aggregated by group; buffers are restricted to response positions; online-softmax recomputation avoids dense HBM writes; and top-K causal pooling reduces critic work to $O(TK)$ with $K = 64$. These choices avoid a separate same-scale value Transformer and its optimizer state; the additional work is confined to attention extraction and the lightweight critic. Exact measured wall-clock, peak-memory, and throughput records are unavailable in the retained experiment logs. We therefore restrict the present efficiency claim to the algorithmic structure—actor-feature reuse, no second same-scale value Transformer, and $O(TK)$ routed pooling—rather than claiming an unmeasured runtime advantage.

5 Properties, Identifiability, and Claim Boundary

The results below condition on a fixed rollout batch. The gates were computed by π_{θ^-} and are detached, which is the regime implemented by the algorithm.

Proposition 1 (Strict reduction)**.** *If $\kappa_t = \gamma$ for every valid transition and the standard critic state is restored, then the computation-conditioned bootstrap, Comp-GAE, and clipped policy update reduce exactly to the standard fixed-γ GAE/PPO estimator path.*

The fixed-gate + aligned-critic ablation is intentionally not vanilla PPO: it preserves the proposed value representation while removing trajectory-specific transport. This is necessary to isolate the gate rather than replacing the entire system at once.

Proposition 2 (Causal transport kernel)**.** *For a finite response, Equation (11) has the unique solution*

$$A_t^{\kappa,\lambda} = \sum_{u=t}^{T} \lambda^{u-t} \left(\prod_{j=t}^{u-1} \kappa_j \right) \delta_u^{\kappa}. \tag{29}$$

Equivalently, $\mathbf{A} = \mathbf{K}^{\kappa,\lambda}\boldsymbol{\delta}$ with entries in Equation (13). If $\kappa_t = \gamma$, the kernel is stationary Toeplitz and equals Equation (7); otherwise it is generally path-dependent and nonstationary.

The proposition formalizes "let credit follow computation." It does not imply nonlocal token attribution: the method still transports evidence through a scalar sequential kernel. What changes is the local retention assigned to each transition, and therefore the sensitivity of every earlier advantage to every later residual.

Proposition 3 (Well-posed detached backup)**.** *For a fixed policy and a gate satisfying $0 \le \kappa_t \le \bar{\kappa} < 1$, the operator in Equation (16) is a sup-norm contraction with modulus at most $\bar{\kappa}$:*

$$\|\mathcal{T}_\kappa V - \mathcal{T}_\kappa W\|_\infty \le \bar{\kappa} \|V - W\|_\infty. \tag{30}$$

It therefore has a unique bounded fixed point.

This is a fixed-policy evaluation statement. It does not establish global convergence of PPO while the behavior policy, critic, and future gates change.

Proposition 4 (Bounded trace)**.** *If $|\delta_t^{\kappa}| \le D$ and $\lambda\bar{\kappa} < 1$, then*

$$|A_t^{\kappa,\lambda}| \le \frac{D}{1 - \lambda\bar{\kappa}}. \tag{31}$$

The trace is therefore controlled by the largest realized gate, not by sequence length directly. TD bootstrapping still provides local residuals: even with $r_t = 0$, δ_t^{κ} can be nonzero whenever the critic detects a value change. We do not claim that terminal reward is transmitted without decay through an arbitrarily long chain.

Transport identifiability. A performance gain from dynamic gates could arise from at least four explanations: a better average horizon, a useful position schedule, trajectory-specific alignment, or critic capacity. The experimental design explicitly separates them. The fixed gate is set to the empirical mean 0.61; a position-only control freezes the mean gate in 20 equal-width position bins; global shuffling preserves the gate values but destroys token correspondence; and the 2×2 factorial crosses transport with critic architecture. These controls do not establish causal token attribution, but they make the trajectory-specific transport claim falsifiable.

Old-policy dependence and staleness. The gate is computed from θ^- and reused during the actor epochs. As θ moves, its current internal computation may differ from the stored coordinate. This resembles other rollout-derived quantities in PPO but introduces a new source of estimator staleness. The present implementation uses two actor epochs and clipping; a future analysis should bound the error in terms of policy-ratio drift, gate drift, and critic error, or recompute the gate per epoch.

Configured envelope versus realized range. The attention instantiation uses the smooth map in Equation (20). With a finite temperature, the configured parameters $(\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}})$ are asymptotic envelope parameters; the actual mathematical range is $[g(0), g(1)] \subset (\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}})$. All propositions use the actual supremum $\bar{\kappa} = \sup_c g(c) < 1$. Descriptive gate statistics reported later focus on mean scale, dispersion, position, and outcome stratification.

What is not proved. We do not prove that attention is a causal explanation, that the chosen concentration statistic is optimal, that every standard critic fails, or that CompPO dominates tuned PPO in every regime. The theoretical contribution is the transport formulation and its fixed-batch properties; the algorithmic claims are supported by controlled experiments in the stated domain.

6 Experimental Design

The experiments are organized around five falsifiable questions rather than a flat list of baselines.

1. **Performance:** does computation-conditioned credit improve over a credible group-relative baseline under a matched search budget?
2. **Mechanism:** are the dynamic gate and computation-aligned critic jointly necessary, or is the result explained by either component alone?
3. **Signal specificity:** does trajectory-specific gate alignment matter beyond position or marginal regularization?
4. **Robustness:** does the method occupy a wider stable region than PPO in a matched stress grid?
5. **Transfer and baseline strength:** do the gains survive frozen evaluation, a second backbone, and a larger GRPO group?

6.1 Task, models, and evaluation

Training uses hard mathematical-reasoning prompts from DAPO-Math-17K with a maximum response length of 16K. The primary backbone is Qwen3-4B (Yang et al., 2025); Llama-3.1-8B-Instruct (AI at Meta, 2024) is used as a transfer control. A deterministic answer extractor supplies terminal correctness reward. Malformed reasoning tags receive a negative terminal reward, and the length penalty is applied only to correct, format-valid responses.

The in-distribution (ID) development/validation set contains 500 held-out problems and is distinct from frozen MATH500. We use the following endpoint conventions throughout:

- **Best**: maximum ID validation accuracy within the predeclared training budget;
- **Final**: the last checkpoint (step 200 for the performance and factorial experiments; step 150 for the PPO stress test);
- **Last-20**: mean over the final 20 checkpoints;
- **AULC**: validation accuracy integrated over training steps and divided by the observed step range.

Final is the primary endpoint. Frozen evaluation always uses the final checkpoint and is never consulted for hyperparameter selection.

Frozen benchmarks are AIME25, AIME2024, MATH500, GSM8K, LiveBench, GPQA, and an Olympiad-style set. We report greedy pass@1 and majority@8. Macro-averages are unweighted across benchmarks and should be interpreted alongside per-benchmark values, especially because AIME sets contain only 30 items.

6.2 Common training configuration

All primary Qwen runs use batch size 4, four responses per prompt, PPO/GRPO minibatch size 4, 16K maximum response length, BF16, AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, zero weight decay, gradient clipping at 1, two update epochs, and sampling temperature/top-p of 1.0/0.7. The actor learning-rate candidates are $\{5{\times}10^{-7}, 10^{-6}, 2{\times}10^{-6}\}$ and KL candidates are $\{0, 10^{-3}, 2 \times 10^{-3}\}$ for both GRPO and CompPO. The selected GRPO configuration is $(10^{-6}, 10^{-3})$; the selected CompPO configuration is $(10^{-6}, 2 \times 10^{-3})$. The CompPO critic uses learning rate 10^{-5}, $\lambda = 0.95$, configured gate-envelope parameters $(0.1, 0.9)$ with $\tau = 4$, and value clamp $[-1, 1]$.

The same $3{\times}3$ grid is screened once on ID validation for each method. The selected configuration is then rerun with five independent seeds $\{17, 42, 123, 256, 2026\}$. The submitted seed is 42. At every learning rate, the selected-KL CompPO endpoint exceeds the selected-KL GRPO endpoint; full grids are reported in Section B.

6.3 Core coefficient-by-critic factorial

We cross two transport rules with two critic classes:

Transport	Critic	Interpretation
fixed $\kappa = 0.61$	standard	mean-matched fixed-coordinate control
κ_t / Comp-GAE	standard	dynamic transport without aligned critic
fixed $\kappa = 0.61$	aligned	aligned critic without trajectory gate
κ_t / Comp-GAE	aligned	full CompPO

The constant 0.61 equals the empirical mean gate of full CompPO. In the fixed/aligned cell, it replaces κ_t both in the GAE recursion and in the local/global value mixture, removing trajectory dependence while preserving all other architecture and stability components. All four cells share the same Qwen checkpoint, hard-DAPO prompt stream, 500-problem ID set, data seeds, response budget, reward, optimizer, KL/clip configuration, stability system, and 200-step budget.

We report policy endpoints and critic diagnostics on a shared held-out token sample: explained variance (EV), Spearman correlation with TD targets, value MAE, and clamp saturation. The seed-matched factorial interaction is

$$\mathcal{I} = \text{Full} - (\text{Gate+Std}) - (\text{Fixed+Aligned}) + (\text{Fixed+Std}). \tag{32}$$

A positive $\mathcal{I}$ indicates complementarity beyond additive main effects.

6.4 Gate controls

We use two negative controls.

Global shuffle. Attention and gates are computed normally, then realized κ_t values are randomly permuted across valid response positions before both Comp-GAE and critic mixing. This preserves the global marginal distribution but destroys token-wise alignment.

Position-only schedule. For each seed, full-training rollout gates are averaged within 20 equal-width bins of normalized response position t/T. The schedule is frozen and used in both Comp-GAE and critic mixing; no validation or frozen benchmark data are used to construct it. This preserves the mean positional profile but removes trajectory-specific variation.

We also report the realized gate distribution from the final checkpoints, including mean, median, interquartile range, response-position thirds, and correct/incorrect trajectories.

6.5 Matched PPO stress grid

The PPO comparison is a robustness stress test, not a claim of universally tuned superiority. PPO and CompPO are evaluated for 150 steps over actor learning rates $\{2\times10^{-7}, 5\times10^{-7}, 10^{-6}\}$ and KL coefficients $\{0, 10^{-3}\}$, with two seeds per cell. No annealing, extra entropy/KL constraint, critic warm-up, or adaptive controller is used. Both methods use the same fixed value-domain bound $[-1, 1]$; separate no-clamp checks show that it acts only during a comparable early critic transient.

A run is declared stable when final accuracy is at least the 42.4% base accuracy, peak-to-final drop is below 15 points, and there is no sequence of five consecutive post-peak checkpoints below base. We report stable runs, mean best/final, and peak-to-final drop. The permitted inference is a wider stable region in this tested grid; it is not that PPO cannot be stabilized by method-specific tuning.

6.6 Stronger GRPO group control

The submitted GRPO uses four prompts and four responses, i.e., 16 trajectories per update. We additionally evaluate a requested 8×8 single-run GRPO control with 64 trajectories per update and all other settings matched. This result is a point estimate without training-seed confidence intervals; it tests baseline sensitivity rather than replacing the five-seed matched-budget comparison.

7 Results

7.1 Five-seed performance against GRPO

The selected GRPO configuration obtains 56.2% best accuracy (95% CI [55.6, 56.9]) and 53.8% final accuracy [52.9, 54.7]. CompPO obtains 61.7% best [61.3, 62.1] and 61.4% final [60.8, 62.0]. Every CompPO seed exceeds every GRPO seed at both endpoints. The method's best-to-final drop is only 0.3 point, compared with 2.4 points for GRPO.

The difference is not an isolated selected cell. At actor learning rates 5×10^{-7}, 10^{-6}, and 2×10^{-6}, each method's preferred-KL endpoint yields CompPO–GRPO gains of +6.8/+7.2, +5.4/+7.6, and +6.8/+8.0 best/final points, respectively. The complete screen appears in Table 4. This pattern argues against a single lucky learning-rate/KL choice.

7.2 The gate and critic form a coupled mechanism

Figure 3b and Table 1 summarize the controlled 2×2. A fixed mean-matched gate with a standard critic reaches 52.6% final. Replacing only the trace by Comp-GAE raises the result to 55.2%, but the tested standard critic still has negative EV (−0.08), weak TD-target Spearman correlation (0.16), and never reaches the submitted GRPO peak of 56.4%. Replacing only the critic raises final accuracy to 56.4%, with EV 0.45, but remains five points below full CompPO. The complete method reaches 61.4%, EV 0.52, and TD-target Spearman 0.57.

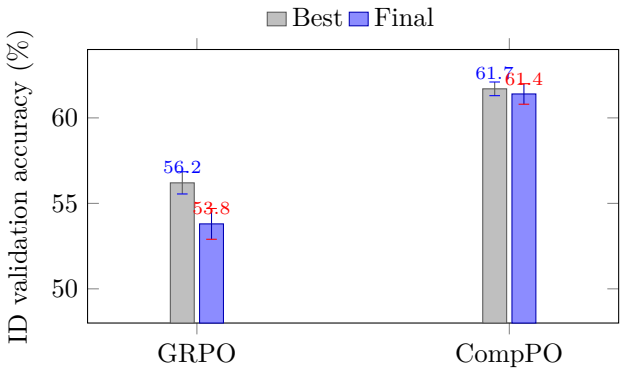
Seed-matched differences sharpen the attribution. Full CompPO exceeds Comp-GAE +standard critic by +5.7 best and +6.2 final points, and exceeds fixed-gate+aligned critic by +4.5 and +5.0. The interaction in Equation (32) is +1.9[1.3, 2.4] for best and +2.4[1.9, 2.9] for final. Thus the components are super-additive under the controlled configuration: the aligned critic is not a free-standing replacement for the gate, and the gate cannot realize its full benefit through the tested standard value head.

The critic diagnostics support the proposed mechanism rather than merely restating endpoint accuracy. Standard-critic cells have negative EV, large MAE, and high clamp saturation. Aligned-critic cells explain target variation, rank TD targets more faithfully, and saturate less. We therefore interpret the standard critic result as an empirical target/representation mismatch in this setting, not as a theorem that every conventional critic must fail.

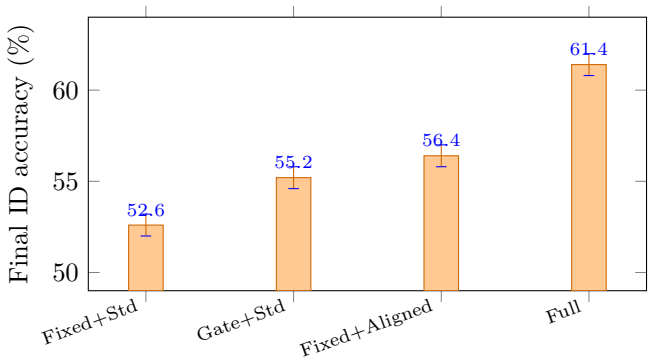
7.3 Trajectory-specific alignment matters

Global shuffling reduces best/final accuracy from 61.7/61.4 to 59.4/58.7, with a 2.7-point final penalty. The position-only schedule recovers part of the effect, reaching 60.2/59.7. It exceeds the matched fixed gate by $+3.0/+3.3$ best/final points and exceeds global shuffle by $+0.8/+1.0$, showing that the mean positional profile is informative. Nevertheless, full CompPO exceeds position-only by a paired +1.5[1.3, 1.7] best and +1.7[1.5, 1.8] final points. The remaining gain therefore depends on trajectory-specific variation rather than position alone.

The realized gate is not constant. Across five final checkpoints, its mean/median are 0.61/0.63 and its interquartile range is 0.18. Mean gates in the early, middle,



(a) Five-seed selected configurations; bars show means and 95% Student-t intervals.



(b) Gate-by-critic factorial. The final interaction is +2.4 points [1.9, 2.9].

Figure 3: Main training evidence. CompPO improves both best and final endpoints over the selected GRPO baseline, and the factorial experiment shows that the trajectory-specific gate and aligned critic are complementary rather than interchangeable.

Variant	Best	Final	Last-20	AULC	Peak–final	EV	TD ρ_s	MAE	Clamp sat.	Step to 56.4
Fixed gate + standard critic	53.4	52.6	52.0	47.4	0.8	−0.15	0.11	0.41	0.24	never
Comp-GAE + standard critic	56.0	55.2	54.7	50.2	0.8	−0.08	0.16	0.37	0.19	never
Fixed gate + aligned critic	57.2	56.4	55.9	53.1	0.8	0.45	0.52	0.20	0.05	112
Full CompPO	**61.7**	**61.4**	**60.8**	**58.9**	**0.3**	**0.52**	**0.57**	**0.18**	**0.03**	**33**

Table 1: Five-seed gate-by-critic factorial. Best and final means have 95% Student-t intervals: Fixed+Std 53.4[52.9, 53.9]/52.6[52.0, 53.2]; Comp-GAE +Std 56.0[55.6, 56.5]/55.2[54.6, 55.8]; Fixed+Aligned 57.2[56.6, 57.8]/56.4[55.8, 57.0]; Full 61.7[61.3, 62.1]/61.4[60.8, 62.0].

and late response thirds are 0.52/0.63/0.68, and correct trajectories have higher mean gate than incorrect trajectories (0.64 versus 0.57). These descriptive differences do not establish causality; they show that the retained coordinate varies meaningfully across positions, trajectories, and outcomes.

7.4 A wider stable region than PPO in the tested grid

Across 12 runs, PPO is stable in 3/12, with mean best/final/drop 47.7/34.4/13.3. CompPO is stable in 10/12, with 52.9/45.6/7.3. The largest difference occurs at aggressive learning rates: at 10^{-6} with KL 10^{-3}, PPO obtains 49.8/22.6 best/final while CompPO obtains 57.2/44.6; in the originally submitted cell, the corresponding single-seed endpoints were PPO 47.6/15.0 and CompPO 58.4/44.2.

The grid also bounds the conclusion. At the conservative 2×10^{-7} learning rate, PPO is stable in one of two seeds for each KL value and can finish above base. At 10^{-6} with no KL, one CompPO cell is also unstable. The evidence therefore supports a wider tested stable region and lower average degradation, not universal dominance over a method-specifically tuned PPO.

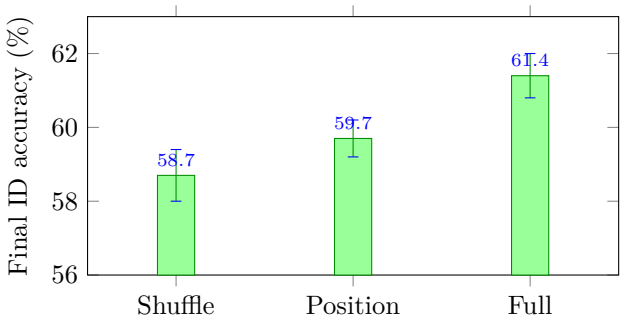
Backbone	Base	GRPO	CompPO	Δ
Qwen3-4B, greedy	58.2	57.5	**61.7**	+4.3
Qwen3-4B, majority@8	69.7	68.7	**71.0**	+2.3
Llama-3.1-8B, greedy	31.1	39.0	**42.8**	+3.9
Llama-3.1-8B, majority@8	36.5	47.8	**52.0**	+4.2

Table 2: Unweighted frozen-benchmark macro-averages. Δ is CompPO minus GRPO. These are final-checkpoint point estimates; complete per-benchmark results are in Tables 10 and 11.

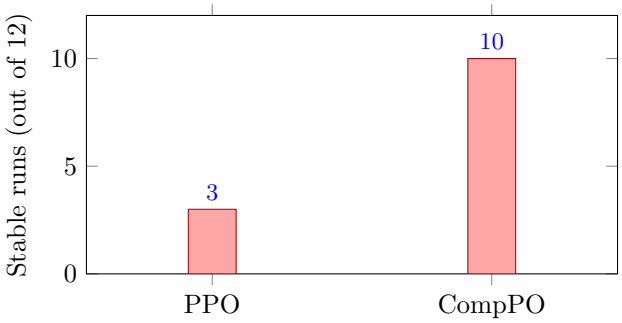
7.5 Frozen transfer across benchmarks and backbones

On Qwen3-4B, CompPO exceeds GRPO by 4.3 greedy and 2.3 majority@8 macro points. AIME25 greedy is the exception: GRPO scores 50.0 while CompPO and base score 43.3; majority@8 ties at 66.7. The strongest Qwen gains occur on AIME2024 (+16.7 greedy), GPQA (+11.6), LiveBench (+3.0), and MATH500 (+3.2). On Llama-3.1-8B-Instruct, all seven greedy differences favor CompPO, for a +3.9 macro improvement.

The Qwen GRPO frozen greedy macro is 0.7 below base despite a +11.4-point final improvement on ID validation. This is a small transfer trade-off, not failed optimization. The Llama control further shows that GRPO is a legitimate learning baseline: it improves frozen greedy macro from 31.1 to 39.0.



(a) Gate controls. Full uses trajectory-specific alignment.



(b) Matched LR–KL stress grid; this is a bounded robustness claim, not tuned PPO ranking.

Figure 4: Specificity and robustness controls. Destroying token-wise alignment or replacing it with a fixed positional profile reduces performance, while CompPO occupies a wider stable region in the tested PPO grid.

7.6 Stronger prompt–group GRPO control

Increasing GRPO from 4×4 to 8×8 improves ID best/final from 56.4/53.8 to 58.0/56.2 and frozen greedy macro from 57.5 to 60.2. The corresponding submitted-seed CompPO reference is 61.8/61.4 ID and 61.7 frozen macro. Thus the stronger group closes the frozen gap to 1.5 points but leaves descriptive gaps of +3.8 best and +5.2 final ID points. Because this control has one training run, we present it as baseline-sensitivity evidence rather than a significance-tested comparison.

Summary. The evidence chain is consistent across levels: matched tuning establishes performance; the factorial isolates coupling; shuffled and position-only controls isolate trajectory-specific alignment; target metrics support the critic interpretation; the stress grid bounds robustness; and frozen/transfer controls show that the effect is not confined to a single ID curve. No individual experiment establishes universal generality, but together they support computation-conditioned credit as a substantive algorithmic variable rather than an attention-themed engineering add-on.

8 Related Work

Fixed and transition-dependent credit in RL. TD learning, eligibility traces, GAE, and PPO provide the classical basis for the estimator used here (Schulman et al., 2015, 2017). Generalized RL task formulations allow transition-dependent discounting (White, 2017), and option-level work has studied per-decision discount factors as temporal-abstraction and bias–variance controls (Harutyunyan et al., 2019). AdaGamma, developed concurrently, learns state-dependent discounts with a return-consistency objective and analyzes the resulting operator (Wang et al., 2026). CompPO differs in interpretation and source: the task reward is unchanged, the gate is a detached statistic of the LLM policy's own computation, and the contribution is its joint use in bootstrap, trace, and critic representation.

LLM policy optimization. PPO-based RLHF trains actor and value models with clipped policy updates and GAE (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022). RLOO and GRPO remove the learned value model through leave-one-out or group-relative baselines (Ahmadian et al., 2024; Shao et al., 2024). DAPO, Dr. GRPO, and GSPO improve large-scale training through asymmetric clipping, dynamic sampling, bias correction, loss normalization, or sequence-level importance ratios (Yu et al., 2025; Liu et al., 2025b; Zheng et al., 2025). These methods primarily change baseline statistics, sampling, or update geometry. Computation-conditioned credit changes the Bellman/advantage transport path and can in principle be combined with those advances.

Fine-grained LLM credit. Process supervision supplies step-level evidence but requires labels or learned reward models (Uesato et al., 2022; Lightman et al., 2023); PRIME learns implicit process rewards online from outcome labels (Cui et al., 2025). VinePPO estimates intermediate state values through additional continuations and documents severe weaknesses of conventional value networks on reasoning trajectories (Kazemnejad et al., 2024). VC-PPO attributes long-CoT PPO collapse to value initialization and reward-signal decay and uses value pretraining plus decoupled GAE (Yuan et al., 2025). TEMPO constructs prefix trees and provides critic-free TD corrections at branching points (Tran et al., 2025); GRPO-λ introduces a critic-free eligibility-trace approximation (Parthasarathi et al., 2025). OAR reshapes group advantages using token perturbation or gradient-based outcome influence (Li et al., 2026c). High-entropy-token methods update only likely reasoning forks (Wang et al., 2025). OPPO derives a Bayesian token-level value recursion using an oracle-conditioned likelihood ratio (Li et al., 2026b). S-trace introduces critic-free selective eligibility

traces (Mou et al., 2026). These works reinforce the importance of nonuniform token credit, but use external rollouts, branch statistics, outcome influence, entropy, oracle scores, or critic-free trace surrogates rather than one policy-internal statistic coupled across bootstrap, trace, and critic. A recent survey organizes this rapidly expanding landscape across reasoning and agentic settings (Zhang, 2026).

Critics for LLM reasoning. The difficulty and cost of LLM value models motivated the shift toward critic-free methods. Recent work reopens this design space from complementary directions. AsyPPO uses lightweight prompt-sharded mini-critics (Liu et al., 2025a); Generative Actor–Critic replaces one-shot scalar prediction with a reasoning critic and in-context conditioning (Shan et al., 2026); VC-PPO improves initialization and target construction (Yuan et al., 2025). CompPO instead asks whether the critic's *state geometry* matches the policy-internal transport rule. Its factorial result supports the narrower claim that the tested standard head does not track the computation-conditioned target, while an aligned actor-feature critic does.

Attention and internal signals as optimization primitives. Attention is a central Transformer mechanism (Vaswani et al., 2017), but its faithfulness as a causal explanation is debated (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019). We use it as an observable routing signal, not as proof of causal attribution. HICRA identifies attention-based planning and anchor tokens and targets their updates (Li et al., 2025); Reinforced Attention Learning directly optimizes internal attention policies for multimodal grounding (Li et al., 2026a). FlowTracer, posted after the initial submission, builds an answer-targeted attention-flow DAG and uses throughput for reward shaping (Dong et al., 2026). These methods are close in spirit because they treat internal computation as useful for RL. The distinctive object in CompPO is the *transition-level credit-retention gate*: the same native statistic parameterizes the one-step value backup, eligibility trace, and critic's local/global state representation while leaving reward and outer policy loss unchanged.

Novelty boundary. We therefore do not claim the first use of attention or the first token-level credit method. The contribution is the coupled construction and the broader abstraction it instantiates: policy-internal computation can be an explicit coordinate of Bellman-style credit transport. Concurrent work makes this boundary more, not less, important by showing that internal-signal credit is becoming a distinct research axis.

9 Discussion: Toward Architecture-Aware RL

9.1 The scientific claim is larger than an attention heuristic

The most conservative reading of CompPO is a new credit estimator that improves long-chain mathematical RL on two backbones. The more consequential reading is that it exposes a design variable that conventional LLM RL leaves implicit. A Transformer is not merely a large black-box policy with token actions; it computes a trajectory-specific internal routing structure at every decision. Once that structure is made available to the estimator, fixed temporal distance and response-level outcome are no longer the only viable coordinates for transporting credit.

This perspective changes the algorithm-design question from

> "Which black-box policy optimizer should be applied to an LLM?"

to

> "Which internal computation of this policy should parameterize its value transport, baseline, trust region, and update?"

The latter is an architecture-aware RL question. Attention concentration is one answer, chosen because it is native, bounded after normalization, inexpensive to extract, and falsifiable through shuffling. It is unlikely to be the final answer.

9.2 Reinterpreting the apparent failure of actor–critic LLM RL

The rise of critic-free methods has often been read as evidence that value learning is a poor fit for LLM reasoning. The evidence in this paper suggests a more specific diagnosis. The standard critic in the factorial experiment fails to track the computation-conditioned target: EV is negative, TD-target correlation is low, and policy gains stall. The aligned critic tracks the target and enables the full algorithm. This does not prove that one architecture is universally necessary, but it supports a field-level hypothesis:

> Conventional LLM critics may fail partly because their representation and training target are mismatched to the policy's history routing, not because token-level actor–critic learning is intrinsically unsuitable.

This distinction matters. If the critic-free turn is driven by a contingent representation failure, then improved value models, generative critics, retrieval-aware critics, and computation-aligned heads may recover the sample-efficiency benefits of TD learning without a second same-scale model.

Family	Evidence or baseline	Transport geometry	Update or value mechanism	Relation to this work
PPO/GAE	learned scalar value	stationary temporal trace	token-ratio clipping; critic	fixed-kernel ancestor
GRPO/RLOO	group or leave-one-out outcome	response-level broadcast	critic-free token update	outcome-coordinate control
DAPO / Dr. GRPO / GSPO	improved sampling and normalization	mainly response/group credit	clipping, loss normalization, or sequence ratios	orthogonal optimizer advances
PRIME/PRM	process or implicit step evidence	reward placement plus downstream estimator	process/reward model	complementary evidence axis
VinePPO / TEMPO	continuation or prefix-tree values	rollout- or tree-based TD	Monte Carlo or nonparametric value	alternative fine-grained transport
OAR / S-trace / GRPO-λ	influence, entropy, or trace proxy	reshaping or selective traces	critic-free update	neighboring token-credit methods
CompPO	native policy computation	path-dependent Bellman/GAE kernel	actor-feature TAC; clipped PPO	coupled architecture-aware transport

Table 3: Positioning by the evidence–transport–update factorization. Categories identify the primary intervention, not an exclusive taxonomy; many systems combine multiple axes.

9.3 An orthogonal primitive rather than another monolithic optimizer

LLM RL algorithms often combine several axes in one recipe. It is useful to separate them:

- **reward evidence**: outcome verifiers, process rewards, preference models, or judges;
- **sampling and baseline**: groups, leave-one-out statistics, dynamic sampling, or trees;
- **credit transport**: fixed GAE, counterfactual values, tree TD, oracle recursion, or computation-conditioned traces;
- **update geometry**: token/sequence ratios, clipping, KL, entropy, and normalization;
- **systems**: rollout scheduling, kernels, memory, and parallelism.

CompPO intervenes primarily in the third axis and co-designs the critic representation needed by that intervention. This makes it potentially composable with DAPO-style clipping, GSPO-style sequence ratios, process rewards, prefix-tree values, or larger group baselines. The present experiments deliberately keep the outer objective unchanged to isolate the new axis.

9.4 From interpretability signals to optimization primitives

Most uses of attention ask whether it explains a prediction. Computation-conditioned credit asks a different question: whether an internal signal can improve the estimator used to train the policy. The burden of proof is therefore empirical alignment, not human interpretability. A signal may be useful for optimization even when it is not a faithful natural-language explanation, provided negative controls show that its trajectory-specific structure matters and its failure modes are bounded.

This shift extends beyond attention. Future policies expose increasingly explicit internal mechanisms:

- MoE models choose experts and routing weights;
- retrieval-augmented models select documents and spans;
- agents read memory, tool observations, and environment states;
- multimodal models route between text, image, audio, and video tokens;
- state-space and recurrent architectures expose gates and memory updates.

Each mechanism can define a candidate computational coordinate. A mature architecture-aware RL stack may use different coordinates for value transport, exploration, KL control, and critic state construction.

9.5 From a scalar gate to a computational Bellman graph

The scalar κ_t is intentionally minimal: it compresses the geometry of the upstream computation into one transition coefficient. A richer extension would transport credit directly over a graph. Let G_t contain historical tokens or latent states as nodes and normalized policy-computation weights as edges. A graph Bellman operator could propagate value along multiple nonlocal routes, while a graph critic could estimate node- or subgraph-level value. Multi-head traces could maintain separate horizons for specialist heads; hierarchical traces could operate at token, reasoning-step, and tool-call scales. The present result should be interpreted as feasibility evidence for this larger program, not as proof that scalar concentration is optimal.

9.6 Where the current evidence stops

The experiments establish the mechanism on mathematical reasoning, two Transformer backbones up to 8B, terminal verifiers, and 16K responses. They do not establish gains in dialogue, tool use, code, retrieval, multimodal trajectories, dense-reward tasks, or larger models. Those

domains are motivated because their dependencies are sparse and nonlocal, but they remain untested.

Attention concentration is a structural proxy. Attention sinks, positional bias, induction heads, and output-layer calibration can corrupt it. The learned relevance gate and negative controls reduce but do not eliminate these risks. The gate is also stale across actor epochs, and the current theory does not provide a global policy-improvement bound for policy-generated gates. Finally, the stronger 8×8 GRPO control and frozen benchmark values are point estimates, while exact measured hardware and wall-clock records are unavailable. These limitations motivate targeted follow-up rather than weakening the central observation that policy-internal computation can carry useful credit information.

10 Conclusion

LLM policies compute with history, but their credit estimators usually do not. This paper identifies that architecture–credit mismatch and introduces computation-conditioned credit: a detached statistic of the policy's own internal computation controls how downstream value crosses a transition. CompPO instantiates the idea with attention concentration, a path-dependent GAE trace, and a computation-aligned actor-feature critic, while preserving the external reward and clipped PPO update.

The empirical evidence supports the mechanism at several levels. Under matched tuning and five seeds, CompPO substantially exceeds GRPO; a controlled gate-by-critic factorial reveals a positive interaction and better critic target tracking; shuffle and position-only controls show that trajectory-specific alignment matters; a PPO stress grid shows a wider stable region; and frozen evaluation transfers across Qwen3-4B and Llama-3.1-8B-Instruct. The contribution is therefore not that attention is universally causal or that one optimizer replaces all others. It is that internal computation can become a first-class variable in RL credit transport. This opens a broader agenda in which attention, routing, retrieval, memory, and modality structure shape Bellman operators, eligibility traces, critics, and trust regions for the policies that actually use them.

References

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. *arXiv preprint arXiv:2402.14740*, 2024.

AI at Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ganqu Cui, Lifan Yuan, Zefan Wang, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Zhichen Dong, Yang Li, Yuhan Sun, et al. How does reasoning flow? tracing attention-induced information flow for targeted rl in llms. *arXiv preprint arXiv:2606.10646*, 2026.

Anna Harutyunyan, Peter Vrancx, Philippe Hamel, Ann Nowe, and Doina Precup. Per-decision option discounting. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2644–2652, 2019.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, 2019.

Amirhossein Kazemnejad et al. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024.

Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. Reinforced attention learning. *arXiv preprint arXiv:2602.04884*, 2026a.

Yang Li, Zhichen Dong, Yuhan Sun, et al. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *arXiv preprint arXiv:2510.13554*, 2025.

Yu Li et al. Oracle-guided proximal policy optimization for fine-grained credit assignment. *arXiv preprint arXiv:2605.21851*, 2026b.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. *arXiv preprint arXiv:2601.07408*, 2026c.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jiashun Liu, Johan Obando-Ceron, Han Lu, Yancheng He, Weixun Wang, Wenbo Su, Bo Zheng, Pablo Samuel Castro, Aaron Courville, and Ling Pan. Asymmetric proximal policy optimization: Mini-critics boost llm reasoning. *arXiv preprint arXiv:2510.01656*, 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.

Chaoli Mou, Zhan Zhuang, Xinning Chen, and Yu Zhang. Beyond uniform credit assignment: Selective eligibility traces for rlvr. *arXiv preprint arXiv:2605.05965*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. Grpo-λ: Credit assignment improves llm reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

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

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

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of ACL*, 2019.

Zikang Shan, Han Zhong, Liwei Wang, and Li Zhao. Bringing value models back: Generative critics for value modeling in llm reinforcement learning. *arXiv preprint arXiv:2604.10701*, 2026.

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.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Hieu Tran, Zonghai Yao, and Hong Yu. Exploiting tree structure for credit assignment in rl training of llms. *arXiv preprint arXiv:2509.18314*, 2025.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Shenzhi Wang, Le Yu, Chang Gao, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Yaomin Wang, Jianting Pan, Ran Tian, Xiaoyang Li, Yu Zhang, Hengle Qin, and Tianshu Yu. Adagamma: State-dependent discounting for temporal adaptation in reinforcement learning. *arXiv preprint arXiv:2605.06149*, 2026.

Martha White. Unifying task specification in reinforcement learning. *arXiv preprint arXiv:1609.01995*, 2017.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, 2019.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. *arXiv preprint arXiv:2503.01491*, 2025.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

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.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2020.

A Full Algorithm

Algorithm 1 Computation-Conditioned Policy Optimization (CompPO)

Require: actor π_θ, reference policy π_{ref}, aligned critic V_ψ, verifier R, gate parameters $(\kappa_{\text{lo}}, \kappa_{\text{hi}}, \tau)$, GAE λ, clip ϵ

for rollout iteration $k = 1, 2, \ldots$ **do**
  Set behavior parameters $\theta^- \leftarrow \theta$
  Sample prompts and responses $y_{1:T} \sim \pi_{\theta^-}$; compute rewards $r_{1:T}$ and valid-token masks $m_{1:T}$
  for each valid response token t **do**
    Extract final-layer GQA-group attention a_{ti} over valid history $\mathcal{I}_t$ and actor hidden states $\{h_t^{(\ell)}\}$
    Renormalize a_{ti} over $i \in \mathcal{I}_t$; set $n_t = |\mathcal{I}_t|$
    $H_t \leftarrow \sum_{i \in \mathcal{I}_t} a_{ti}^2$
    $c_t \leftarrow \log(n_t H_t)/\log n_t$ if $n_t > 1$, else $1/2$
    $\kappa_t \leftarrow \kappa_{\text{lo}} + (\kappa_{\text{hi}} - \kappa_{\text{lo}})\sigma(\tau(c_t - 1/2))$
    Detach and store κ_t, attention summaries, and required actor features
    Fuse actor features $\bar{h}_t \leftarrow \sum_\ell \omega_\ell h_t^{(\ell)}$
    Compute value-refined scores s_{ti} by Equations (23) and (24)
    Pool routed history h_t^G by Equation (26)
    Evaluate $V_t \leftarrow \text{clip}(\kappa_t V^G(h_t^G) + (1 - \kappa_t) V^L(\bar{h}_t), -1, 1)$
  Set $A_{T+1} = 0$ and $V_{T+1} = 0$
  for $t = T, T-1, \ldots, 1$ **do**
    $\delta_t \leftarrow r_t + \kappa_t V_{t+1} m_{t+1} - V_t m_t$
    $A_t \leftarrow \delta_t + \lambda \kappa_t A_{t+1} m_{t+1}$
    $G_t \leftarrow A_t + V_t m_t$
  Normalize valid-token advantages $\widehat{A}_t$
  for each optimization epoch **do**
    Update actor with clipped objective in Equation (17) plus reference-policy KL
    Update critic by clipped regression to detached G_t
    Apply the predeclared stability phase controller if enabled

Gradient convention. The behavior-policy attention and κ_t are detached for the whole rollout batch. The critic consumes actor features from the same forward computation and cannot alter the detached, stored transport gate for the current batch. The exact backbone-gradient routing must match the released implementation; all controlled variants use the same convention. The outer actor update consumes detached normalized advantages, as in ordinary PPO.

Fixed-gate controls. For a fixed-gate ablation, every occurrence of κ_t in the TD residual, trace, and value mixture is replaced by the same constant. The main factorial uses $\kappa = 0.61$, the empirical mean of the full method, to avoid confounding trajectory dependence with mean scale.

B Complete Experimental Tables

B.1 Common tuning screen

Method	Actor LR	KL	Best	Final
GRPO	5e−7	0	52.8	51.2
GRPO	5e−7	1e−3	54.0	52.4
GRPO	5e−7	2e−3	53.2	51.8
GRPO	1e−6	0	54.6	52.8
GRPO	1e−6	1e−3	**56.4**	**53.8**
GRPO	1e−6	2e−3	55.0	53.2
GRPO	2e−6	0	53.8	51.6
GRPO	2e−6	1e−3	54.2	52.0
GRPO	2e−6	2e−3	53.0	51.4
CompPO	5e−7	0	59.8	58.4
CompPO	5e−7	1e−3	60.4	59.2
CompPO	5e−7	2e−3	60.8	59.6
CompPO	1e−6	0	60.6	59.0
CompPO	1e−6	1e−3	61.2	60.2
CompPO	1e−6	2e−3	**61.8**	**61.4**
CompPO	2e−6	0	59.6	58.0
CompPO	2e−6	1e−3	60.8	59.4
CompPO	2e−6	2e−3	61.0	60.0

Table 4: Single-seed common 3×3 screen used to select the five-seed configurations. Frozen benchmarks were not used.

B.2 Five-seed confirmation

Seed	GRPO Best	GRPO Final	CompPO Best	CompPO Final
17	55.6	52.9	61.2	60.8
42	56.4	53.8	61.8	61.4
123	55.8	53.4	61.5	61.2
256	56.6	54.2	61.9	61.6
2026	56.8	54.7	62.1	62.0
Mean	56.2	53.8	61.7	61.4
95% CI	[55.6,56.9]	[52.9,54.7]	[61.3,62.1]	[60.8,62.0]

Table 5: Five-seed selected-configuration confirmation.

B.3 Per-seed factorial results

	Fixed+Std		Comp-GAE +Std		Fixed+Aligned		Full	
Seed	Best	Final	Best	Final	Best	Final	Best	Final
17	52.8	51.9	55.6	54.6	56.6	55.7	61.2	60.8
42	53.8	53.1	56.6	55.8	57.8	56.9	61.8	61.4
123	53.2	52.4	55.8	54.9	57.0	56.1	61.5	61.2
256	53.6	52.8	56.2	55.4	57.4	56.6	61.9	61.6
2026	53.6	52.8	56.0	55.3	57.2	56.7	62.1	62.0

Table 6: Seed-level gate-by-critic factorial endpoints.

B.4 Gate controls and descriptive statistics

Variant	Best	Final	Last-20	AULC
Global shuffle	59.4 [58.9,59.9]	58.7 [58.0,59.4]	58.1	57.2
Position-only	60.2 [59.8,60.6]	59.7 [59.2,60.2]	–	–
Full CompPO	61.7 [61.3,62.1]	61.4 [60.8,62.0]	60.8	58.9

Table 7: Five-seed gate controls. Brackets are 95% Student-t intervals.

Statistic	Value	Statistic	Value
Mean / median	0.61 / 0.63	IQR	0.18
Early / middle / late	0.52 / 0.63 / 0.68	Correct / incorrect	0.64 / 0.57

Table 8: Realized gate statistics from five final full-method checkpoints. Archived tail and boundary-proximity summaries are excluded pending reconciliation of the exact implementation map with the printed smooth-map equation.

B.5 Matched PPO stress grid

Method	Actor LR	KL	Best	Final	Stable seeds
PPO	2e−7	0	48.2	44.8	1/2
PPO	2e−7	1e−3	49.0	46.2	1/2
PPO	5e−7	0	46.8	31.0	0/2
PPO	5e−7	1e−3	48.4	43.6	1/2
PPO	1e−6	0	44.2	18.4	0/2
PPO	1e−6	1e−3	49.8	22.6	0/2
CompPO	2e−7	0	52.4	48.6	2/2
CompPO	2e−7	1e−3	53.0	49.2	2/2
CompPO	5e−7	0	51.6	47.0	2/2
CompPO	5e−7	1e−3	52.2	47.8	2/2
CompPO	1e−6	0	50.8	36.4	0/2
CompPO	1e−6	1e−3	57.2	44.6	2/2

Table 9: Two seeds per LR–KL cell. Across 12 runs, PPO is stable in 3 and CompPO in 10.

B.6 Frozen Qwen3-4B evaluation

		Greedy pass@1			Majority@8		
Benchmark	N	Base	GRPO	CompPO	Base	GRPO	CompPO
AIME25	30	43.3	50.0	43.3	66.7	66.7	66.7
AIME2024	30	63.3	53.3	70.0	76.7	73.3	80.0
MATH500	500	67.2	66.6	69.8	79.8	79.2	81.0
GSM8K	1319	91.7	93.1	94.1	93.5	94.3	94.5
LiveBench	200	46.5	50.5	53.5	61.0	61.5	67.0
GPQA	198	47.0	40.4	52.0	57.6	54.5	55.6
Olympiad	674	48.5	48.2	49.4	53.0	51.6	52.1
Macro	–	58.2	57.5	**61.7**	69.7	68.7	**71.0**

Table 10: Frozen Qwen3-4B final-checkpoint point estimates.

B.7 Frozen Llama-3.1-8B-Instruct evaluation

	Greedy pass@1			Majority@8		
Benchmark	Base	GRPO	CompPO	Base	GRPO	CompPO
AIME25	3.3	13.3	16.7	6.7	23.3	30.0
AIME2024	13.3	20.0	23.3	16.7	33.3	36.7
MATH500	53.4	64.2	69.6	64.4	76.6	80.6
GSM8K	81.3	86.4	87.6	85.6	90.4	91.2
LiveBench	22.5	35.0	39.0	29.0	44.5	49.0
GPQA	26.8	32.8	40.4	32.3	39.4	47.0
Olympiad	17.1	21.1	23.3	21.1	26.7	29.8
Macro	31.1	39.0	**42.8**	36.5	47.8	**52.0**

Table 11: Frozen Llama-3.1-8B-Instruct final-checkpoint point estimates.

B.8 Stronger GRPO group control

Method	Trajectories/update	ID Best	ID Final	Frozen greedy macro
GRPO 4×4	16	56.4	53.8	57.5
GRPO 8×8	64	58.0	56.2	60.2
CompPO reference	16	61.8	61.4	61.7

Table 12: Single-run stronger-group control. The CompPO row is the submitted-seed reference.

C Reproducibility Details

C.1 Primary configuration

Item	GRPO	CompPO
Backbone	Qwen3-4B	same
Transfer backbone	Llama-3.1-8B-Instruct	same
Maximum response	16K	16K
Batch / responses / minibatch	4 / 4 / 4	same
Actor learning rate	10^{-6}	10^{-6}
Critic learning rate	–	10^{-5}
KL coefficient	10^{-3}	2×10^{-3}
Policy clip	0.20	0.20
Credit estimator	group-relative	κ_t backup; $\lambda = 0.95$
Gate map	–	configured envelope $(0.1, 0.9)$, $\tau = 4$
Value clamp	–	$[-1, 1]$
Optimizer	AdamW $(0.9, 0.999)$	same
Weight decay / grad clip	0 / 1	same
Update epochs	2	2
Sampling temperature / top-p	1.0 / 0.7	same
Precision	BF16	BF16

Table 13: Selected primary settings. Both methods were screened on the same actor-LR/KL grid.

C.2 Phase controller

The full performance experiments and all four factorial cells use the same closed-loop controller. It changes only after the predeclared stability trigger persists; escalation requires two consecutive trigger events and de-escalation requires five clear steps.

Phase	KL	Clip	Actor LR	Value-loss clip	Critic LR mult.
Stable	0.0020	0.20	10^{-6}	0.5	1.0
Warning	0.0035	0.16	7×10^{-7}	0.4	0.9
Strong	0.0045	0.14	5×10^{-7}	0.3	0.7
Hard-stop	0.0070	0.09	2×10^{-7}	0.2	0.5

Table 14: Adaptive stability phases. The same controller is held fixed across the factorial, so it cannot induce the gate-by-critic interaction.

The monitored signals are policy KL, gradient norm, clip fraction, entropy, and ID validation accuracy. Because validation participates in phase control, the 500-problem set is a development/held-out-ID set rather than a final test set. Frozen benchmarks are never used by the controller, checkpoint selection, or hyperparameter search.

Controller inputs and hysteresis. The controller maintains exponential moving averages with coefficient 0.4 for policy clip fraction, PPO KL, KL loss, and gradient norm, and coefficient 0.2 for entropy. It also reads raw response-clip ratio and ID development accuracy when evaluated. Entropy decline is measured against the oldest four values in an eight-step window. The target phase is the maximum severity triggered by the following rules:

- entropy EMA below 8.0/7.5/6.5 requests warning/strong/hard-stop;
- clip fraction at least 0.14/0.18/0.25 requests warning/strong/hard-stop;

- PPO KL at least 0.08/0.12/0.22 requests warning/strong/hard-stop;
- KL loss at least 2.85/3.20/4.00 requests warning/strong/hard-stop;
- gradient norm at least 40/80 requests warning/strong;
- development accuracy drop from the running best of at least 0.04/0.08/0.20 requests warning/strong/hard-stop;
- an entropy drop of at least 1.5/3.0, when clip fraction or PPO KL is already in warning territory, requests strong/hard-stop; and
- response-clip ratio at least 0.65 requests hard-stop.

Escalation requires two consecutive steps with a higher target phase; de-escalation requires five consecutive clear steps. Fresh runs start in the stable phase. On fresh phase-0 runs, steps 25–40 linearly interpolate stable to warning knobs. If a screened launch KL differs from 0.002, later phases preserve the same multipliers 1.75×, 2.25×, and 3.5× relative to that launch value.

Additional phase-controlled quantities. The stable/warning/strong/hard-stop phases use advantage clips 5.0/4.0/3.5/2.5, actor gradient clips 1.0/0.8/0.7/0.5, actor epochs 2/2/2/1, and critic gradient clips 1.0/0.8/0.7/0.5. A reward schedule, applied in parallel rather than triggered by phase, ramps the format-penalty coefficient from 0.2 to 1.0 over steps 0–40; the length penalty is enabled at step 20 and ramps from 3×10^{-5} to 8×10^{-5} over steps 20–60. The PPO stress experiment disables critic warm-up and this controller; the fixed value-domain clamp remains enabled for the applicable value path.

C.3 Value clamp control

The clamp is a fixed output-domain constraint, not a method-specific adaptive intervention. In the most collapse-prone cell (Qwen3-4B, actor LR 10^{-6}, KL 0), PPO and CompPO were each repeated three times without the clamp. The last out-of-range critic prediction occurred around step 37 for PPO and step 32 for CompPO on average. In the bounded grid, activation was therefore confined to comparable early critic transients and then became the identity. The stress-grid comparison applies the same bound to both methods.

C.4 Attention extraction and memory

The implementation uses final-layer GQA-group attention. Only the final layer exposes chunked attention; earlier layers retain fused attention kernels. Response-window-only buffers avoid prompt-heavy storage, online softmax/recomputation avoids writing dense attention to HBM, and prior-guided causal top-K pooling uses $K = 64$. The resulting critic adds fewer than 0.5% trainable parameters and does not require a second same-scale Transformer or a second full optimizer state.

Exact GPU model/count, wall-clock time, peak allocated/reserved HBM, generated-token count, and throughput were not retained in the experiment record available for this version. They cannot be reconstructed reliably from public model specifications and are therefore not estimated. Consequently, the present paper does not make a measured wall-clock or peak-memory superiority claim.

C.5 Statistical conventions

Training intervals are two-sided 95% Student-t intervals over five independent runs. The same seed set is used across factorial cells, enabling seed-matched interaction intervals. Best is reported for completeness, but final is primary and last-20/AULC are secondary. Frozen results are currently point estimates from final checkpoints; they reflect finite benchmark sampling and checkpoint variance that is not quantified by the training-seed intervals.

C.6 Data and result availability

Editable CSV versions of every numerical table are included with the source archive. The public code release should additionally include the rollout configuration, evaluator scripts, checkpoint-selection rules, and commands needed to reproduce the training and frozen-evaluation tables.

D Proofs and Limiting Cases

D.1 Strict reduction

If $\kappa_t = \gamma$ for all valid t, then Equation (16) becomes

$$(\mathcal{T}_\kappa V)(h_t) = \mathbb{E}[r_t + \gamma V(h_{t+1}) \mid h_t] = (\mathcal{T}_\gamma V)(h_t).$$

Substitution into Equations (10) and (11) yields the standard TD residual and GAE recursion in Equation (5). The outer objective Equation (17) is unchanged by construction, so the fixed-gate estimator path recovers the standard GAE/PPO path when the critic is also restored to the standard control.

D.2 Contraction of the detached operator

For bounded functions V, W and any history h_t,

$$\begin{aligned} |\mathcal{T}_\kappa V(h_t) - \mathcal{T}_\kappa W(h_t)| &= |\mathbb{E}\left[\kappa_t (V(h_{t+1}) - W(h_{t+1})) \mid h_t\right]| \\ &\le \mathbb{E}\left[\kappa_t |V(h_{t+1}) - W(h_{t+1})| \mid h_t\right] \\ &\le \kappa_{\mathrm{hi}} \|V - W\|_\infty. \end{aligned}$$

Taking the supremum over h_t proves the contraction. Banach's fixed-point theorem gives a unique bounded fixed point for a fixed policy and detached gate.

D.3 Bounded trace

Unrolling Equation (11) gives Equation (29). If $|\delta_t^\kappa| \le D$ and $\kappa_t \le \kappa_{\mathrm{hi}}$, then

$$\begin{aligned} |A_t^{\kappa,\lambda}| &\le \sum_{\ell=0}^{T-t} \lambda^\ell \left(\prod_{j=0}^{\ell-1} \kappa_{t+j} \right) D \\ &\le D \sum_{\ell=0}^{\infty} (\lambda \kappa_{\mathrm{hi}})^\ell = \frac{D}{1 - \lambda \kappa_{\mathrm{hi}}}. \end{aligned}$$

D.4 Length normalization of attention concentration

For a probability vector over n_t valid historical positions,

$$\frac{1}{n_t} \le H_t = \sum_i a_{ti}^2 \le 1,$$

where the lower bound follows from Cauchy–Schwarz and is achieved by the uniform vector, while the upper bound is achieved by a point mass. Therefore $n_t H_t \in [1, n_t]$, so

$$0 \le \frac{\log(n_t H_t)}{\log n_t} \le 1.$$

At uniform attention, $n_t H_t = 1$ and $c_t = 0$; at a point mass, $n_t H_t = n_t$ and $c_t = 1$. Because logarithm is monotone, the transformation preserves the ordering of H_t at each n_t while aligning the extremes across positions.

D.5 Why the coefficient is not a post-hoc loss weight

Multiplying a policy loss by an importance score changes the magnitude of the gradient at token t but leaves the value target and propagation path unchanged. In CompPO, κ_t enters both terms that define the estimator:

$$\delta_t^\kappa = r_t + \kappa_t V_{t+1} - V_t, \qquad A_t^{\kappa,\lambda} = \delta_t^\kappa + \lambda \kappa_t A_{t+1}^{\kappa,\lambda}.$$

It therefore changes the one-step bootstrap and every later residual's retention weight in Equation (29). Its additional use in Equation (27) changes the state representation used to estimate the corresponding target. The policy loss consumes the resulting advantage without a direct multiplicative attention weight.

E Limitations, Open Questions, and Broader Impact

Attention is a proxy, not causal attribution. Concentration can be distorted by attention sinks, positional biases, delimiter tokens, induction heads, or final-layer calibration. The method does not establish that highly attended tokens caused the outcome. A future version should compare layers and heads, mask known sinks, test intervention-based relevance, and use graph-valued rather than scalar signals.

Policy dependence and staleness. The gate is computed by the behavior policy and held fixed through multiple actor epochs. This is operationally simple and avoids differentiating the estimator through attention, but it creates staleness as the actor moves. Theoretical work should quantify the interaction among gate drift, PPO ratios, critic error, and clipping; empirical work should compare recomputation, fewer epochs, and explicit consistency regularization.

Choice of concentration statistic. Herfindahl concentration is bounded, monotone, inexpensive, and length-normalizable, but it is one point in a larger space. Shannon/Rényi entropy, top-k coverage, head-specialist statistics, multi-layer aggregation, learned bounded maps, and attention-flow measures may be stronger. The present paper establishes that one trajectory-specific policy-internal statistic is useful, not that it is optimal.

Reward quality. Computation-conditioned transport can amplify a bad reward just as easily as a good one. Checker false negatives, formatting artifacts, or answer conventions may cause the algorithm to transport spurious evidence. Because the gate is correlated with the policy computation, reward artifacts that themselves dominate attention could be reinforced more strongly. Robust verifiers and audit sets remain essential.

Scope. The current evidence is mathematical reasoning on Qwen3-4B and Llama-3.1-8B-Instruct, with sparse terminal reward and responses up to 16K. Dialogue, tools, code, retrieval, multimodal generation, agents, dense rewards, and larger models are not experimentally established. The framework applies syntactically to these settings, but usefulness depends on whether the chosen internal statistic captures value-relevant computation.

Statistical and systems limitations. Training uncertainty is reported for the principal Qwen comparisons, but frozen evaluations and the 8×8 GRPO control are point estimates. Exact hardware, wall-clock, throughput, and peak-memory records are unavailable, so the current evidence does not support a measured systems-efficiency comparison.

Potential positive impact. More accurate credit transport may reduce rollout and training requirements, improve the viability of small actor-feature critics, and make long-horizon reasoning or agent training more sample efficient. The architecture-aware framing may also encourage modular combinations of credit estimators with better sampling, verifiers, and update geometry.

Potential risks. The method is a general optimizer for capable generative models and can improve downstream systems with both beneficial and harmful uses. Internal-signal optimization may also create new forms of reward hacking: a policy could learn attention or routing patterns that manipulate the estimator without improving task behavior. Detached gates and current controls do not rule out such co-adaptation over many training iterations. Future deployments should monitor gate distributions, test counterfactual stability, retain external behavioral evaluation, and avoid treating internal routing as a safety guarantee.