LET CREDIT FOLLOW COMPUTATION: ARCHITECTURE-AWARE CREDIT TRANSPORT FOR LLM REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-language-model reinforcement learning has substantially improved reward modeling, rollout sampling, baselines, and policy-update geometry, but it usually leaves one component architecture-agnostic: the operator that transports delayed evidence into token-level credit. Fixed-discount generalized advantage estimation (GAE) applies the same geometric kernel to every trajectory, while group-relative methods broadcast a response-level statistic to all tokens. We identify this discrepancy as an *architecture–credit mismatch* and introduce computation-conditioned credit transport (CCT), in which a detached statistic of the behavior policy's internal computation parameterizes a trajectory-specific causal kernel. Its concrete instantiation, COMPPO, maps native attention concentration to a bounded retention gate, uses the gate in both the one-step bootstrap and a path-dependent advantage trace (COMP-GAE), and co-designs a 14.4M transport-aligned critic (TAC) that reuses actor hidden states and the policy KV cache. The external reward and clipped PPO objective remain unchanged, and a constant gate recovers fixed-coefficient GAE. Across five Qwen3-4B seeds, COMPPO reaches 61.4% final held-out accuracy (95% CI $[60.8, 62.0]$), versus 53.8% $[52.9, 54.7]$ for GRPO from the same 3×3 screen, under the same 4×4 trajectory budget, with a best-to-final gap of 0.3 versus 2.4 points. Under that shared budget, COMPPO reaches GRPO's 56.4% peak at step 33. A controlled gate-by-critic factorial yields a $+2.4$-point final interaction $[1.9, 2.9]$; shuffling and position-only controls show that trajectory-specific alignment contributes beyond mean scale and position. COMPPO is stable in $10/12$ runs in a matched PPO stress grid versus $3/12$ for PPO, and improves frozen greedy macro accuracy over GRPO by 4.2 and 3.8 points on Qwen3-4B and Llama-3.1-8B-Instruct. These results establish policy-internal computation as a useful variable for RL credit estimation and motivate architecture-aware Bellman operators, traces, and critics.

1 INTRODUCTION

Reinforcement learning with verifiable or preference-derived rewards is a central route to stronger LLM reasoning and decision making. The relevant methods have progressed rapidly: learned-value PPO supplies token-level bootstrapping (Schulman et al., 2017; 2016); critic-free objectives use response groups or leave-one-out baselines (Ahmadian et al., 2024; Shao et al., 2024); process supervision and counterfactual rollouts densify feedback (Lightman et al., 2024; Cui et al., 2025; Kazemnejad et al., 2025); and recent optimizers improve clipping, normalization, sampling, and importance-ratio geometry (Yu et al., 2025; Liu et al., 2025b; Zheng et al., 2025). Yet these developments usually leave the policy's internal computation unused when converting delayed evidence into token-level advantages.

That abstraction is increasingly restrictive in long trajectories. A causal Transformer does not process every token transition in the same way: different heads retrieve different parts of the prefix, and the realized dependency structure varies across examples. A late proof step may depend on an early substitution, a tool action on a distant observation, or a code token on an earlier declaration. The policy computation is therefore history-structured and trajectory-specific, whereas the credit estimator is commonly architecture-agnostic. Fixed-discount GAE (Schulman et al., 2016) transports value

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.

according only to token distance. Group-relative methods assign one outcome-derived statistic to every token in a response. We call this discrepancy the **architecture–credit mismatch**.

We formalize the mismatch by separating a policy-gradient pipeline into three objects:

$$\underbrace{\mathbf{e}}_{\text{credit evidence}} \xrightarrow{\mathbf{K}} \underbrace{\mathbf{A}}_{\text{token credit}} \xrightarrow{\mathcal{U}} \underbrace{\Delta\theta}_{\text{policy update}} . \tag{1}$$

Credit evidence includes terminal rewards, process rewards, judges, rollout estimates, and TD residuals. The transport operator $\mathbf{K}$ determines how that evidence influences individual decisions. The update geometry $\mathcal{U}$ contains likelihood ratios, clipping, KL penalties, entropy, and normalization. This decomposition distinguishes complementary advances: a better verifier changes evidence, GRPO changes the response-level baseline (Shao et al., 2024), and GSPO changes update geometry (Zheng et al., 2025); none uniquely determines how evidence should be transported through the policy's realized computation.

Our central hypothesis is that the policy's internal computation can provide a useful coordinate for credit transport. We introduce **computation-conditioned credit transport** (CCT), in which a detached statistic of the behavior policy parameterizes a causal kernel. With $D_\kappa = \mathrm{diag}(\kappa)$ and forward shift S,

$$\mathbf{K}^{\kappa,\lambda} = (I - \lambda D_\kappa S)^{-1}, \qquad K^{\kappa,\lambda}_{t,u} = \mathbf{1}[u \geq t]\lambda^{u-t} \prod_{j=t}^{u-1} \kappa_j, \tag{2}$$

rather than the stationary GAE operator $(I - \lambda\gamma S)^{-1}$. Two same-length responses can therefore transport the same later residual differently when the behavior policy computes different gates. The residuals themselves also change, because κ_t enters the one-step bootstrap.

We instantiate the framework as **Computation-Conditioned Policy Optimization** (CompPO). Native attention concentration is converted into a bounded retention gate. The same gate enters the one-step value bootstrap, a path-dependent generalized-advantage trace (Comp-GAE), and the local/global mixture of a transport-aligned critic (TAC) that reuses actor hidden states and routing information. The external reward, rollout interface, reference-policy KL, and clipped PPO objective remain unchanged. Setting the gate to a constant recovers fixed-coefficient GAE; restoring the standard critic recovers the standard estimator path.

The gate changes both the local TD target and the influence of every later residual on earlier advantages. This differs from placing a post-hoc attention weight on the policy loss. Concentration is a structural proxy for routing rather than a causal attribution, and its usefulness must persist under controls for critic capacity, position, and the marginal gate distribution.

We evaluate these claims on long-horizon mathematical reasoning with responses up to 16K tokens. Under a shared 3×3 search and five-seed confirmation, CompPO reaches 61.4% final ID accuracy versus 53.8% for GRPO on the same 4×4 trajectory budget, with a 14.4M critic rather than a second 4B value model. The five-seed ID curves separate under this matched update budget, so the comparison is a sample-efficiency difference rather than a longer run. A gate-by-critic factorial, shuffle and position-only controls, a PPO stress grid, and two-backbone frozen evaluation test whether the gain is the coupled transport mechanism rather than critic size, mean horizon, extra rollouts, or one selected cell.

Contributions.

- We identify the architecture–credit mismatch and formulate credit assignment as evidence, transport, and update geometry, exposing transport as an independent LLM-RL design axis.
- We define CCT as the path-dependent kernel $\mathbf{K} = (I - \lambda D_\kappa S)^{-1}$, with reduction to fixed GAE, a retention (not task-discount) interpretation, contraction, and bounded-trace properties.
- We introduce CompPO, coupling an attention-derived retention gate, Comp-GAE, and a 14.4M actor-feature TAC (0.36%) that reuses the actor KV cache, remaining in GRPO's training-memory class while avoiding a full-scale PPO critic.
- We provide matched tuning, five-seed endpoint estimates, a gate-by-critic factorial, critic diagnostics, shuffle and position-only controls, a PPO stress grid, a stronger GRPO group control, and two-backbone frozen evaluation, showing a sample-efficiency gain under a shared trajectory budget.

2 Related Work

LLM policy optimization. PPO-based RLHF combines a learned value function, GAE, clipped policy updates, and reference-model regularization (Schulman et al., 2016; 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). RLOO and GRPO remove the learned critic through leave-one-out or group-relative response baselines (Ahmadian et al., 2024; Shao et al., 2024). DAPO, Dr. GRPO, and GSPO respectively refine dynamic sampling and clipping, remove biased length/std normalization, and replace token-level importance ratios with sequence-level ones (Yu et al., 2025; Liu et al., 2025b; Zheng et al., 2025). These methods primarily modify baseline statistics, sampling, or update geometry. CCT instead changes the Bellman/GAE transport path and can in principle be combined with those improvements; CompPO is one instance.

Fine-grained credit for reasoning. Process supervision supplies intermediate evidence but requires labels or learned reward models (Uesato et al., 2022; Lightman et al., 2024); PRIME learns implicit process rewards from outcomes (Cui et al., 2025). VinePPO estimates intermediate values with extra continuations (Kazemnejad et al., 2025); TEMPO uses prefix trees, while GRPO-λ, OAR, OPPO, and S-trace construct critic-free traces or token scores from λ-returns, outcome influence, oracle-conditioned likelihoods, or selective eligibility (Tran et al., 2025; Parthasarathi et al., 2025; Li et al., 2026d;c; Mou et al., 2026). High-entropy-token methods focus updates on likely reasoning forks (Wang et al., 2025). These approaches, surveyed by Zhang (2026), establish that response-uniform credit is often inadequate. CCT differs by letting one policy-internal coordinate parameterize the bootstrap, the eligibility-trace path, and—in the CompPO instance—the critic state.

Critics and transition-dependent continuation. Conventional LLM critics can be expensive and inaccurate on partial reasoning states, motivating critic-free algorithms. Recent work revisits value modeling through value pretraining, lightweight critics, or generative critics (Yuan et al., 2025; Liu et al., 2025a; Shan et al., 2026). Classical RL formulations permit transition-dependent continuation or discounting (White, 2017; Harutyunyan et al., 2019); AdaGamma learns state-dependent discounts with a consistency objective (Wang et al., 2026). In CCT the task reward is unchanged and the gate is detached from the behavior policy. CompPO additionally uses the same coordinate in the value representation.

Attention and internal signals. Attention is a central Transformer routing mechanism (Vaswani et al., 2017), but its faithfulness as causal explanation remains debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019). Attention-based preplan-and-anchor credit, Reinforced Attention Learning, and FlowTracer use internal routing for token reweighting, attention-policy optimization, or flow-based reward shaping (Li et al., 2026b;a; Dong et al., 2026). We likewise treat internal computation as useful for optimization, but not as causal ground truth. The distinctive object in CCT is a transition-level retention gate for the value bootstrap and GAE path kernel. CompPO additionally couples that gate to an aligned critic while leaving the external reward and outer policy loss unchanged. Table 6 compares these families along evidence, transport, and update axes.

3 Computation-Conditioned Credit Transport

Setup. Let x be a prompt and $y_{1:T}$ a response generated by behavior policy π_{θ^-}, with history $h_t = (x, y_{<t})$. Let $\mathbf{e}$ collect local evidence and $\mathbf{A}$ the token advantages consumed by an update $\mathcal{U}$. We write $\mathbf{A} = \mathbf{K}\mathbf{e}$ with causal $\mathbf{K}$. Standard GAE (Schulman et al., 2016) uses

$$\delta_t^\gamma = r_t + \gamma V(h_{t+1}) m_{t+1} - V(h_t) m_t, \tag{3}$$

$$A_t^{\gamma,\lambda} = \delta_t^\gamma + \gamma\lambda A_{t+1}^{\gamma,\lambda} m_{t+1}, \tag{4}$$

or, stacking residuals and writing S for the forward shift $(Sv)_t = v_{t+1}$,

$$\mathbf{A} = (I - \lambda\gamma S)^{-1}\boldsymbol{\delta}, \qquad K_{t,u}^{\gamma,\lambda} = \mathbf{1}[u \geq t](\lambda\gamma)^{u-t}. \tag{5}$$

The kernel is stationary Toeplitz: it depends only on token distance. A canonical group-relative estimator assigns response i a normalized outcome b_i and sets $\mathbf{A}^{(i)} = b_i\mathbf{1}$. Within a response this is a rank-one broadcast; if group rewards coincide, the standardized evidence can vanish. Neither geometry contains a variable for how the policy computed the realized actions. These three

coordinates are complementary rather than exclusive: process evidence can be transported temporally or computationally, and a sequence-level update can consume an architecture-aware advantage.

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 assume that internal computation is a faithful causal explanation. It identifies an available policy-side signal whose usefulness must be tested empirically.

Computation-conditioned kernel. Let $\mathcal{G}_{\theta^-}(h_t)$ be a behavior-policy computation object and

$$z_t = \Phi(\mathcal{G}_{\theta^-}(h_t)), \qquad \kappa_t = g(z_t), \qquad 0 \le \kappa_t \le \bar{\kappa} < 1. \tag{6}$$

The gate is detached and stored with the rollout. COMPPO uses the simplest scalar summary Φ—attention concentration; other summaries define other CCT estimators and are not evaluated here. The computation-conditioned backup and trace are

$$\delta_t^{\kappa} = r_t + \kappa_t V(h_{t+1}) m_{t+1} - V(h_t) m_t, \tag{7}$$

$$A_t^{\kappa,\lambda} = \delta_t^{\kappa} + \lambda \kappa_t A_{t+1}^{\kappa,\lambda} m_{t+1}. \tag{8}$$

With $D_\kappa = \mathrm{diag}(\kappa_1, \ldots, \kappa_T)$,

$$\mathbf{A} = \boldsymbol{\delta} + \lambda D_\kappa S \mathbf{A} = \mathbf{K}^{\kappa,\lambda} \boldsymbol{\delta}, \qquad \mathbf{K}^{\kappa,\lambda} = (I - \lambda D_\kappa S)^{-1}. \tag{9}$$

Because S is nilpotent on a finite response, the inverse is a finite causal series with entries

$$K_{t,u}^{\kappa,\lambda} = \mathbf{1}[u \ge t] \lambda^{u-t} \prod_{j=t}^{u-1} \kappa_j = \frac{\partial A_t}{\partial \delta_u}. \tag{10}$$

Fixed-discount GAE is the special case $D_\kappa = \gamma I$. The gate is an estimator-side retention coefficient: it does not redefine the environment reward or the normative task discount.

Retention, not termination. A latent $c_t \in \{0, 1\}$ with $\Pr(c_t = 1 \mid \mathcal{G}_{\theta^-}(h_t)) = \kappa_t$ yields $(\mathcal{T}_\kappa V)(h_t) = \mathbb{E}[r_t + \kappa_t V(h_{t+1}) \mid h_t]$ after marginalization. The environment does not terminate with probability $1 - \kappa_t$. The estimator retains an expected fraction κ_t of downstream value across the current computational transition. Transition-dependent discounts can specify task horizon (White, 2017; Harutyunyan et al., 2019; Wang et al., 2026). CCT keeps the external task fixed, reads the gate from behavior-policy computation, and carries the same coordinate through the trace (and, in COMPPO, the critic).

Design space. A CCT estimator specifies (i) a computation object $\mathcal{G}$ and summary Φ; (ii) a bounded map g or a graph-valued kernel; (iii) a local evidence vector (TD residuals, process rewards, or tree values); (iv) a transport-aligned baseline or critic; and (v) an outer update. Attention concentration instantiates (i)–(ii). We fix (v) to clipped PPO (Schulman et al., 2017) to isolate the transport axis.

Properties and claim boundary.

Proposition 1 (Strict reduction)**.** *If $\kappa_t = \gamma$ for every valid transition and the same value representation is used, Equations* (7) *and* (8) *reduce exactly to fixed-discount TD residuals and GAE.*

Proposition 2 (Causal kernel)**.** *On a finite response, Equation* (8) *has the unique solution Equation* (10)*, or equivalently $\mathbf{A} = \mathbf{K}^{\kappa,\lambda}\boldsymbol{\delta}$ with Equation* (9)*. If $\kappa_t = \gamma$ the kernel is stationary Toeplitz; otherwise it is path-dependent.*

Proposition 3 (Contraction and bounded trace)**.** *For a fixed policy and detached gates with $\kappa_t \le \bar{\kappa} < 1$, $\mathcal{T}_\kappa$ is a $\bar{\kappa}$-contraction in the sup norm. If $|\delta_t^{\kappa}| \le D$ and $\lambda\bar{\kappa} < 1$, then $|A_t^{\kappa,\lambda}| \le D/(1 - \lambda\bar{\kappa})$.*

Proofs are in Section D. The propositions concern a fixed detached batch. Empirically, a dynamic-gate gain could still arise from mean horizon, a position schedule, token alignment, or critic capacity; Section 5 separates those explanations. We do not claim an unbiased current-policy gradient or global improvement: the implemented object is a detached, rollout-conditioned surrogate. The configured envelope $(\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}})$ is asymptotic, so the contraction uses the attained $\bar{\kappa} = \sup g < 1$. A descriptor that does not pass these controls lies outside the claim.

4 CompPO

4.1 Attention concentration as a retention gate

Figure 1 shows the resulting architecture. For valid response token t, let $\mathcal{I}_t$ be the historical positions exposed to the selected causal-attention layer and $n_t = |\mathcal{I}_t|$. Final-layer GQA attention is aggregated across query heads and renormalized over $\mathcal{I}_t$ to obtain $a_{ti} \geq 0$ with $\sum_{i \in \mathcal{I}_t} a_{ti} = 1$. We summarize routing sharpness with Herfindahl concentration

$$H_t = \sum_{i \in \mathcal{I}_t} a_{ti}^2 \in [1/n_t, 1], \qquad c_t = \frac{\log(n_t H_t)}{\log n_t} \in [0, 1], \tag{11}$$

for $n_t > 1$, with $c_t = 1/2$ at the first valid position. Uniform attention maps to $c_t = 0$ and a point mass to $c_t = 1$, removing the deterministic dependence of the minimum concentration on context length. The retention gate is

$$\kappa_t = \kappa_{\text{lo}} + (\kappa_{\text{hi}} - \kappa_{\text{lo}})\sigma(\tau(c_t - 1/2)), \tag{12}$$

with envelope parameters $(\kappa_{\text{lo}}, \kappa_{\text{hi}}, \tau) = (0.1, 0.9, 4)$ in the main experiments. Because the sigmoid temperature is finite, the attained gate range is a strict subinterval of the configured envelope. A larger gate retains more downstream value across the current transition under the chosen transport model.

The statistic describes the concentration of upstream support, not the identity or causal responsibility of individual tokens.

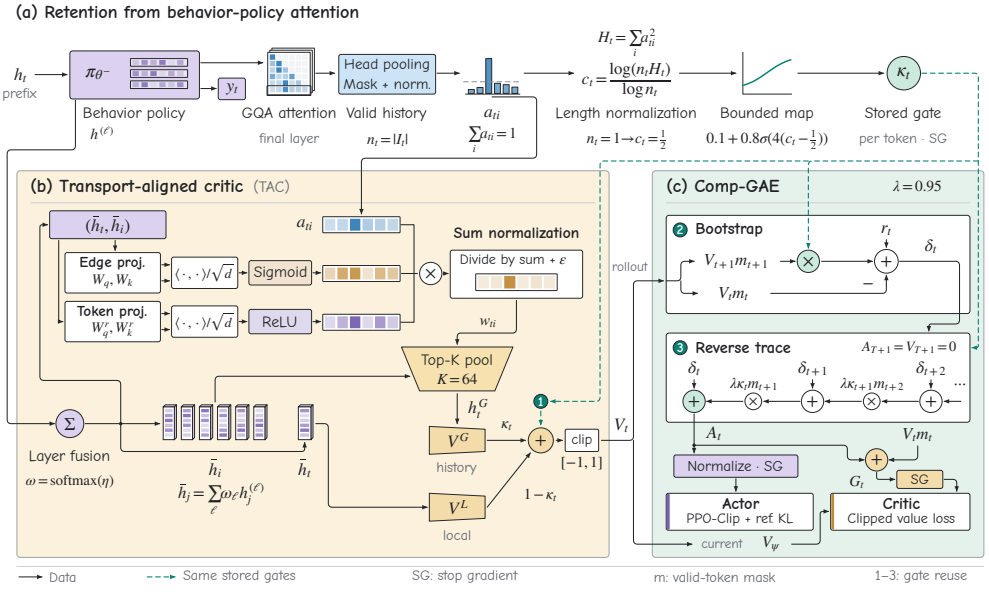

Figure 1: CompPO couples credit transport with value representation.

4.2 Transport-aligned critic

Comp-GAE requires values that track the computation-conditioned targets in Equation (7). Rather than train a second same-scale Transformer, TAC reuses actor activations and adds 14,428,163 parameters (0.36% of the Qwen3-4B actor (Yang et al., 2025)).

For selected upper layers,

$$\bar{h}_t = \sum_{\ell=1}^{L'} \omega_\ell h_t^{(\ell)}, \qquad \omega = \text{softmax}(\eta). \tag{13}$$

Native attention is then refined for value prediction. A bilinear gate modulates each edge,

$$\beta_{ti} = \frac{\langle W_{\beta,q}\bar{h}_t, W_{\beta,k}\bar{h}_i \rangle}{\sqrt{d}}, \tag{14}$$

$$\widetilde{a}_{ti} = a_{ti}\sigma(\beta_{ti}), \tag{15}$$

and a nonnegative token-relevance score

$$m_{ti} = \mathrm{ReLU}\left(\frac{\langle W_{r,q}\bar{h}_t, W_{r,k}\bar{h}_i\rangle}{\sqrt{d}}\right) \tag{16}$$

filters routing correlations that are not predictive of value. We call m_{ti} a token-relevance gate: it is a learned nonnegative filter, not a mutual-information estimator. The routed-history representation is

$$s_{ti} = \widetilde{a}_{ti} m_{ti} \mathbf{1}[i \in \mathcal{I}_t], \qquad h_t^G = \sum_{i \in \mathcal{I}_t} \frac{s_{ti}}{\sum_{j \in \mathcal{I}_t} s_{tj} + \varepsilon} \bar{h}_i. \tag{17}$$

Local and routed-history heads are combined with the same retention regime used by the target:

$$V_\psi(s_t^C) = \mathrm{clip}\big(\kappa_t V_\psi^G(h_t^G) + (1-\kappa_t) V_\psi^L(\bar{h}_t), -1, 1\big)\,. \tag{18}$$

The shared gate is a co-design hypothesis: when routing is concentrated, the value estimate relies more heavily on explicitly pooled history. The gate-by-critic factorial tests whether this alignment provides more than critic capacity alone.

4.3 Objectives and optimization

The actor uses normalized detached Comp-GAE advantages in the standard clipped objective (Schulman et al., 2017)

$$\mathcal{L}_{\mathrm{clip}}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho_t \widehat{A}_t, \mathrm{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\widehat{A}_t\Big)\Big]\,, \tag{19}$$

where $\rho_t = \pi_\theta(y_t \mid h_t)/\pi_{\theta^-}(y_t \mid h_t)$. The critic regresses to the detached λ-return $\widehat{G}_t = A_t^{\kappa,\lambda} + V_\psi(s_t^C)m_t$ with clipped value loss. Behavior-policy attention and gates are detached for the rollout batch and reused across the update epochs. Reward computation, rollout sampling, reference-model KL, and policy clipping are unchanged.

A fixed-gate control replaces every occurrence of κ_t in the bootstrap, trace, and value mixture by 0.61, the empirical mean of full CompPO. This isolates trajectory dependence from average horizon. The complete algorithm is given in Section A.

4.4 Long-context implementation

Final-layer chunked attention, GQA-group aggregation, response-window buffers, online-softmax recomputation, and causal top-K pooling ($K = 64$) avoid dense T^2 materialization and reduce critic work to $O(TK)$. The critic reuses actor features and the actor KV cache rather than a second same-scale value Transformer or a second attention cache. Section C.5 reports the configuration-derived parameter, memory, and FLOP budget.

5 Experimental Design

Training uses verifiable mathematical-reasoning prompts from DAPO-Math-17K (Yu et al., 2025) with a 16K response limit. The primary backbone is Qwen3-4B (Yang et al., 2025); Llama-3.1-8B-Instruct (Llama Team, AI @ Meta, 2024) is a transfer control. Reward is terminal exact-match; malformed reasoning tags receive a negative terminal reward, and the length penalty applies only to correct, format-valid responses. A 500-problem held-out ID set, distinct from MATH500, is used for selection and controller feedback. Frozen evaluation reports greedy pass@1 and majority@8 on AIME25, AIME2024, MATH500, GSM8K, LiveBench, GPQA, and an Olympiad-style set. Shared batch, sampling, optimizer, and selected hyperparameters appear in Section B; both methods use the same 3×3 actor-LR/KL screen. The experiments then test matched GRPO performance, gate–critic complementarity, trajectory-specific alignment, PPO-grid stability, and frozen transfer across backbones and group size.

Protocol. Each method is screened once; selected cells are rerun on seeds $\{17, 42, 123, 256, 2026\}$. Final ID accuracy at step 200 is primary; best, last-20 mean, AULC, and best-to-final drop are secondary. Intervals are two-sided 95% Student-t intervals over the five seeds. Frozen benchmarks are never used for hyperparameter or checkpoint selection.

Mechanism tests. These controls isolate the explanations left open in Section 3: mean horizon, position schedule, token alignment, and critic capacity. A 2×2 crosses dynamic versus mean-matched fixed transport ($\kappa = 0.61$) with a standard top-layer head versus TAC, yielding Fixed+Standard, COMP-GAE +Standard, Fixed+TAC, and full COMPPO. All four cells share data, seeds, reward, budget, controller, and horizon. Besides policy endpoints we record critic EV, TD-target Spearman correlation, value MAE, clamp saturation, and steps to 56.4%. The seed-matched interaction is

$$\Delta_{\text{int}} = (Y_{G,A} - Y_{F,A}) - (Y_{G,S} - Y_{F,S}), \tag{20}$$

where G/F are dynamic/fixed transport and A/S aligned/standard critics.

Global shuffle permutes realized gates over valid tokens before both COMP-GAE and critic mixing, keeping the empirical multiset but removing token correspondence. Position-only replaces each gate by a frozen 20-bin mean profile of normalized response position. A matched PPO stress grid (150 steps; actor LR $\{2\times10^{-7}, 5\times10^{-7}, 10^{-6}\}$ and KL $\{0, 10^{-3}\}$; two seeds per cell; no adaptive controller) tests bounded stability: final accuracy at least the 42.4% base, peak-to-final drop below 15 points, and no five consecutive post-peak checkpoints below base. A single-run GRPO 8×8 control (64 trajectories/update) probes group-size sensitivity against the default 4×4 setting.

6 RESULTS

6.1 MATCHED TUNING AND FIVE-SEED PERFORMANCE

Table 1 reports the primary comparison. Figure 2 shows the corresponding ID learning curves and a recorded PPO stress-grid trajectory. Every COMPPO seed exceeds every GRPO seed at both endpoints, and the best-to-final drop shrinks from 2.4 to 0.3 points. The five-seed mean curves in Figure 2a separate early and remain separated through step 200, so the gap is a sample-efficiency difference under a shared update budget rather than a longer COMPPO run. The same ordering holds at every actor learning rate in the shared screen, so the result is not an isolated-cell artifact. Both selected cells use the same 4×4 trajectory budget: the gap is not purchased with a larger GRPO group or a second 4B critic (Section C.5).

Table 1: Matched performance. Panel A: five-seed selected configurations with 95% Student-t intervals. Panel B: final ID accuracy in the common single-seed 3×3 screen.

A. Five-seed selected configurations

Method	Best	Final	Best–final	Seeds
GRPO	56.2 [55.6, 56.9]	53.8 [52.9, 54.7]	2.4	5
COMPPO	**61.7** [61.3, 62.1]	**61.4** [60.8, 62.0]	**0.3**	5

B. Final accuracy in the common LR/KL screen

Method / actor LR	KL $= 0$	KL $= 10^{-3}$	KL $= 2\times10^{-3}$	Row best
GRPO / 5×10^{-7}	51.2	52.4	51.8	52.4
GRPO / 10^{-6}	52.8	53.8	53.2	53.8
GRPO / 2×10^{-6}	51.6	52.0	51.4	52.0
COMPPO/ 5×10^{-7}	58.4	59.2	59.6	59.6
COMPPO/ 10^{-6}	**59.0**	**60.2**	**61.4**	**61.4**
COMPPO/ 2×10^{-6}	58.0	59.4	60.0	60.0

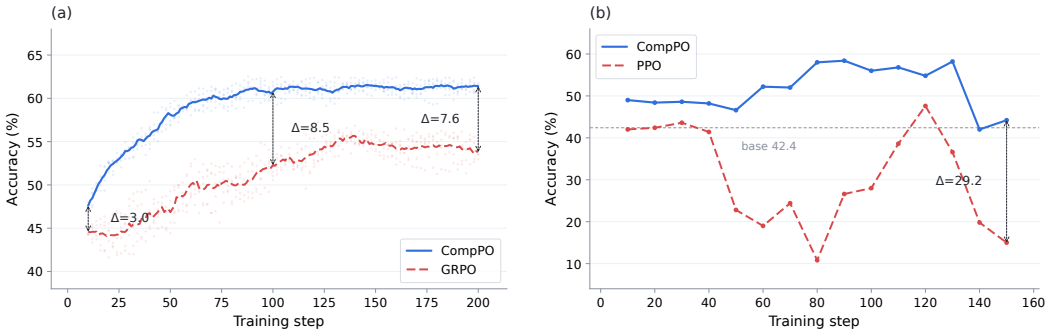

Figure 2: (a) CompPO versus GRPO held-out ID accuracy on Qwen3-4B. Solid lines are five-seed means; markers are seed-level observations. (b) Single-run PPO stress-grid trajectories at actor LR 10^{-6}, KL 10^{-3} (150 steps, no adaptive controller).

6.2 THE TRANSPORT GATE AND CRITIC ARE COMPLEMENTARY

Table 2 separates transport from critic architecture. Neither dynamic transport with the tested standard head nor TAC with a mean-matched fixed gate reproduces full COMPPO. The seed-matched final interaction in Equation (20) is $+2.4$ points $[1.9, 2.9]$ (best-score interaction $+1.9$ $[1.3, 2.4]$; Table 11). Full COMPPO exceeds COMP-GAE +Standard by 6.2 final points and Fixed+TAC by 5.0. The standard head has negative explained variance and never reaches the GRPO peak of 56.4% (step-to-$56.4 > 200$). Fixed+TAC reaches it at step 112; full COMPPO at step 33. TAC restores target tracking and is required for the trace to improve the policy; the same coupling also accelerates learning to the GRPO peak. The gain is this coupling, beyond critic capacity or a dynamic coefficient alone.

Table 2: Five-seed gate $\times$ critic factorial. Panel A: policy endpoints with 95% Student-t intervals. Panel B: critic diagnostics (point estimates). Seed-level endpoints are in Table 10.

A. Policy endpoints

Variant	Best	Final	Last-20	AULC	Drop
Fixed + Standard	53.4 [52.9, 53.9]	52.6 [52.0, 53.2]	52.0	47.4	0.8
COMP-GAE + Standard	56.0 [55.6, 56.5]	55.2 [54.6, 55.8]	54.7	50.2	0.8
Fixed + TAC	57.2 [56.6, 57.8]	56.4 [55.8, 57.0]	55.9	53.1	0.8
Full COMPPO	**61.7** [61.3, 62.1]	**61.4** [60.8, 62.0]	**60.8**	**58.9**	**0.3**

B. Critic diagnostics

Variant	EV	TD ρ_s	MAE	Clamp sat.	Step to 56.4
Fixed + Standard	-0.15	0.11	0.41	0.24	> 200
COMP-GAE + Standard	-0.08	0.16	0.37	0.19	> 200
Fixed + TAC	0.45	0.52	0.20	0.05	112
Full COMPPO	**0.52**	**0.57**	**0.18**	**0.03**	**33**

6.3 TRAJECTORY-SPECIFIC ALIGNMENT

Table 3 retains progressively less of the realized gate structure. Shuffle keeps the empirical multiset but removes token alignment; position-only keeps only the average schedule. Full COMPPO exceeds position-only by a paired $+1.7$ final points $[1.5, 1.8]$. Realized gate marginals (Table 12) therefore cannot explain the gap: the gain is not a favorable mean coefficient or a fixed position profile.

Table 3: Five-seed gate-alignment controls. Brackets are 95% intervals. The full-method row repeats Table 1A.

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]	59.0	57.8
Full COMPPO	**61.7** [61.3, 62.1]	**61.4** [60.8, 62.0]	**60.8**	**58.9**

6.4 ROBUSTNESS AND STRONGER GROUP-RELATIVE BASELINE

Table 4 is a bounded stress test rather than a fully tuned comparison; Figure 2b is one recorded $10^{-6}/10^{-3}$ trajectory. PPO is stable in $3/12$ runs and COMPPO in $10/12$. The no-KL 10^{-6} COMPPO cell is also unstable, so the evidence supports a wider stable region in this grid, not superiority under every learning rate and KL coefficient.

Increasing GRPO from 4×4 to 8×8 raises ID final accuracy from 53.8 to 56.2 and frozen greedy macro from 57.5 to 60.2 in a single run that uses four times as many trajectories per update and, under Section C.5, four times the rollout FLOP and response-window KV. The gap to the 16-trajectory COMPPO reference (61.4 ID final, 61.7 frozen macro) narrows but remains: even a $4\times$ GRPO rollout budget does not match the COMPPO learning curve. This single-run comparison is not a significance test.

Table 4: PPO stress grid (two seeds per cell). Drop is best minus final. The last row is the unweighted mean over the six cells ($3/12$ vs $10/12$ stable).

Actor LR / KL	PPO final / stable	COMPPO final / stable	PPO best / drop	COMPPO best / drop
2×10^{-7} / 0	44.8 / 1/2	**48.6 / 2/2**	48.2 / 3.4	**52.4** / 3.8
2×10^{-7} / 10^{-3}	46.2 / 1/2	**49.2 / 2/2**	49.0 / 3.8	**53.0** / 3.8
5×10^{-7} / 0	31.0 / 0/2	**47.0 / 2/2**	46.8 / 15.8	**51.6 / 4.6**
5×10^{-7} / 10^{-3}	43.6 / 1/2	**47.8 / 2/2**	48.4 / 4.8	**52.2 / 4.4**
10^{-6} / 0	18.4 / 0/2	**36.4** / 0/2	44.2 / 25.8	**50.8 / 14.4**
10^{-6} / 10^{-3}	22.6 / 0/2	**44.6 / 2/2**	49.8 / 27.2	**57.2 / 12.6**
Mean	34.4 / 3/12	**45.6 / 10/12**	47.7 / 13.3	**52.9 / 7.3**

6.5 FROZEN BENCHMARK TRANSFER

Table 5 reports every benchmark. On Qwen3-4B, AIME25 greedy is the only negative difference; AIME2024 and GPQA contribute the largest gains. GRPO's unweighted greedy macro (57.5) is below the 58.2 base, while COMPPO reaches 61.7 (Tables 13 and 14). On Llama-3.1-8B-Instruct, all seven greedy differences favor COMPPO. These are final-checkpoint point estimates and do not quantify training-seed uncertainty per benchmark.

Table 5: Frozen final-checkpoint evaluation. B/G/C denote the base model, GRPO, and COMPPO. Sample counts are in Section B.

	Qwen3-4B						**Llama-3.1-8B-Instruct**					
	Greedy			Maj.@8			Greedy			Maj.@8		
Benchmark	B	G	C	B	G	C	B	G	C	B	G	C
AIME25	43.3	**50.0**	43.3	**66.7**	**66.7**	**66.7**	3.3	13.3	**16.7**	6.7	23.3	**30.0**
AIME2024	63.3	53.3	**70.0**	76.7	73.3	**80.0**	13.3	20.0	**23.3**	16.7	33.3	**36.7**
MATH500	67.2	66.6	**69.8**	79.8	79.2	**81.0**	53.4	64.2	**69.6**	64.4	76.6	**80.6**
GSM8K	91.7	93.1	**94.1**	93.5	94.3	**94.5**	81.3	86.4	**87.6**	85.6	90.4	**91.2**
LiveBench	46.5	50.5	**53.5**	61.0	61.5	**67.0**	22.5	35.0	**39.0**	29.0	44.5	**49.0**
GPQA	47.0	40.4	**52.0**	**57.6**	54.5	55.6	26.8	32.8	**40.4**	32.3	39.4	**47.0**
Olympiad	48.5	48.2	**49.4**	**53.0**	51.6	52.1	17.1	21.1	**23.3**	21.1	26.7	**29.8**
Macro	58.2	57.5	**61.7**	69.7	68.7	**71.0**	31.1	39.0	**42.8**	36.5	47.8	**52.0**

7 DISCUSSION AND CONCLUSION

The experiments support a unified mechanism: trajectory-dependent credit transport improves learning when the value representation is grounded in the same computational coordinates. A controlled factorial isolates this coupling from critic capacity or dynamic scaling alone, while shuffle and position-only controls rule out mean horizon and fixed schedules. Crucially, by separating credit transport $\mathbf{K}$ from the outer policy update $\mathcal{U}$ and outcome evidence $\mathbf{e}$, CCT establishes an orthogonal axis that natively composes with advancements in sequence-level clipping, dynamic sampling, and process supervision, bypassing confounding algorithmic bake-offs.

This perspective re-evaluates the prevailing trend of abandoning learned critics in LLM RL. The empirical failures of conventional value heads often stem not from a failure of value modeling itself, but from a severe representation–target mismatch over long contexts. Once internal routing is exposed to the critic, learned values faithfully track computation-conditioned targets, demonstrating that token-distance decay and group broadcasting are unnecessary constraints.

The resulting coupling translates directly into superior sample efficiency rather than compute inflation. Under an identical 4×4 budget, five-seed COMPPO surpasses the GRPO peak at step 33 (versus step 112 for Fixed+TAC and > 200 for standard heads), and quadrupling GRPO rollout trajectories (8×8) still fails to match the 16-trajectory COMPPO curve. Notably, TAC preserves the persistent-memory class of critic-free methods without introducing a second same-scale value model (Section C.5).

We delineate the scope of these claims: attention concentration serves as a structural proxy rather than causal attribution, detached gates introduce off-policy staleness beyond our contraction guarantee, and benchmarks are evaluated on reasoning tasks up to 8B models. Frozen benchmarks and the 8×8 baseline remain point estimates.

LLMs compute conditionally over history, yet conventional credit estimators treat that history as an unstructured sequence. CCT bridges this divide by turning internal computation into an explicit coordinate of credit transport via $\mathbf{K} = (I - \lambda D_\kappa S)^{-1}$. COMPPO validates this principle using an attention retention gate and a 14.4M aligned critic under standard clipped PPO. Rather than claiming Herfindahl concentration is optimal, we view this minimal scalar gate as the starting point for architecture-aware Bellman operators and trace design in foundation models.

AI USE STATEMENT

Language models appear in this work only as the experimental policies being trained, not as a substitute for authorship. In this work, we have not used generative AI tools for any task requiring disclosure: generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results; these required-disclosure uses are not applicable. Additionally, we used generative AI tools solely to edit the manuscript to improve readability. We have reviewed all AI-assisted edits against the authors' source text and the experimental record. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of all reported findings, this work explicitly references the implementation, theoretical foundations, data processing, and empirical configurations across the manuscript, appendices, and supplementary materials in compliance with ICLR guidelines. For novel models and algorithms, the mathematical formulation and architectural design of COMPPO are detailed in Section 4, with complete procedural pseudocode documented in Section A (Algorithm 1); an anonymous downloadable source code package containing the credit-transport operator, gating mechanism, distributed training pipelines, and evaluation harnesses is provided in the supplementary materials. For theoretical results, all operational assumptions, operator contraction properties, matrix kernel path expansions, bounded credit traces, and strict reductions to standard GAE are formally proven in Section D. For datasets and experimental protocols, the public mathematical reasoning benchmarks, prompt templates, data processing procedures, and deterministic ground-truth verification rules are specified in Section 5 and Section C, with hyperparameter search grids and dynamic controller configurations tabulated in Tables 16 to 18. Finally, all per-seed and per-cell results across five random seeds, factorial ablations, and frozen cross-architecture evaluations are compiled in Section B, while configuration-derived parameter, memory, and FLOP budgets are formalized in Section C.5.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 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, Weixun Wang, Yijia Luo, Zinian Peng, Taiheng Ye, Chao Yang, Wenbo Su, Yu Cheng, Bo Zheng, and Junchi Yan. How does reasoning flow? tracing attention-induced information flow for targeted RL in LLMs. In *International Conference on Machine Learning*, 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*, pp. 2644–2652, 2019.

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

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590, 2025.

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, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, Bo Zheng, and Junchi Yan. Attention illuminates LLM reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. In *International Conference on Machine Learning*, 2026b.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. OPPO: Bayesian value recursion for token-level credit assignment in LLM reasoning. *arXiv preprint arXiv:2605.21851*, 2026c.

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*, 2026d.

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.

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.

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

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, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 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. In *International Conference on Learning Representations*, 2016.

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, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from 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, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 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. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3742–3750, 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*, 2019.

A Full Algorithm

Algorithm 1 Computation-Conditioned Policy Optimization (CompPO)

Require: actor π_θ, reference policy π_{ref}, critic V_ψ, verifier R, gate parameters $(\kappa_{\text{lo}}, \kappa_{\text{hi}}, \tau)$, trace 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 attention a_{ti} over valid history $\mathcal{I}_t$ and selected actor hidden states $\{h_t^{(\ell)}\}$
    Renormalize a_{ti} over $\mathcal{I}_t$; 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 by Equations (15) and (16)
    Pool routed history h_t^G by Equation (17)
    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 Equation (19) plus reference-policy KL
    Update critic by clipped regression to detached G_t
    Apply the predeclared phase controller when enabled

B Additional Numerical Results

B.1 Related-work family axes

These families are separated by the credit coordinate they primarily change.

Table 6: Primary intervention of representative LLM-RL families. Categories mark the dominant axis, not an exclusive taxonomy.

Family	Evidence / baseline	Transport	Update / value
PPO/GAE	learned scalar value	stationary temporal trace	token clip + critic
GRPO/RLOO	group/LOO outcome	response-level broadcast	critic-free token update
DAPO/GSPO	group outcome + sampling	mainly group credit	clip, norm, sequence ratio
PRM/PRIME	process/implicit evidence	reward placement + estimator	process/reward model
VinePPO/TEMPO	continuation or tree values	rollout- or tree-based TD	MC or nonparametric value
OAR/OPPO/S-trace	influence or oracle signal	reshaping or selective trace	critic-free update
CompPO	terminal / verifier reward	path-dependent Bellman/GAE kernel	actor-feature TAC + PPO

B.2 TRAINING CONFIGURATION AND MATCHED PERFORMANCE

The matched comparison uses one shared screen. The selected cells are then rerun on the five confirmation seeds.

Table 7: Primary training configuration shared by the matched comparison. The PPO stress grid disables the adaptive controller and uses the smaller LR/KL set in Section 5.

Item	GRPO	COMPPO
Backbone / max response	Qwen3-4B / 16K	same
Batch / responses / minibatch	4 / 4 / 4	same
Actor LR candidates	$\{5\times10^{-7}, 10^{-6}, 2\times10^{-6}\}$	same
KL candidates	$\{0, 10^{-3}, 2\times10^{-3}\}$	same
Selected actor LR / KL	$10^{-6}/10^{-3}$	$10^{-6}/2\times10^{-3}$
Policy clip / update epochs	0.20 / 2	same
Sampling temperature / top-p	1.0 / 0.7	same
Credit estimator	group-relative	κ_t bootstrap; $\lambda = 0.95$
Critic LR / value clamp	–	10^{-5} / $[-1, 1]$
Gate envelope / temperature	–	$(0.1, 0.9)$ / 4
Precision / optimizer	BF16 / AdamW	same

Table 8: Seed-level selected-configuration 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 9: Best ID accuracy in the common single-seed screen. Final values are reported in the main text.

Method	Actor LR	KL $= 0$	KL $= 10^{-3}$	KL $= 2\times10^{-3}$
GRPO	5×10^{-7}	52.8	54.0	53.2
GRPO	10^{-6}	54.6	**56.4**	55.0
GRPO	2×10^{-6}	53.8	54.2	53.0
COMPPO	5×10^{-7}	59.8	60.4	60.8
COMPPO	10^{-6}	60.6	61.2	**61.8**
COMPPO	2×10^{-6}	59.6	60.8	61.0

B.3 FACTORIAL ENDPOINTS AND GATE STATISTICS

Every factorial cell uses the same five seeds. The gate summaries come from the five full-method checkpoints.

Table 10: Seed-level gate-by-critic factorial endpoints.

	Fixed gate standard critic		COMP-GAE standard critic		Fixed gate TAC		Full COMP-GAE + TAC	
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
Mean	53.4	52.6	56.0	55.2	57.2	56.4	61.7	61.4
95% CI	[52.9, 53.9]	[52.0, 53.2]	[55.6, 56.5]	[54.6, 55.8]	[56.6, 57.8]	[55.8, 57.0]	[61.3, 62.1]	[60.8, 62.0]

Table 11: Seed-matched gate–critic interactions from Equation (20), computed from Table 10. Intervals are 95% Student-t.

Seed	Final Δ_{int}	Best Δ_{int}
17	2.4	1.8
42	1.8	1.2
123	2.6	1.9
256	2.4	1.9
2026	2.8	2.5
Mean	2.4	1.9
95% CI	[1.9, 2.9]	[1.3, 2.4]

Table 12: Gate statistics from five final full-method checkpoints. Early, middle, and late are response tertiles; correct and incorrect are outcome-conditioned means.

	Mean	Median	IQR	Early	Middle	Late	Correct	Incorrect
Retention gate κ	0.61	0.63	0.18	0.52	0.63	0.68	0.64	0.57

B.4 FROZEN EVALUATION AND STRONGER GROUP CONTROL

Frozen entries are final-checkpoint point estimates. The 8×8 GRPO comparison is a single run.

Table 13: Frozen Qwen3-4B final-checkpoint point estimates, including base models.

		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 14: Frozen Llama-3.1-8B-Instruct final-checkpoint point estimates, including base models.

		Greedy pass@1			Majority@8		
Benchmark	N	Base	GRPO	CompPO	Base	GRPO	CompPO
AIME25	30	3.3	13.3	16.7	6.7	23.3	30.0
AIME2024	30	13.3	20.0	23.3	16.7	33.3	36.7
MATH500	500	53.4	64.2	69.6	64.4	76.6	80.6
GSM8K	1319	81.3	86.4	87.6	85.6	90.4	91.2
LiveBench	200	22.5	35.0	39.0	29.0	44.5	49.0
GPQA	198	26.8	32.8	40.4	32.3	39.4	47.0
Olympiad	674	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 15: Single-run stronger-group control. The CompPO row is the primary 4×4 configuration.

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

C Reproducibility Details

C.1 Optimization and phase controller

The performance experiments and all four factorial cells use the same closed-loop stability controller. It monitors policy KL, gradient norm, clip fraction, entropy, and ID development accuracy. Escalation requires two consecutive trigger events; de-escalation requires five clear steps.

Table 16: Adaptive stability phases used in the full performance and factorial experiments.

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

The controller uses exponential moving averages with coefficient 0.4 for policy clip fraction, PPO KL, KL loss, and gradient norm, and 0.2 for entropy. The trigger map and the remaining phase knobs are listed below. A format-penalty coefficient ramps 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 grid disables critic warm-up and the adaptive controller; the fixed value clamp remains applied to the relevant value path.

Table 17: Controller trigger thresholds. Each cell is the cutoff that requests the named phase.

Signal	Warning	Strong	Hard-stop
Entropy EMA (below)	8.0	7.5	6.5
Clip fraction (at least)	0.14	0.18	0.25
PPO KL (at least)	0.08	0.12	0.22
KL loss (at least)	2.85	3.20	4.00
Gradient norm (at least)	40	80	–
ID-accuracy drop (at least)	0.04	0.08	0.20
Response-clip ratio (at least)	–	–	0.65

Table 18: Additional knobs attached to each controller phase.

Quantity	Stable	Warning	Strong	Hard-stop
Advantage clip	5.0	4.0	3.5	2.5
Actor gradient clip	1.0	0.8	0.7	0.5
Critic gradient clip	1.0	0.8	0.7	0.5
Actor epochs	2	2	2	1

C.2 VALUE-CLAMP CONTROL

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 stress grid, clamp activation was confined to comparable early critic transients and then became the identity.

C.3 ATTENTION EXTRACTION AND MEMORY STRUCTURE

The implementation uses final-layer GQA-group attention. Only the final layer exposes chunked attention; earlier layers retain fused kernels. Response-window buffers avoid prompt-heavy storage, online-softmax recomputation avoids dense attention writes, and prior-guided causal top-K pooling uses $K = 64$. The TAC module contains $14{,}428{,}163$ parameters (0.359% of the Qwen3-4B actor). It reuses actor hidden states and the actor KV cache, so reading routing for the gate and value mix does not allocate a second attention cache or a second same-scale value Transformer.

C.4 STATISTICS AND EVALUATION

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. Final is primary; best, last-20 mean, and AULC are secondary. Frozen benchmark results are final-checkpoint point estimates. Macro-averages are unweighted across the seven benchmarks and should be interpreted with the per-benchmark values because the AIME sets contain 30 examples each.

C.5 CONFIGURATION-DERIVED SYSTEMS BUDGET

The tables below are a *configuration-derived* parameter, memory, and FLOP budget under Table 7: Qwen3-4B, maximum response 16K, batch 4, 4 responses per prompt, minibatch 4, two update epochs, BF16, and AdamW. Selected GRPO and COMPPO cells use nonzero KL, so a frozen reference copy is included. Sequence quantities use the configured 16K response cap and the response-window convention; prompt length is omitted. These figures are analytical budgets, not measured runtime or peak-memory results.

The Qwen3-4B actor has $4{,}022{,}468{,}096$ parameters. Following the standard veRL PPO layout, the learned critic is a separate token-classification copy of the same backbone with a scalar score head ($4{,}022{,}470{,}657$ parameters) and its own AdamW state. GRPO has no critic. COMPPO adds the given TAC count $14{,}428{,}163$. The factorial standard top-layer head in Table 2 is a mechanism control; it is not the veRL PPO column.

AdamW is counted as two FP32 moment tensors plus BF16 gradients, with no offload and no extra FP32 master copy: 12 bytes per trainable parameter and 2 bytes per frozen reference parameter. One veRL step includes decode generation, an actor log-probability prefill, a reference prefill, and two update epochs. PPO adds a critic value prefill and a two-epoch critic update without sharing the actor KV cache. COMPPO adds only the TAC forward and the same two-epoch critic update on those 14.4M parameters. FLOPs use per-layer QKV/MLP matmuls plus GQA attention at $T = 16{,}384$; a backward pass is counted as $2\times$ a forward pass.

Table 19: Exact parameter counts under the primary Qwen3-4B recipe. The reference is frozen. The PPO critic follows the veRL token-classification layout.

	GRPO	veRL PPO	CompPO
Actor (trainable)	4,022,468,096	4,022,468,096	4,022,468,096
Reference (frozen)	4,022,468,096	4,022,468,096	4,022,468,096
Critic	0	4,022,470,657	14,428,163
Trainable total	4,022,468,096	8,044,938,753	4,036,896,259

Table 20: Persistent HBM in GiB (BF16 weights and gradients, AdamW FP32 moments, no offload). Response-window KV is the 16K cap for the listed trajectory count.

	GRPO 4×4	veRL PPO 4×4	CompPO 4×4
Trainable weights	7.49	14.98	7.52
Reference weights	7.49	7.49	7.49
AdamW	29.97	59.94	30.08
Gradients	7.49	14.98	7.52
Persistent total	52.45	97.40	52.61
vs. GRPO	–	+44.95	+0.16
Response KV (16 seqs)	36.00	36.00	36.00 (reused)

Table 21: Architecture-exact PFLOP per update at the 16K response cap. Shared work is generation, actor/reference prefills, and the two-epoch actor update.

	GRPO 4×4	veRL PPO 4×4	CompPO 4×4
Shared actor/reference	40.51	40.51	40.51
Critic value + update	0	31.07	0.05
Total	40.51	71.58	40.57
vs. GRPO	–	+76.7%	+0.13%

GRPO rollout HBM and FLOPs scale with the group (B, G). Persistent actor/reference memory does not. The 8×8 control uses 64 trajectories per update. With minibatch 4 and two epochs this is exactly four times the default 4×4 generation, KV footprint, and update work (Table 22). That control improves ID final accuracy from 53.8 to 56.2 but remains below the 16-trajectory CompPO reference (61.4). PPO's extra cost is the second 4B critic and does not grow with G; CompPO's extra cost is the 14.4M TAC and likewise does not grow with G.

Table 22: Group-size scaling for GRPO against the matched 4×4 CompPO recipe. ID final / frozen greedy macro are the paper endpoints; 8×8 is a single run.

Setting	Traj./update	KV (GiB)	PFLOP/step	ID final / greedy macro
GRPO 4×4	16	36.00	40.51	53.8 / 57.5
GRPO 8×8	64	144.00	162.05	56.2 / 60.2
CompPO 4×4	16	36.00	40.57	61.4 / 61.7

D Proofs and Limiting Cases

D.1 Strict reduction

If $\kappa_t = \gamma$ for every valid transition, Equation (7) becomes the standard TD residual and Equation (8) becomes standard GAE. The outer objective Equation (19) is unchanged. Restoring the standard critic therefore recovers the standard fixed-γ GAE/PPO estimator path. The Fixed+TAC factorial cell is intentionally not called PPO because it preserves the proposed value representation while removing trajectory-specific transport.

D.2 Matrix kernel and path expansion

Let $D_\kappa = \mathrm{diag}(\kappa_1, \ldots, \kappa_T)$ and $(Sv)_t = v_{t+1}$ on a finite response, so S is strictly upper-triangular and nilpotent. The vector form of Equation (8) is $\mathbf{A} = \boldsymbol{\delta} + \lambda D_\kappa S \mathbf{A}$, hence $\mathbf{K}^{\kappa,\lambda} = (I - \lambda D_\kappa S)^{-1}$

as in Equation (9). Repeated substitution yields

$$A_t^{\kappa,\lambda} = \delta_t^\kappa + \lambda\kappa_t\delta_{t+1}^\kappa + \lambda^2\kappa_t\kappa_{t+1}\delta_{t+2}^\kappa + \cdots,$$

which gives Equation (10). The causal series is finite and the solution is unique.

D.3 Contraction of the detached operator

For bounded V, W and any history h_t,

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

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

D.4 Bounded trace

If $|\delta_t^\kappa| \le D$ and $\kappa_t \le \bar{\kappa}$, then

$$|A_t^{\kappa,\lambda}| \le D\sum_{\ell=0}^{\infty}(\lambda\bar{\kappa})^\ell = \frac{D}{1-\lambda\bar{\kappa}}.$$

D.5 Length normalization

For a probability vector over n_t positions, Cauchy–Schwarz gives $1/n_t \le H_t \le 1$. Hence $n_t H_t \in [1, n_t]$ and $c_t = \log(n_t H_t)/\log n_t \in [0, 1]$. Uniform attention maps to zero and a point mass to one independently of context length.

D.6 Difference from loss weighting

Multiplying the policy loss by a token score changes only the gradient magnitude at that token. In CompPO, κ_t changes the TD bootstrap $r_t + \kappa_t V_{t+1}$ and every later residual's path coefficient in Equation (10); it also changes the value representation in Equation (18). The policy loss receives the resulting advantage without a direct multiplicative attention weight.

E Extended Limitations and Broader Impact

Attention is a proxy. Concentration can be distorted by attention sinks, delimiter tokens, positional bias, induction heads, or output-layer calibration. The method does not establish that heavily attended tokens caused the outcome. Intervention-based relevance, head selection, multi-layer aggregation, and graph-valued transport are important extensions.

Policy dependence and staleness. The gate is computed by the behavior policy and held fixed through the update epochs. This avoids differentiating through the estimator but creates staleness as the actor changes. Recomputing gates, reducing update epochs, or regularizing gate drift may improve fidelity.

Concentration statistic. Herfindahl concentration is bounded, inexpensive, and length-normalizable, but it is one of many possibilities. Shannon or Rényi entropy, top-k coverage, learned bounded maps, attention flow, retrieval provenance, or expert routing may provide stronger coordinates.

Reward dependence and co-adaptation. Computation-conditioned transport can amplify a bad reward as easily as a good one. Checker errors and format artifacts may be propagated through high-gate transitions. A policy may also learn internal routing patterns that manipulate the estimator without improving external behavior. External evaluation and monitoring of gate distributions remain necessary.

Empirical scope. The evidence covers mathematical reasoning, two backbones up to 8B, terminal verifiers, and responses up to 16K. Dialogue, tools, code, retrieval, multimodal generation, dense rewards, and larger models are not established. Frozen results and the stronger group control are point estimates. Systems quantities in Section C.5 are configuration-derived parameter, memory, and FLOP budgets.

Potential impact. More accurate credit transport may improve sample efficiency and make lightweight actor-feature critics practical for long-horizon training. The same capability can also improve general-purpose generative agents with harmful applications. The method is not a safety mechanism, and internal routing should not be treated as a guarantee of alignment or faithfulness.

F USE OF LANGUAGE MODELS

Language models appear in this work only as the experimental policies being trained. Generative AI tools were used solely to edit the manuscript to improve readability. They were not used for any task requiring disclosure under the ICLR LLM-use policy, including synthetic data generation, theory development, mathematical claims or proofs, hypothesis refinement, experimental design, method implementation, translation, dataset cleaning, qualitative analysis, or result interpretation. All AI-assisted edits were reviewed against the authors' source text and the experimental record. The authors retain responsibility for the scientific claims, derivations, citations, and empirical record.