How Parameter Correspondence Shapes Cross-Layer Optimization

Anonymous authors
Paper under double-blind review

Abstract

Cross-layer optimization couples updates from different Transformer layers and therefore requires a correspondence between their parameters. We introduce bounded depth tempering (BDT), which reshapes AdamW proposals across depth by attenuating high-energy depthwise modes while preserving their common-depth component and pooled centered norm. We compare BDT with correspondence-altered controls matched in common component, centered norm, and angular departure from AdamW at each control branch's own state. In controlled 124M language models, BDT lowers validation loss relative to both AdamW and these matched controls, showing that correspondence-derived direction carries useful information beyond these measures of update strength. Pretrained layer copying then gives a known relationship between descendants of the same block. Preserving that ancestry retains the benefit, while breaking the ancestry match loses it. Ancestry alone is not sufficient. True descendant pairing outperforms false pairing when update differences are amplified, and the ordering reverses under suppression. We then separate magnitude allocation across ancestry-defined sectors from direction within them. Crossed controls identify magnitude allocation as the more consequential difference. Under a prospectively specified shared window-smoothed direction construction, full-BDT allocations outperform structured-rule allocations, with the ordering reproduced on fresh parents. Relative to fixed ancestry amplification, retained diagnostics show that BDT assigns a larger share of centered proposal energy to descendant differences early in training and a smaller share late. Together, these results establish correspondence, operation, and magnitude allocation as distinct design choices in depth-coupled optimization.

1 Introduction

Recent optimizer constructions can couple update information across Transformer depth. Gradient Smoothing, for example, explicitly smooths layerwise updates across neighboring layers (Meng et al., 2026). Any such cross-layer operation requires a correspondence, meaning a rule for which parameters in one layer are compared with which parameters in another.

The most immediate correspondence is to match same-shaped parameter tensors coordinatewise by their stored indices. This correspondence depends on how the network is parameterized. Consistently permuting a feed-forward network's hidden units preserves its function while changing which coordinates a cross-layer rule compares. This issue is closely related to work on permutation alignment and optimizer symmetry (Ainsworth et al., 2023; Xie et al., 2025; Singh, 2026; Lau & Su, 2026).

Pretrained layer copying provides a case in which one cross-layer relationship is known by construction. When a pretrained block is copied to multiple depths, its descendants share an origin while subsequently receiving different gradients. Model-growth and recurrent architectures can create or exploit related structure across depth (Dehghani et al., 2019; Wang et al., 2023; Du et al., 2024; Bae et al., 2025), but our question is different: does the inherited correspondence itself contain useful information for optimization, and how should an optimizer act on it? Identifying related parameters alone does not determine whether their update differences should be amplified, attenuated, or redirected.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce **bounded depth tempering (BDT)**, a depth-coupled transformation of AdamW proposals. In our experiments, BDT improves validation loss over AdamW. It estimates high-energy depthwise modes of recent proposals, attenuates them with bounded conditioning, preserves the common-depth component, and restores the pooled centered norm. Its construction also makes the correspondence used by the optimizer directly manipulable without changing model storage.

BDT is related to existing structured update transformations. Gradient Smoothing mixes proposals from neighboring layers through a local depthwise window (Meng et al., 2026). BDT instead estimates a data-dependent operator over the depth axis from recent proposal statistics. This depth-axis construction differs from matrix-coordinate methods such as SOAP (Vyas et al., 2024) and other spectral update transformations (Shumaylov et al., 2026).

We use BDT to ask what underlies its benefit. First, does the useful update direction depend on the chosen correspondence, or only on the strength of BDT's departure from AdamW? A correspondence-scrambled control matches the common component, pooled centered norm, and angular departure from AdamW at its own state. BDT lowers validation loss relative to both AdamW and this matched control, showing that correspondence-derived direction carries useful information beyond these measures of intervention strength.

We next ask which cross-layer relationships carry that information. Pretrained layer copying gives a known ancestry relation between pairs of descendants. We use the same coordinate maps equally often, but change whether descendants of the same pretrained block receive the same map. Ancestry-preserving assignments retain the benefit, whereas reassignment loses it. This identifies useful inherited structure beyond canonical coordinate labels.

The relationship alone, however, does not determine a useful operation. True descendant pairing outperforms false pairing when update differences are amplified, but the ordering reverses under suppression. Although the pooled centered norm is fixed, its distribution between ancestry-defined components can vary, changing the overall update direction. We therefore distinguish the magnitude assigned to moving descendants together or apart from the direction taken within each component.

We compare full BDT with a structured alternative that retains the magnitude allocation of fixed amplification while modifying the pair-difference direction. In crossed development controls, the loss gap tracks the allocation source more closely than the within-sector direction source.

We then test prospectively whether this allocation advantage persists under a third direction construction that smooths proposals across neighboring layers. With both alternatives using the same within-sector direction rule, full-BDT allocations outperform structured-rule allocations on all four reused development parents. A separate confirmation reproduces the ordering on four fresh pretrained parents. Retained diagnostics from the original prospective test characterize the allocation: relative to fixed amplification, BDT assigns a larger share of centered proposal energy to descendant differences early in training and a smaller share late.

Together, these experiments give BDT two complementary roles. It is a constructive optimizer transformation that improves over AdamW in the tested setting, and its controlled structure exposes three distinct choices in depth-coupled optimization: which parameters correspond, what operation acts on that relationship, and how update magnitude is distributed across it.

2 BOUNDED DEPTH TEMPERING AND CORRESPONDENCE CONTROLS

BDT acts on AdamW proposals across depth while leaving the Transformer's forward computation unchanged. We first define the proposal representation and its correspondence, then construct BDT and a control that changes correspondence at matched update strength.

2.1 REPRESENTING PROPOSALS ACROSS DEPTH

Consider the same parameter group, such as the gate projection of a feed-forward network (FFN), at R depths. Let $U_t \in \mathbb{R}^{R\times P}$ stack the signed AdamW data increments at update t, with one layer's flattened P-dimensional proposal in each row. The learning rate is included in U_t, and decoupled weight decay is applied separately (Loshchilov & Hutter, 2019).

Let H be the $R \times R$ centering matrix across depth. With I_R the identity matrix and $\mathbf{1} \in \mathbb{R}^R$ the all-ones vector, define

$$H = I_R - \mathbf{1}\mathbf{1}^\top/R, \qquad U_{c,t} = HU_t, \qquad U_{m,t} = (I_R - H)U_t, \tag{1}$$

Multiplying by H subtracts the across-depth mean from each layer's proposal. Every row of $U_{m,t}$ contains that mean, while the rows of $U_{c,t}$ contain the deviations from it and sum to zero across depth. BDT modifies these deviations while retaining the shared update. Their Frobenius norm, $\|U_{c,t}\|_F$, is the square root of the sum of squared entries and pools all depths and parameter coordinates.

Correspondence. The rows of U_t identify the layers being compared. A *coordinate chart* specifies which entries within those layers correspond. The *canonical (stored-index) chart* matches entry k in one layer with entry k in another.

In a SwiGLU FFN, consistently permuting gate/up rows and down columns preserves the represented function. Let u_r and u_s be the flattened proposals at two depths, and let P_r, P_s denote the coordinate permutations induced by their hidden-unit maps. Their inner product becomes $u_r^\top P_r^\top P_s u_s$. A shared permutation cancels. Different permutations generally change the comparison. Our controls change these maps with model storage fixed (Appendix B).

2.2 BOUNDED DEPTH TEMPERING

BDT estimates which patterns of variation across depth have recently carried the most proposal energy. Let $B \in \mathbb{R}^{R\times(R-1)}$ contain an orthonormal basis for the centered depth subspace and let $Z_t = B^\top U_{c,t}$ express the centered proposal in that basis.

The matrix C_t tracks an exponential moving average of the normalized proposal second moment. Every tenth update,

$$C_t = 0.9C_{t^-} + 0.1\frac{Z_t Z_t^\top}{\|U_{c,t}\|_F^2}, \qquad C_0 = 0, \tag{2}$$

where t^- is the preceding observation step. A zero centered proposal leaves the statistic unchanged. We collect statistics during the first 200 AdamW updates, then increase the BDT treatment to full strength over the next 200 updates. For grown models, this schedule begins after copying.

At an observation step, write $C_t = V\,\mathrm{diag}(\lambda)V^\top$, where the columns of V are eigenvectors and λ_i are their eigenvalues. Larger eigenvalues identify depth directions with more normalized proposal energy. The statistic and eigensystem are held between observations. For twelve layers, C_t is only 11×11.

Let $g_i(t)$ be the attenuation factor for eigenmode i, let $g(t)$ be their vector, and let $M_t \in \mathbb{R}^{R\times R}$ be the resulting operator in depth coordinates. The ramp $a_t = \mathrm{clip}((t-200)/200, 0, 1)$ controls activation, and $\kappa \geq 1$ bounds relative attenuation. For $a_t > 0$, largest eigenvalue $\lambda_{\max} > 0$, and $U_{c,t} \neq 0$, define

$$g_i(t) = \left[1 + (\kappa^2 - 1)a_t\lambda_i/\lambda_{\max}\right]^{-1/2}, \qquad M_t = BV\,\mathrm{diag}(g(t))V^\top B^\top. \tag{3}$$

M_t applies gain $g_i(t)$ along each centered depth eigenmode. The gains depend on relative energy $\lambda_i/\lambda_{\max}$ and lie between $1/\kappa$ and 1. At full strength the highest-energy mode receives gain $1/\kappa$. The principal matched-direction experiments use $\kappa = 3$ (Appendices A and E).

Applying M_t changes both direction and norm. BDT restores the original centered norm and adds back the common component, giving the canonical-chart proposal V_t^{can}:

$$V_t^{\mathrm{can}} = U_{m,t} + \frac{\|U_{c,t}\|_F}{\|M_t U_{c,t}\|_F} M_t U_{c,t}. \tag{4}$$

This preserves $U_{m,t}$ and $\|U_{c,t}\|_F$, hence also the full proposal norm. BDT redistributes centered magnitude across depth directions at fixed total magnitude.

Let W collect the corresponding model parameters across depths. After BDT constructs a transformed proposal $\widetilde{U}_t$, we apply it with ordinary decoupled weight decay,

$$W \leftarrow (1 - \eta_t \omega) W + \widetilde{U}_t.$$

Here η_t is the learning rate and ω is the weight-decay coefficient. BDT modifies only the applied proposal. Adam's first- and second-moment states remain in the model's original parameter coordinates.

BDT reduces to AdamW in three cases. An inactive treatment ($a_t = 0$), a statistic with no nonzero mode ($\lambda_{\max} = 0$), or a proposal with no depth variation ($U_{c,t} = 0$) leaves the proposal unchanged, $V_t^{\mathrm{can}} = U_t$.

The transformation is bounded. Let $V_{c,t}^{\mathrm{can}} = H V_t^{\mathrm{can}}$ be the centered canonical proposal. On the centered subspace, M_t is positive definite with condition number at most κ. Using the Frobenius inner product, the centered proposal satisfies

$$\cos(V_{c,t}^{\mathrm{can}}, U_{c,t}) \geq \frac{2\sqrt{\kappa}}{\kappa + 1}. \tag{5}$$

Thus κ limits both relative attenuation and rotation away from AdamW. Appendix B gives the derivation. Proposal energy determines the transformation, and its optimization utility is tested empirically.

Gate, up, and down projections maintain separate histories. The original treatment leaves non-FFN parameters on AdamW. Implementation and computational costs are detailed in Appendices A–B.

2.3 Changing correspondence at matched update strength

Changing correspondence can alter both proposal orientation and the strength of departure from AdamW. Our control retains the principal measures of that departure while changing the correspondence-derived direction.

An experimental *branch* is a training continuation under one update rule. At the control branch's current state, let U be its raw AdamW proposal and V^{can} its hypothetical canonical-BDT proposal. Write $u = HU$ for the centered component and $U_m = (I_R - H)U$ for the common component. Fixed independent hidden-unit permutations define an alternative chart with its own BDT history. Transforming in that chart and returning to stored coordinates gives the scrambled candidate W^{scr}.

We extract the part of its displacement orthogonal to centered AdamW. For $u \neq 0$, let $\Pi_u^{\perp}$ project a displacement X onto the centered subspace orthogonal to u, and let d_{scr} be the resulting unit scrambled direction:

$$\Pi_u^{\perp}(X) = HX - \frac{\langle HX, u \rangle_F}{\|u\|_F^2} u, \qquad d_{\mathrm{scr}} = \frac{\Pi_u^{\perp}(W^{\mathrm{scr}} - U)}{\|\Pi_u^{\perp}(W^{\mathrm{scr}} - U)\|_F}. \tag{6}$$

The projection removes the common-depth and AdamW-parallel components before normalization.

Let $\theta = \angle(u, HV^{\mathrm{can}})$ be canonical BDT's centered angular departure at this state. The matched proposal V^{match} uses the scrambled orthogonal direction with the magnitude needed to reproduce that angle:

$$V^{\mathrm{match}} = U_m + \cos\theta\, u + \|u\|_F \sin\theta\, d_{\mathrm{scr}}. \tag{7}$$

The three terms have a direct interpretation. U_m restores the common-depth component. The $\cos\theta\, u$ term retains the component parallel to centered AdamW. The final term supplies the required orthogonal component in the scrambled direction.

The resulting proposal matches canonical BDT in common component, pooled centered norm, and centered angle from AdamW. Its displacement magnitude is consequently matched as well:

$$\|HV^{\mathrm{match}} - u\|_F = 2\|u\|_F \sin(\theta/2) = \|HV^{\mathrm{can}} - u\|_F. \tag{8}$$

What changes is the AdamW-orthogonal orientation supplied by the correspondence.

All matching quantities are computed at each branch's current state. Each control proposal is matched to the hypothetical canonical BDT proposal at that state, and trajectories can subsequently diverge. Individual-layer norm allocation and temporal direction consistency are not separately matched. Appendix B gives degeneracy rules and numerical checks.

This construction tests correspondence-derived orientation beyond the matched measures of update strength. Section 3 evaluates the comparison and uses known ancestry to identify the relationships supporting the effect.

3 Correspondence reflects ancestry, not privileged labels

3.1 Shared-state tests of directional utility

We compare AdamW, BDT, and the matched control in two 124M language-model settings. Explicit-cloned S1 combines depth-specific attention with a shared FFN branch and separately stored private FFNs initialized by copying. The second setting uses ordinary Dense Transformers enlarged from six to twelve layers by copying pretrained blocks. Each layer has separate storage and no shared FFN branch. BDT treats gate, up, and down proposals in both settings.

Shared-state protocol. Each family supplies one update-200 state, including weights, Adam moments, proposal histories, and random state. All three continuations restore it and train to update 3,000 on S1-B FineWeb-Edu (Penedo et al., 2024). For grown models, these update numbers count from copying. Final reporting uses a fixed validation slice after every run in the cohort reaches its prescribed endpoint. Paired Student-t intervals use seeds or parents as independent units, with adjustment for two primary contrasts (Appendices A, O).

Matched-direction result. BDT beats both comparators on all six fresh S1 seeds (Table 1). Its improvement over AdamW establishes constructive benefit, and the matched comparison identifies useful correspondence-derived direction beyond the specified strength measures. All 50,400 transformations satisfy the matching tolerances (Appendix M).

Transfer after pretrained copying. Six pretrained six-layer parents are stacked into twelve-layer models with independently stored weights and Adam moments. Using the same FFN-only BDT treatment, BDT outperforms both AdamW and the matched control on all six parents. The same ordering also holds on a separate evaluation sample of 4,096 screened documents (Table 1, Appendix E).

Table 1: **Constructive benefit and matched-correspondence evidence.** BDT-minus-comparator loss (nats/token). Paired 97.5% intervals adjust two primaries. Document evaluation reuses growth models. Standalone GS and cost measurements appear in Appendices L and N.

Setting / evaluation	Comparator	Mean difference	Adjusted 97.5% interval
S1 / final	AdamW	-0.002697	$[-0.004215, -0.001178]$
	Matched	-0.006099	$[-0.008102, -0.004096]$
Growth / final	AdamW	-0.002204	$[-0.002289, -0.002119]$
	Matched	-0.003148	$[-0.003229, -0.003067]$
Growth / documents	AdamW	-0.002219	$[-0.002270, -0.002168]$
	Matched	-0.003215	$[-0.003335, -0.003095]$

The combined FFN-and-attention BDT policy retains its improvement over AdamW after separate learning-rate selection and across the tested fast and memory-saving execution stacks (Appendices I and S).

3.2 ANCESTRY SEPARATES RELATIONSHIPS FROM LABELS

To isolate ancestry from chart repetition, we use six distinct charts twice each. Layers j and $j+6$ share a pretrained ancestor. Let P_j be the first descendant's chart and σ a fixed derangement, with $\sigma(j) \neq j$. Only the second descendant's chart changes:

	first descendant of j	second descendant of j
preserved	P_j	P_j
reassigned	P_j	$P_{\sigma(j)}$

(9)

BDT still transforms all depths jointly (Figure 1). The maps are inverted before own-state matching by Equation (7), with model storage fixed (Appendix F).

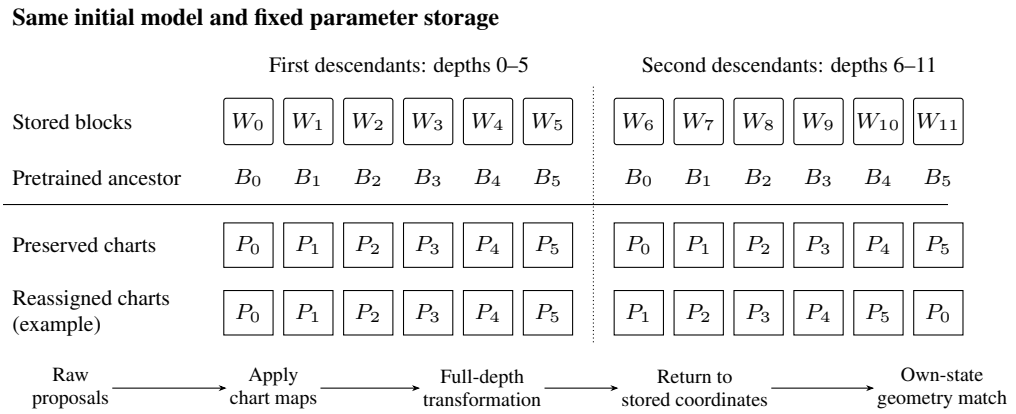

Figure 1: **Preserving ancestry versus reassigning the same charts.** Each chart appears twice. Preservation gives true descendants a shared chart. Reassignment breaks the ancestry match with the same counts. The cyclic shift is one example.

Fresh-parent confirmation. Six independently pretrained parents, each with one preassigned chart draw, confirm the balanced contrast. Reassignment is worse on every parent, by 0.003300 nats/token on average, with paired 95% interval $[0.002972, 0.003629]$. Preservation improves on AdamW by 0.002174, whereas reassignment worsens it by 0.001126. Preserved-minus-canonical loss is -0.000003, descriptively (Figure 2).

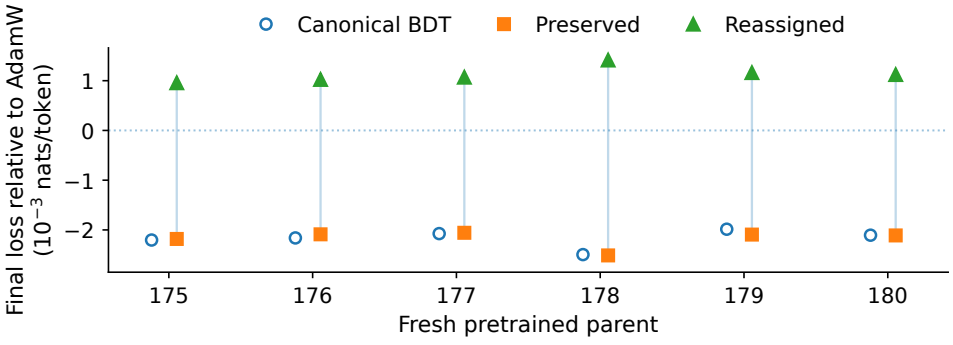

Figure 2: **Relationships, not privileged labels.** Final loss relative to each parent's AdamW anchor. Lines connect preserved and reassigned outcomes. Hollow canonical markers are offset. Preservation and canonical BDT improve on AdamW for every parent.

The ancestry contrast also holds under adjacent copying, removal of the imposed angular match, and a longer horizon. Longer-horizon superiority to AdamW remains unresolved (Appendices W, Q, T).

4 A RELATIONSHIP IS USEFUL ONLY THROUGH AN OPERATION

The balanced comparison identifies a useful ancestry relationship under BDT. Correspondence specifies related parameters but leaves open how their updates should interact. Descendants can move together or relative to one another. We test how the value of ancestry depends on that operation.

For a fixed pairing of all layers, let A repeat each pair's centered mean and D contain its equal-and-opposite deviations, so $U_c = A + D$. The *pair-average component* A can vary between pairs, whereas U_m is shared across all depths. A coefficient ρ sets the relative strength of the orthogonal components through

$$A + \rho D,$$

followed by restoration of the pooled centered norm and the unchanged U_m. Amplification ($\rho = 2$) emphasizes differences relative to the pair average, and suppression ($\rho = 0.5$) reduces them. We compare true ancestry with a fixed false pairing under the same ramp (Appendices R, W.2). These fixed ancestry components differ from BDT's learned spectral modes. Attenuating those modes need not suppress ancestry differences after norm restoration.

In a two-parent development study, the result is a *crossover*. True ancestry beats false pairing under amplification, while the ordering reverses under suppression. Mean false-minus-true loss is $+0.004554$ nats/token at $\rho = 2$ and -0.004556 at $\rho = 0.5$ (Figure 3). These policies preserve the common component and pooled centered norm, but do not match realized angle or layerwise norms.

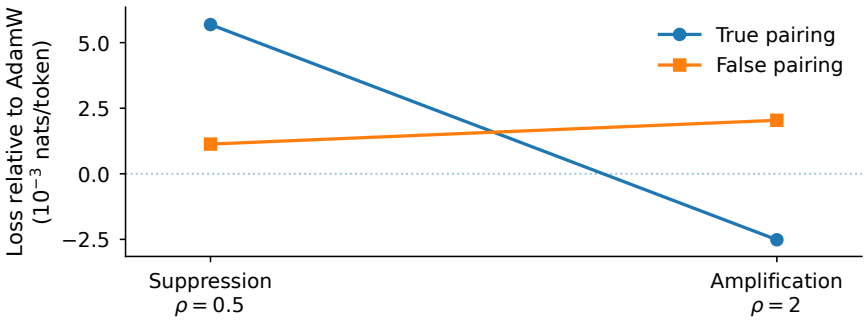

Figure 3: **The operation reverses the relative value of ancestry.** Mean development loss relative to same-parent AdamW in the two-parent study. Lines connect the two tested operations. Appendix R gives all endpoints and parent provenance.

True-pair amplification finishes close to combined FFN-and-attention BDT on the two development parents above, providing a simple reference for testing what full BDT adds. For the reduction and allocation studies, the two copies of each pretrained block occupy adjacent depths, giving ancestry pairs $(2j, 2j + 1)$. In this layout, fixed amplification improves on AdamW by 0.005609 nats/token across four families.

We then ask what BDT adds beyond fixed amplification. The *structured FFN rule* applies bounded conditioning within the pair-difference sector while retaining the sector magnitudes produced by fixed amplification. Adding BDT's attention treatment gives the *assembled comparator*. The combined-BDT reference uses the same attention treatment and full-depth BDT for the FFNs. Across four adjacent-copy parents, the assembled comparator improves on fixed amplification by 0.001199 nats/token on average but finishes slightly worse than combined BDT (Appendix W).

The remaining difference motivates separating the magnitudes assigned to ancestry sectors from the directions within them (Section 5).

5 FROM SECTOR ALLOCATION TO A PROSPECTIVE PREDICTION

5.1 SEPARATING MAGNITUDE ALLOCATION FROM WITHIN-SECTOR DIRECTION

The earlier controls preserve the total centered norm while allowing its distribution across ancestry sectors to vary. Changing that distribution changes the overall update direction. We now separate this allocation effect from changes in direction within each sector.

Let P_A and P_D be the orthogonal projectors onto the pair-average and pair-difference sectors, with $P_A + P_D = H$. For a proposal V retaining the common component U_m, let b_q be the norm (*budget*) in sector $q \in \{A, D\}$ and d_q its unit direction. For nonzero sectors,

$$V = U_m + b_A d_A + b_D d_D, \qquad b_q = \|P_q V\|_F, \quad d_q = P_q V / b_q, \quad q \in \{A, D\}. \tag{10}$$

Orthogonality makes $b_A^2 + b_D^2$ the squared pooled centered norm.

We exchange the FFN sector budgets and directions of full BDT and the structured rule independently, with attention BDT held fixed.

Table 2: **Crossed budget and direction sources.** Four-parent mean development loss relative to combined BDT. These families also enter the prospective test. Rows select budgets and columns select directions.

	Within-sector direction source	
Sector-budget source	Full BDT	Structured rule
Full BDT	0 (reference)	-0.000015
Structured rule	$+0.000514$	$+0.000478$

Full-BDT budgets nearly recover combined BDT with either direction source. Structured budgets retain most of the loss difference. This identifies allocation as the more consequential ingredient here and predicts its advantage under a third direction construction.

5.2 A COMMON DIRECTION CONSTRUCTION

Let K_t be the depthwise window-smoothing matrix and $G_t(U) = K_t U$ its output. It mixes neighboring AdamW proposals with coefficient $0.1a_t$, repeating the first and last layers' own proposals for their missing neighbors (Meng et al., 2026).

For raw proposal U, write $U_m = (I_R - H)U$. Let X denote the full-BDT or structured rule at the current branch state. Its candidate $X(U)$ supplies sector budgets $b_q^X = \|P_q X(U)\|_F$. We normalize the smoothed sectors to unit directions and rescale to these budgets, forming the hybrid proposal V^X:

$$V^X = U_m + b_A^X \frac{P_A G_t(U)}{\|P_A G_t(U)\|_F} + b_D^X \frac{P_D G_t(U)}{\|P_D G_t(U)\|_F}. \tag{11}$$

Both policies share the smoothing rule and attention treatment. Each computes its budgets at its own evolving state and preserves the common component and pooled centered norm. The design was fixed before training (Appendix X).

5.3 THE BUDGET ORDERING TRANSFERS

Full-BDT budgets lower final loss on all four original test parents and all four fresh confirmation parents, with both paired intervals below zero (Table 3). The first test changes the direction construction. The second reproduces the result on new initializations and pretraining streams under the unchanged recipe. The cohorts remain separately analyzed (Appendices X, AA).

Table 3: **Allocation ordering under a third direction construction.** Full-BDT-budget minus structured-budget final loss (nats/token). All four signs favor full budgets in each separately analyzed cohort.

Cohort	Mean difference	Paired 95% interval
Prospective test (4 development families)	-0.000618	$[-0.000772, -0.000464]$
Fresh-parent confirmation (4 new families)	-0.000594	$[-0.000633, -0.000555]$

5.4 THE SUCCESSFUL ALLOCATION CHANGES WITH TRAINING

The preceding tests identify sector magnitude allocation as consequential. We now examine the allocation BDT chooses over training. Let $q_D = b_D^2/(b_A^2 + b_D^2)$ denote the fraction of centered proposal energy assigned to moving descendants relative to one another. A larger q_D assigns a larger share to descendant differences. At each retained state from the original prospective trajectories, both candidates use the same incoming proposal with their respective histories. Their fractions therefore compare allocations of the same centered energy.

Full BDT assigns relatively more to differences at phase 400 and less at phase 3,000, for every parent and projection on both branch-state families. Gate means along full-budget branches change from 0.368 versus 0.322 to 0.460 versus 0.509 (Figure 4). Both fractions rise. Their relative ordering reverses.

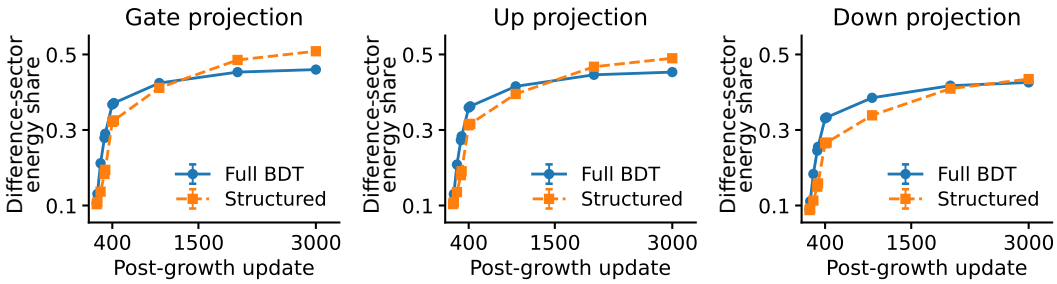

Figure 4: **BDT changes its relative allocation across ancestry sectors over training.** Same-state q_D along the prospective test's full-budget branches. Markers show four-parent means, bars parent ranges, and lines ten retained phases.

Changing raw proposals can vary q_D even under fixed amplification. We therefore also compute the amplification coefficient that reproduces BDT's sector-norm ratio on the same raw proposal. It falls from 2.21 to 1.81 for gate and from 2.34 to 1.96 for down between phases 400 and 3,000, against the structured rule's post-ramp coefficient two (Appendix Z).

These diagnostics characterize the allocation profile rather than establish that its time variation is necessary for the gain.

6 DISCUSSION AND CONCLUSION

Depth-coupled optimization begins with a choice: which parameters across layers are treated as corresponding. BDT makes that choice directly testable. Changing correspondence changes optimization outcomes at matched common component, centered magnitude, and angular departure from AdamW. Pretrained copying identifies a useful relationship. Shared maps retain the benefit when they connect descendants of the same block. Balanced reassignment loses it, locating useful structure in inherited relationships beyond the coordinate labels.

That relationship is the first part of the explanation. The amplification–suppression crossover shows that its relative value depends on the operation. Decomposing the update separates magnitude allocation from within-sector direction. In the crossed controls, the stronger effect follows allocation. The predicted ordering survives a third direction construction and is confirmed separately on four fresh pretrained parents. The original prospective trajectories further reveal a dynamic profile, emphasizing differences relatively more early and less late than fixed amplification.

BDT contributes an optimizer transformation that improves on AdamW in the tested settings and exposes consequential choices in cross-layer optimization. A depth-coupled rule acts on a chosen relationship, through a chosen operation, with a chosen distribution of update magnitude. The interventions separate these choices and establish their consequences.

Once an optimizer couples layers, correspondence becomes part of its optimization geometry. The correspondence map determines what cross-layer structure the optimizer can exploit.

AI USE STATEMENT

Generative AI assisted experiment planning, software implementation, analysis, literature retrieval, and manuscript preparation. Empirical values originate in recorded training and evaluation outputs; they are not generated estimates. The author is responsible for verification, attribution, and the claims in the submitted work.

REPRODUCIBILITY STATEMENT

The appendices specify model families, restored states, data realizations, update rules, checks, statistical units, fixed endpoints, and per-parent outcomes. Table/figure regeneration scripts and authenticated score extracts accompany this revision. Appendix X documents distributed local verification and REPORT evaluation with the frozen numerical kernel; Appendices U, V, and Y distinguish source-record and diagnostic checks from end-to-end reproduction. Appendix AA adds the fresh-parent allocation confirmation and distinguishes the supplied final report from the earlier live record audit. Reused parents and recovery compute are accounted for separately.

ETHICS STATEMENT

The experiments use public web-derived language data and inherit their provenance and content limitations. Validation cross-entropy does not assess factual reliability, deployment safety, or downstream capability. The study evaluates training procedures rather than deploying a user-facing system.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive transformers: Effective parameter sharing with layer-wise LoRA. In *International Conference on Learning Representations*, 2025.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Wenyu Du, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. Stacking your transformers: A closer look at model growth for efficient LLM pre-training. *arXiv preprint arXiv:2405.15319*, 2024.

Tim Tsz-Kit Lau and Weijie Su. Symmetry-compatible principle for optimizer design: Embeddings, LM heads, SwiGLU MLPs, and MoE routers, 2026. arXiv:2605.18106v4, June 22, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Haoming Meng, Anton Sugolov, and Vardan Papyan. Gradient smoothing: Coupling layer-wise updates for improved optimization. *arXiv preprint arXiv:2606.30813*, 2026.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Devender Singh. The loss does not see the basis, but adam does, 2026. arXiv:2608.05136v1, August 5, 2026.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Peihao Wang, Rameswar Panda, Lucas Torroba Hennigen, Philip Greengard, Leonid Karlinsky, Rogerio Feris, David Daniel Cox, Zhangyang Wang, and Yoon Kim. Learning to grow pretrained models for efficient transformer training. *arXiv preprint arXiv:2303.00980*, 2023.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity, 2025. arXiv:2410.08198v3, June 11, 2025.

SUPPLEMENTARY MATERIAL

Reading guide. Appendices A–B (pp. 12–15) specify BDT and its matched control. Appendices C–G (pp. 16–22) give the fresh-seed, pretrained-growth, and ancestry results. For the operation and allocation argument, see Appendix R (p. 38) for the crossover, W (p. 47) for structural refinement, X (p. 49) for prospective transfer, and AA (p. 57) for fresh-parent confirmation. Appendix Z (p. 53) characterizes the original prospective trajectories.

The remaining appendices cover attention scope, initialization, comparison with standalone Gradient Smoothing, cost, horizon, execution compatibility, and supporting mechanistic interventions. Each retains its original cohort and evaluation protocol.

Reading conventions. Study IDs link the descriptive experiments to their source records. DEV labels development evaluation, and REPORT labels final reporting after the relevant cohort endpoints are locked. An endpoint is a study's fixed terminal checkpoint. A cohort lock records all required endpoint identities before reporting. Protocol details distinguish record-level checks from fresh source, checkpoint, or training reproduction. LR denotes learning rate, EMA an exponential moving average, and RNG the random-number-generator state.

A MATCHED-DIRECTION EXPERIMENT: IMPLEMENTATION AND PROVENANCE

The matched-direction comparison (CMS1) supplies the fresh-seed evidence in Section 3.1. This appendix records the model, training recipe, shared-state construction, and evaluation protocol used for that comparison.

Display tables round losses and differences to six decimal places unless a smaller effect requires additional digits. Companion numerical files retain their recorded precision. Display precision describes the presentation rather than measurement accuracy.

Reproducibility and assistance. The retained experiment archives include per-parent endpoints, recomputation scripts, protocols and source identifiers, and an explicit verification boundary. Record checks are distinguished from fresh checkpoint-tensor audits. The public web-derived corpora inherit content and provenance limitations. AI assisted software, analysis, and manuscript preparation. Completed confirmations and reused-parent development studies retain their separate evidential status. The new document evaluation supplements rather than replaces earlier reporting.

Model details. All model families use RMS normalization, rotary attention, tied token/output embeddings, and bias-free SwiGLU FFNs. At S1 depth r, let x be the incoming residual state and e_r its depth embedding. The intermediate states x' and y are the attention-updated residual and normalized FFN input. The layer output is x_{r+1}, computed as

$$\begin{aligned} x' &= x + \mathrm{Attn}_r(\mathrm{RMSNorm}_1(x + e_r)), \\ y &= \mathrm{RMSNorm}_2(x' + e_r), \\ x_{r+1} &= x' + \mathrm{SharedFFN}(y, \mathrm{route}(y)) + \mathrm{PrivateFFN}_r(y). \end{aligned}$$

The 124M shared bank has 3,016 units and activates 256 per depth; private width is 1,792. A shared router with hidden width 128 chooses among twelve fixed masks. Training uses hard straight-through Gumbel routing, temperature 1.5 to 0.5, balance coefficient 0.01, and router-logit penalty 0.0001. Evaluation uses deterministic argmax. Shared normalizations and depth embeddings remain part of S1. Explicit private FFNs have independent storage initialized from the same base matrices.

In the 501.9M transfer, the bank has 8,386 units, shared active width 704, private width 4,928, model width 1,536, and twelve heads. Ordinary Dense has separate attention, FFN, and normalization modules per layer, without the shared bank or router.

Data. The main realization, S1-B, comprises three billion training IDs and five million validation IDs, represented as unsigned 16-bit arrays. The exact lengths are six billion and ten million bytes. A second realization, OWN-S1-R1, is used for the matched-random/reflected controls and the longer-budget/growth screens. Comparisons stay within each realization. Neither absolute losses nor

seed-level effects are pooled across them. These are independently hashed token realizations of the same named corpus, not distinct external benchmark datasets.

Evaluation labels and slices. DEV denotes development evaluation and REPORT denotes final reporting after the study's endpoint lock. Both labels identify evaluation roles within the retained validation data. Exact slices and any additional document screening are specified by study.

CMS1's DEV slice is $[0, 1{,}050{,}624)$ and REPORT is $[2{,}500{,}000, 3{,}550{,}624)$ in the S1-B validation array. Each contains 2,048 sequences with 513 input IDs and 512 predicted tokens. This establishes disjoint indexed segments, not independently collected text or verified document-level separation. The historical exact-fork studies use their archived fixed seed-indexed evaluation convention. The ordinary-Dense factorial's evaluator uses the same function of seed and batch index for both initialization conditions. The local full-step audit uses its separately specified 2,048-token, context-256 probe.

Principal training hyperparameters. CMS1 uses context 512, microbatch 16, four accumulation steps, BF16 forward/backward computation with FP32 parameter storage, and 3,000 updates. AdamW has $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, and gradient clipping at 1. Peak LR is 0.0012, with 200-update warmup and cosine decay to one tenth of peak; BDT uses cap three and its separately defined activation ramp. A full path contains 98,304,000 training-token exposures. Forming and applying each depth operator costs $O(R^2P)$; persistent statistics cost $O(R^2)$ per projection, with $O(R^3)$ eigendecomposition at observation steps.

Data and random-state pairing. CMS1 draws one logical batch per update from a deterministic function of seed and update, and splits it into four microbatches. The shared parent includes the model, optimizer, RNG, canonical covariance history, scrambled-chart history, and permutations. CMS1 records the complete 2,800-update batch/RNG history and full state/proposal hashes at designated checkpoints. Initial correspondence is fixed by the model construction. Later candidate operators depend on each branch's evolving state. Within each seed, histories and first-active-step pre-treatment inputs match, but active-policy model trajectories are expected to diverge.

Strict execution and validation. The strict runtime enables strict deterministic algorithms, the cuBLAS setting `CUBLAS_WORKSPACE_CONFIG=:4096:8`, cuDNN determinism, and automatic attention-backend selection while disabling benchmarking and matrix-multiplication TF32. The preparatory execution study observed nonidentical repeated gradients under the preceding runtime, including among ordinary Adam executions, and identical recorded prefixes under the strict setting. That finding localizes the earliest observed discrepancy upstream of the optimizer but does not isolate a unique responsible kernel.

The CMS1 target-GPU gate uses existing engineering seed 127: two 200-update parents, six active continuations (two per scientific policy), and two reference continuations through update 420. One scrambled continuation is deliberately interrupted and resumed in another process. The gate requires exact recorded repeat/resume comparisons and archived AdamW/BDT agreement, plus equal first-active-step gradients and raw proposals. The gate has 2,160 optimizer updates. Scientific execution comprises six 200-update parents plus eighteen 2,800-update continuations, totaling 51,600 executed optimizer updates. A modeled training path still represents 3,000 updates. Shadow candidates and validation add compute.

Registered order, stopping, and evaluation. The within-seed arm orders are Adam–BDT–matched (147), BDT–matched–Adam (148), matched–Adam–BDT (149), matched–BDT–Adam (150), BDT–Adam–matched (151), and Adam–matched–BDT (152). Failures of source, state, nonfinite, or matching invariants stop the queue. OOM retries preserve the recipe and last committed state. They do not change batch size or learning rate. The final eighteen result hashes are locked before the first REPORT evaluation. The final endpoint is update 3,000, not a selected checkpoint. No seed or trial is dropped from the registered analysis.

Record audit. The final audit recovered six parent result files, eighteen continuation files, eighteen REPORT files, contracts, initialization records, checkpoint-pointer metadata, the cohort lock, registration, cache identities, source manifests, and logs. Per-seed state links and histories, result-to-report links, and all registered contrasts reproduce. Fifteen study-manifest entries and twenty frozen-backend

entries were downloaded and hash-verified. This is a record/source audit: full checkpoint tensor payloads and all gate sample arrays were not freshly reloaded during the final inspection. The source-manifest digest, protocol, and raw-archive checksums are retained in the experiment archive.

B MATCHING CONSTRUCTION AND ALGEBRAIC CHECKS

Permutation transport. For an FFN with weights (W_g, W_u, W_d) and hidden permutation matrix P_r, the reindexed tensors are $(P_r W_g, P_r W_u, W_d P_r^\top)$. Since elementwise activation and product commute with hidden-unit permutation, the represented FFN function is unchanged. The same transport applies to signed proposal tensors. CMS1 maintains original model storage. A separate collection of independently chosen P_r defines the scrambled proposal chart. Permutations are fixed per seed with a SHA256-derived local CPU generator, and have a shared hidden index across gate/up rows and down columns within a depth.

The scrambled second moment is accumulated from transported raw proposals throughout the same inactive prefix. At an active state, the transformed scrambled candidate is transported back before centering and matching. Centering independently in incompatible charts would change the meaning of the common component. Equations (6)–(7) explicitly restore the canonical common component and remove both the common and Adam-parallel components of the candidate displacement.

Numerical implementation. The frozen BDT candidate and proposal statistics use the archived proposal-precision implementation. The small covariance eigendecomposition is symmetrized and performed in FP64 before storing the eigensystem at proposal precision. Matching geometry uses FP64 reductions and is cast back to FP32 for application. The orthogonal scrambled residual is computed from the difference $W^{\mathrm{scr}} - U$ before centering, which reduces cancellation near identity, and is projected a second time to suppress numerical common/parallel residuals.

If the canonical candidate is exactly the raw proposal, matching returns the raw proposal exactly. A canonical sine at most 10^{-7} is a logged numerical-zero event. Otherwise the orthogonal scrambled residual divided by $\|U_c\|_F$ must exceed 10^{-7}. An undefined direction stops the study rather than receiving a random replacement. Applied-value tolerances are 10^{-5} for relative centered norm and normalized common-component error, 5×10^{-6} radians for angle, and $2 \times 10^{-6} + 2 \times 10^{-4} d$ for normalized displacement with target d. All actual scrambled continuations complete without either zero-angle category after onset.

The largest recorded amplification of the scrambled orthogonal residual is 34.3817 in CMS1 and 128.6800 in GCT2. This is an explicit property of converting a weak raw-scrambled candidate into a matched-strength directional control. The experiment does not assert that the control preserves the raw operator's spectrum or every layer's individual update norm. Table 25 reports the achieved matching-summary maxima; Appendix M adds sparse distribution summaries. These are logged diagnostics, not a newly reconstructed tensor-level audit.

B.1 ONE MATCHED-STRENGTH PROPOSAL

Algorithm 1: One matched-strength proposal at one current state.

Input: raw AdamW increment U, canonical and alternative-chart histories, and fixed chart maps.

Observe: update each history from the same raw proposal in its own chart at the prescribed cadence.

Construct: form canonical V^{can} and transported scrambled W^{scr}; measure $\theta = \angle(HU, HV^{\mathrm{can}})$.

Match: project $W^{\mathrm{scr}} - U$ into the centered subspace orthogonal to HU, normalize it, and apply Equation (7).

Check and apply: enforce the fixed identity/degeneracy rules and cast-proposal tolerances, then apply the assigned policy. Weight decay is separate. All arms compute both candidates. Only the applied proposal differs.

B.2 PRESERVATION AND ROTATION

The first argument bounds the rotation produced by BDT. The second checks the geometry of the matched control.

For a positive definite operator on the centered subspace, let m and M bound its eigenvalues $g_i \in [m, M]$. Let p_i be the normalized squared mass of the centered proposal U_c in each eigendirection, summed over feature coordinates. The gain mean and second moment are $\mu = \sum_i p_i g_i$ and $\nu = \sum_i p_i g_i^2$. Since $(g_i - m)(g_i - M) \le 0$, $\nu \le (m + M)\mu - mM$, giving

$$\cos^2(M_t U_c, U_c) = \frac{\mu^2}{\nu} \ge \frac{\mu^2}{(m + M)\mu - mM} \ge \frac{4mM}{(m + M)^2}.$$

The last expression follows by minimizing the rational function at $\mu = 2mM/(m + M)$. BDT's gains have ratio at most κ, so taking square roots yields Equation (5). Positive norm restoration does not change the angle. Since $\mathbf{1}^\top B = 0$, the transformed component stays centered; orthogonality of common and centered components proves full-norm preservation. These are operator invariants, not a quality theorem.

For the matched control, $\langle u, d_{\mathrm{scr}} \rangle_F = 0$, $\|d_{\mathrm{scr}}\|_F = 1$, and $H d_{\mathrm{scr}} = d_{\mathrm{scr}}$. Equation (7) consequently has centered norm $\|u\|_F$ and cosine $\cos\theta$ with u. Expanding its squared difference from u gives $2\|u\|_F^2(1 - \cos\theta)$, proving Equation (8). These quantities concern the signed data proposal before decay, not a loss-space or full-function distance.

Why shared and independent permutations differ. For a fixed linear depth operator S, let P_r permute the coordinates of flattened layer block r. Let $T = \mathrm{diag}(P_1, \ldots, P_R)$ collect those permutations, and let I be the identity on a single flattened block. The (i, j) block of the commutator $T(S \otimes I) - (S \otimes I)T$ is $S_{ij}(P_i - P_j)$. A common within-block permutation commutes; independent layer permutations generally do not whenever the corresponding layers are coupled. This algebra explains why choosing a chart can change a coupling rule. It does not predict the sign of the loss effect. CMS1 supplies an empirical test with a data-dependent operator and controlled proposal geometry, not a claim that this commutator is a novel theorem.

B.3 BASE-OPTIMIZER EQUIVARIANCE AND CROSS-DEPTH CORRESPONDENCE

Write w, g, m, v for parameters, gradients, and the first and second Adam moments, and let $D(m, v)$ denote the Adam data increment. Let P permute coordinates within parameter groups with equal hyperparameters, and transport (w, g, m, v) to (Pw, Pg, Pm, Pv). Elementwise squaring, square root, bias correction, and division commute with P. Consequently the Adam data increment obeys $D(Pm, Pv) = PD(m, v)$, and scalar decoupled weight decay also commutes. This exact-arithmetic statement requires consistent optimizer-state transport and equivalent stochastic inputs. It is not a full-production finite-precision null test. Arbitrary orthogonal transformations generally do not commute with coordinatewise adaptivity (Xie et al., 2025; Singh, 2026).

For SwiGLU, the hidden-neuron transformation is $(W_g, W_u, W_d) \mapsto (PW_g, PW_u, W_d P^\top)$. The nonlinearity and elementwise product do not permit substituting arbitrary hidden rotations for P. Block-specific symmetry groups therefore matter (Lau & Su, 2026). A cross-depth inner product $u_r^\top u_s$ changes to $u_r^\top P_r^\top P_s u_s$ under independent layer permutations. An equivariant base optimizer need not make an additional cross-depth operator independent of these maps. Simultaneously relabeling parameters and explicit correspondence maps is distinct from changing the maps while retaining model storage. The latter is the principal directional intervention.

C ALL PRINCIPAL ENDPOINTS AND STATISTICAL UNITS

Each row below is an independent training seed. The two contrasts answer the constructive and correspondence questions separately.

Table 4: All six CMS1 final REPORT endpoints and registered paired differences. Values are rounded only for display.

Seed	AdamW	BDT	Matched	BDT − Adam	BDT − matched
147	3.891970	3.887365	3.896162	−0.004605	−0.008796
148	3.899320	3.896562	3.901058	−0.002758	−0.004496
149	3.897200	3.896253	3.901105	−0.000947	−0.004852
150	3.898831	3.895917	3.901761	−0.002914	−0.005844
151	3.895459	3.892799	3.899592	−0.002660	−0.006793
152	3.890453	3.888156	3.893967	−0.002297	−0.005811

Let δ_s be the paired loss difference for seed s, with mean $\bar{\delta}$ and sample standard deviation s_δ. Write $t_{\nu,p}$ for the Student-t quantile with degrees of freedom ν and cumulative probability p. The ordinary interval is $\bar{\delta} \pm t_{5,0.975} s_\delta / \sqrt{6}$. For the two registered comparisons, each 97.5% interval uses $t_{5,0.9875}$; Bonferroni gives family coverage at least 95% under the assumptions of the individual intervals. The SDs are 0.00117587534 for BDT–Adam and 0.00155114687 for BDT–matched. No eighteen-independent-sample analysis or per-token significance test is used. Intervals condition on this fixed recipe, seed-generating procedure, and reused corpus realization. They do not account for a new hyperparameter search or distribution shift. A failure to resolve another contrast would not imply equivalence.

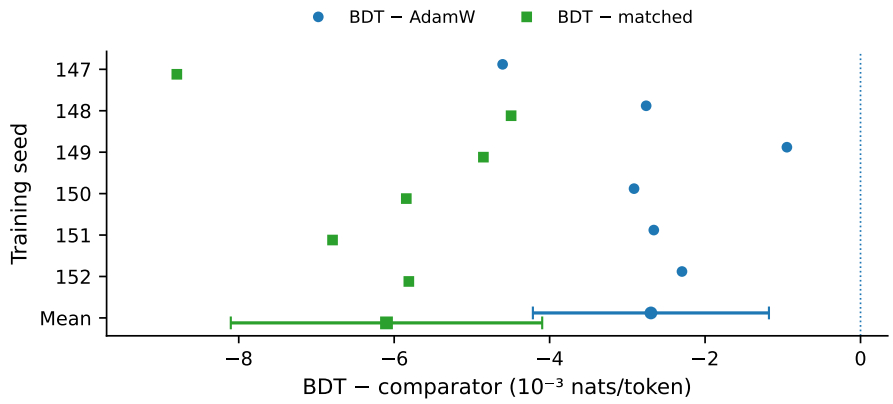

Figure 5: **Direction and constructive benefit in one comparison.** Each seed contributes two paired final REPORT contrasts. Negative values favor canonical BDT. The bottom row shows their means and the registered multiplicity-adjusted intervals. All six seed-level differences in both comparisons favor BDT. The larger control separation is not the practical improvement over AdamW.

D PRETRAINED-GROWTH TRANSFER: PROTOCOL AND CONTROLS

The two-parent growth pilot (GCT1) tests the same three optimizer policies after pretrained copying. Its model and state-transfer protocol also define the later fresh-parent confirmation.

GCT1 uses fresh parent-training seeds 159–160 on S1-B. Each six-layer Dense parent has 81,074,688 parameters and is trained for 7,630 updates at peak LR 0.0008, 200-update warmup, cosine decay to 0.1 of peak, and 32,768 exposures per update. The whole-block mapping is $[0, 1, 2, 3, 4, 5, 0, 1, 2, 3, 4, 5]$, producing a 123,551,232-parameter Dense12 model. Every mapped parameter and both Adam moments have independent target storage. Embeddings and final normalization are copied once. No learned growth mapping is introduced.

The mapping checks record 86 mapped named parameters, exact transported weights and moments, independent storage, and preserved Adam age 7,630. RNG is inherited and the next global data update is 7,631. Both twelve-depth covariance histories start at zero, with centered dimension eleven. Their histories and fixed hidden-unit permutations are accumulated/saved during one 200-update post-growth Adam prefix. The three continuations restore the full saved prefix, including its optimizer and random states. Post-growth phase counts updates since copying, while global Adam age includes parent pretraining. Their final phase is 3,000 and final global Adam age is 10,630. Post-growth peak LR is 0.0006. Tempering acts over all twelve FFNs, not separately normalized copied pairs: with two roles, a positive operator on the one-dimensional centered subspace is scalar, and centered-norm restoration would cancel it.

The operation is not function-preserving. The seed-159 small-parent, immediate-grown, and shared-prefix DEV losses are 3.760558831, 4.212327722, and 3.907647571. Seed 160 gives 3.765389726, 4.253142221, and 3.909670882. The immediate jumps, +0.451768892 and +0.487752495, are recorded rather than used to select parents. All six terminal losses are in Table 5.

The independent replication unit is the small-parent training seed; eighteen thousand projection diagnostics do not create additional independent replications. The two BDT–Adam differences are -0.002174401656 and -0.002115920186; BDT–matched gives -0.003190658987 and -0.003116741776. Both means are favorable, with no confirmatory significance test performed. The pilot was not extended on the basis of these outcomes.

Within each seed, all three recorded prefix identities, complete 2,800-step continuation batch/RNG histories, and first-active-step raw inputs match. Each actual scrambled continuation records 8,400 matched projection events, no identity/numerical-zero events, and maximum residual amplification 84.3453 (159) or 122.3189 (160). Maximum logged angular mismatch is 3.84×10^{-11} radians. The amplified orthogonal residual is part of the specified matched-direction construction. No claim is made that it preserves the original scrambled operator spectrum or individual layer norms. The full-size gate tests growth/state transport, active-policy repeats, archived-reference agreement and saved-state resume through the ramp. Final record verification checks contracts and checkpoint links, not a fresh reconstruction of every applied tensor.

Table 5: **Pretrained-growth transfer (GCT1).** Final DEV loss after the same grown parent and post-growth prefix. Both seeds favor BDT against both comparators. This is a two-parent pilot, not a confirmatory REPORT cohort.

Seed	AdamW	BDT	Matched	BDT − Adam	BDT − matched
159	3.661593	3.659418	3.662609	−0.002174	−0.003191
160	3.664348	3.662232	3.665349	−0.002116	−0.003117

E SIX-PARENT GROWTH CONFIRMATION AND DOCUMENT-LEVEL SUPPLEMENTAL EVALUATION

E.1 FIXED COHORT, UNCHANGED TREATMENT

GCT2 uses six new independently initialized and trained six-layer parents, seeds 161–166, on S1-B. The parameters, optimizer settings, growth mapping $[0, 1, 2, 3, 4, 5, 0, 1, 2, 3, 4, 5]$, optimizer-age transport, and token indexing are those of the pilot. Each small parent trains for 7,630 updates at

peak LR 0.0008. The twelve-layer phase trains for 3,000 updates at peak LR 0.0006, including a shared 200-update prefix. The branch orders are Adam–BDT–matched (161), Adam–matched–BDT (162), BDT–matched–Adam (163), BDT–Adam–matched (164), matched–Adam–BDT (165), and matched–BDT–Adam (166). Each of two lanes handles three complete parent families. Every policy occupies each position once within a lane.

The original FFN-only treatment and symmetric hypothetical-candidate calculations are unchanged. Neither the faster diagnostics-lite implementation nor the later attention extension is substituted. Scientific execution contains 45,780 six-layer parent updates, 1,200 twelve-layer shared-prefix updates, and 50,400 twelve-layer continuation updates. These 97,380 mixed-size campaign updates are not the cost of one model: each complete path pays 7,630 parent and 3,000 post-growth updates, totaling 348,323,840 token exposures. Engineering tests and retries are additional.

Every final training record and checkpoint link is locked before the eighteen original REPORT evaluations. The fixed slice is $[2{,}500{,}000, 3{,}550{,}624)$, with 2,048 records and 1,048,576 scored targets. The same prespecified paired-t analysis is used as in CMS1: ordinary 95% and individual 97.5% intervals for the directional and constructive contrasts. Both adjusted upper bounds are negative. The pilot parents 159–160 do not enter this analysis. Small between-parent standard deviations are reported as observed properties of this fixed experiment, not assumed population invariance or a power guarantee.

Table 6: All original GCT2 REPORT endpoints. Six independently pretrained parents; lower is better.

Parent	AdamW	BDT	Matched	BDT − Adam	BDT − matched
161	3.619368	3.617222	3.620442	−0.002146	−0.003220
162	3.622452	3.620120	3.623263	−0.002331	−0.003143
163	3.619942	3.617763	3.620913	−0.002179	−0.003150
164	3.621547	3.619336	3.622550	−0.002211	−0.003213
165	3.618084	3.615900	3.618957	−0.002184	−0.003056
166	3.623472	3.621299	3.624404	−0.002172	−0.003104

E.2 Newly reserved document sample

FEV1 is a separately specified supplemental evaluation, not a change to GCT2's reporting endpoint. The sample comes from FineWeb-Edu crawl CC-MAIN-2025-26 at corpus revision `87f09149ef4734204d70ed1d046ddc9ca3f2b8f9`, with the GPT-2 tokenizer pinned at revision `607a30d783dfa663caf39e06633721c8d4cfcd7e`. Selection is independent of model losses. Eligible documents have 513–16,384 tokens. One deterministic 513-token within-document window scores 512 targets. Thus the estimand concerns this eligible-document sample, not an unrestricted crawl average.

The builder uses a fixed stream shuffle (seed 20260911, buffer 10,000), candidate limits, and source-order selection. Entire candidate texts are scanned for exact 64-token matches against the full three-billion-token training and five-million-token validation arrays, respecting EOS boundaries. A single rolling-hash hit conservatively rejects a candidate. A collision can add an exclusion, not miss an exact match. Candidate IDs, normalized URLs/text, and candidate-to-candidate overlap are checked. The final sample locks 4,096 documents and 2,097,152 targets, with source IDs, URLs, lengths, window locations, content/token hashes, and the selection evidence retained. Historical source identifiers were not retained, so exact-token screening does not prove URL-level separation or eliminate every near duplicate. The original DEV/REPORT segments are indexed-disjoint. No stronger document-level independence is inferred from the offsets alone.

The evaluator first waits for original GCT2 training, cohort locking, REPORT scoring, and analysis. It then locks and scores all eighteen checkpoints on the same document set after an outcome-blind evaluator repeatability/model-state gate. Per-document losses are retained, but the independent training units remain six parents. The two-contrast adjustment applies separately to this prespecified supplemental evaluation, not to a claimed four-contrast family combining it with REPORT. Neither text set is used to choose the later scope policies.

Table 7: All FEV1 endpoints for the same eighteen GCT2 models. These are not six additional training parents.

Parent	AdamW	BDT	Matched	BDT − Adam	BDT − matched
161	3.361918	3.359635	3.362993	−0.002282	−0.003357
162	3.368506	3.366302	3.369401	−0.002204	−0.003099
163	3.364237	3.362048	3.365291	−0.002189	−0.003243
164	3.363376	3.361167	3.364399	−0.002209	−0.003232
165	3.361202	3.359023	3.362148	−0.002179	−0.003125
166	3.366596	3.364345	3.367578	−0.002251	−0.003233

Table 8: Paired uncertainty over the six parent seeds, conditional on each fixed evaluation set.

Evaluation / contrast	Ordinary 95% interval	Adjusted 97.5% interval
Original REPORT / AdamW	$[-0.002273, -0.002135]$	$[-0.002289, -0.002119]$
Original REPORT / Matched	$[-0.003214, -0.003082]$	$[-0.003229, -0.003067]$
Document sample / AdamW	$[-0.002261, -0.002178]$	$[-0.002270, -0.002168]$
Document sample / Matched	$[-0.003312, -0.003118]$	$[-0.003335, -0.003095]$

Verification depth. The preserved final audit reads all eighteen original training records and checkpoint-metadata links, eighteen REPORT records, eighteen supplemental score records and document-loss arrays, and the locked document/record bytes. It verifies six distinct small-parent model hashes, shared-prefix identities within families, complete 2,800-update batch/RNG histories, first-active pre-treatment inputs, and recomputed endpoint contrasts. Full model-checkpoint payloads and the complete old token cache were not reloaded during that final audit. The summaries here retain the verified reporting values rather than claiming a new tensor-level rerun. Original and supplemental cohort-lock digests and evidence paths accompany the manuscript.

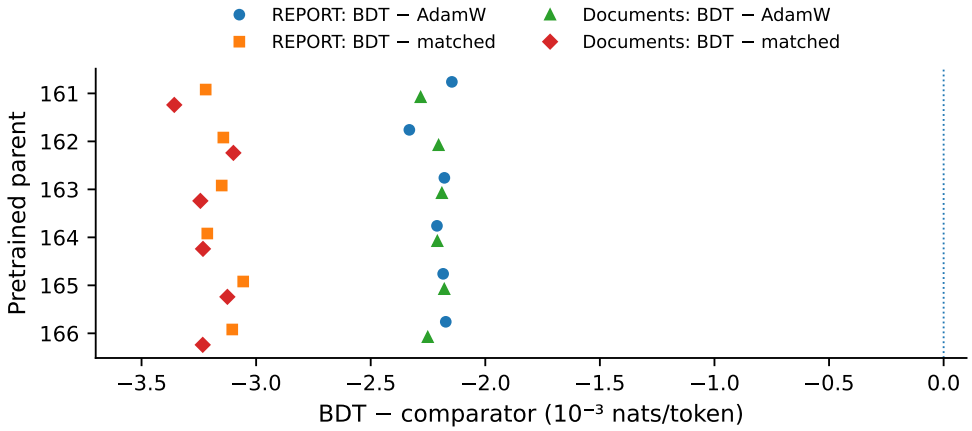

Figure 6: **Growth transfer survives the change of evaluation sample.** Each symbol is one paired final loss contrast for one independently pretrained parent. Both evaluation sets preserve the constructive and directional ordering. Symbols are offset vertically for legibility. The same model families are evaluated twice and the evaluations are not pooled.

F INHERITED CORRESPONDENCE: DEVELOPMENT PROTOCOLS AND ENDPOINTS

F.1 GRADED ANCESTRY DISRUPTION (PCT1)

PCT1 reuses the two GCT1 development parents, seeds 159–160, on S1-B. It retains the phase-zero copying map, named Adam moments and global age 7,630, then uses a shared 200-update prefix and eight 2,800-update continuations per parent. The arms are AdamW, canonical FFN BDT, and two fixed draws at zero, three, and six broken pairs. Peak LR is 0.0006, cap three, and the final DEV slice scores 1,048,576 targets at offset zero. This is not a new REPORT or FEV1 evaluation.

For each draw, first-copy charts P_j have no fixed hidden coordinates. An additional Q_j for each source is constructed to disagree with identity and P_j at every coordinate. A fixed nested pair order chooses which second-copy charts change from P_j to Q_j. Each chart consistently transports gate/up rows and down columns; trained model storage remains unchanged. Alternative histories are accumulated from the common prefix and subsequent raw proposals, never borrowed from another branch. Common-component, pooled-centered-norm and angle matching use each branch's own hypothetical canonical proposal. Neither layerwise allocation nor temporal consistency is constrained to be equal.

Table 9: **All completed PCT1 development endpoints.** Anchors repeat only to locate the two fixed draws within each parent; repeated anchors are not extra executions or replicates. All values are final DEV nats/token.

Parent	AdamW	BDT	Draw	Preserve all	Break three	Break six
159	3.661593	3.659418	0	3.659452	3.661167	3.662597
159	3.661593	3.659418	1	3.659432	3.660350	3.662589
160	3.664348	3.662232	0	3.662207	3.662584	3.665411
160	3.664348	3.662232	1	3.662276	3.663183	3.665393

The primary contrast averages the two fixed draw differences within each parent. The parent contrasts are $+0.003150784411$ and $+0.003160959110$, with descriptive mean $+0.003155871760$. Breaking three pairs instead gives mean $+0.000979364850$ relative to preserving all; breaking the remaining three adds $+0.002176506910$. Every individual draw has the graded ordering. There are two reused development parents, not four independent parent replications. No confirmatory interval is computed.

Record verification. All sixteen original final-result records, contract hashes, final-checkpoint metadata, prefix links, initial identities, full batch/RNG histories and first-active inputs pass the recorded audit. The recomputed contrasts agree with the global analysis. Tensor payloads and full token caches were not newly reloaded in this final readout.

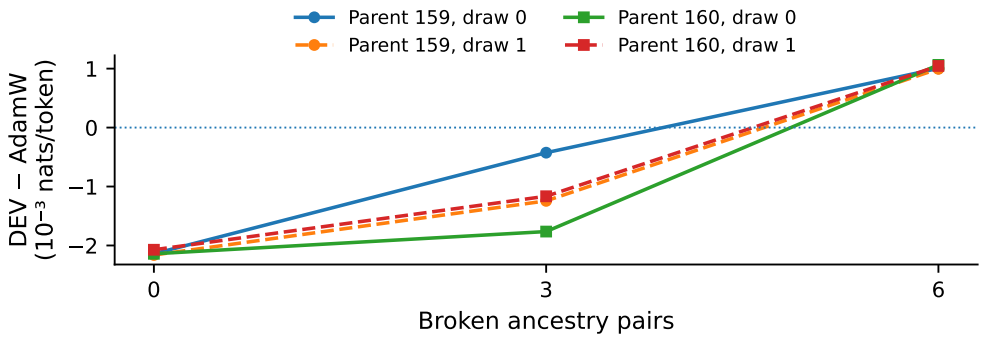

Figure 7: **Graded ancestry disruption in development (PCT1).** Final DEV loss relative to each parent's AdamW anchor. All four fixed parent/draw curves worsen as zero, three, and six pairs are broken. Draws are averaged within parent, not treated as independent replications.

F.2 BALANCED-ASSIGNMENT DEVELOPMENT (BAA1)

BAA1 retains six distinct hidden charts, each appearing exactly twice in both assignments. First descendants use P_j; second descendants use either P_j (preserved) or $P_{\sigma(j)}$ (reassigned), where σ is a fixed derangement of source indices. Distinct charts may incidentally agree at individual coordinates; that agreement is measured rather than rejected. Unlike PCT1, the chart multiset and multiplicities are fixed. Full twelve-layer FFN tempering and own-state common/norm/angle matching are unchanged. Independently tempering and normalizing each two-layer pair would cancel positive scalar tempering in its one-dimensional centered subspace; that is not the operation used here.

Both registered development parents, 159–160, and all twelve arms are complete. Each parent supplies AdamW, canonical FFN BDT, and preserving/reassigned policies for two fixed draws. Draws are averaged within parent, not counted as additional independent parents.

Table 10: **All completed BAA1 development endpoints.** Final S1-B DEV loss. Repeated anchor entries locate each draw and are not new replications.

Parent	Draw	AdamW	Canonical BDT	Preserved	Reassigned
159	0	3.661593	3.659418	3.659485	3.662580
159	1	3.661593	3.659418	3.659467	3.662596
160	0	3.664348	3.662232	3.662230	3.665426
160	1	3.664348	3.662232	3.662174	3.665375

Reassigned-minus-preserved is $+0.003112372011$ for parent 159 and $+0.003198526800$ for parent 160, after draw averaging. The descriptive two-parent mean is $+0.003155449405$. All four draw-level contrasts are positive. Every preserving draw beats its parent's AdamW anchor, whereas every reassigned draw is worse. The preserving-minus-canonical mean is $+0.000013707671$, which is not an equivalence test or a new method gain.

Record verification. The completed record audit passes 124 identity/arithmetic checks, covering both prefixes, all twelve result/checkpoint links, full paired histories, and chart-multiset/multiplicity/assignment records. It does not freshly reload all checkpoint tensors or token-cache bytes.

The two-parent development cohort is analyzed separately from the fresh families in the next section. The one-parent applied-direction diagnostics in Appendix M remain exploratory measurements of parent 160.

G FRESH-PARENT BALANCED-ASSIGNMENT CONFIRMATION

G.1 FIXED CONTRAST, COHORT, AND REPORTING PROCEDURE

BAC1 confirms the balanced-assignment contrast on six newly pretrained Dense parents, 175–180, excluding the development parents 159–160. Each six-layer parent has width 768, FFN width 2,048, twelve heads, context 512, and vocabulary 50,257. It trains for 7,630 updates at peak LR 0.0008 before its blocks and named Adam moments are stacked into the twelve-layer architecture. All mapped tensors have independent storage. Global Adam age and data indexing continue. The post-growth schedule runs for 3,000 updates at peak LR 0.0006, with a shared 200-update AdamW prefix, cap three, and a 200-update treatment ramp. The batch has 32,768 tokens (microbatch 16, accumulation four, context 512); precision is BF16. Covariance observations occur every ten updates with EMA coefficient 0.9.

The four arms are AdamW, canonical FFN BDT, preserved assignment, and reassignment. Each fresh parent receives one preassigned chart draw, using the inherited BAA1 chart generator and fixed derangement. The six charts each appear twice in both chart treatments. All source, state, and own-state geometric-matching checks are retained. The independent unit is the *fresh pretrained parent with its fixed draw*. The six draws are not six extra units beyond those parents.

BAC1 primary comparison. For parent s, $\Delta_s = L^{\mathrm{REPORT}}_{s,\,reassigned} - L^{\mathrm{REPORT}}_{s,\,preserved}$ at post-growth update 3,000. Positive is the prespecified sign. The sole primary estimate is the mean of the six paired Δ_s, with a two-sided 95% Student-t interval (five degrees of freedom). AdamW and canonical-BDT comparisons are descriptive anchors.

All 24 training endpoints are locked before any scientific REPORT evaluation. REPORT uses the existing S1-B validation slice at offset 2,500,000, with 2,048 records and 1,048,576 scored target tokens. This set was used earlier in the project. The confirmation adds fresh training parents, not newly collected evaluation text. The fixed analysis retains all parents and does not extend the cohort or substitute an intermediate checkpoint.

The preparation followed positive development evidence from parent 160. Before the complete BAA1 outcome was available, release was changed from a global positive-outcome dependency to an outcome-independent local completion/engineering dependency. This did not change the six seeds, charts, arms, LR, horizon, primary contrast, or all-endpoints-before-REPORT rule. The confirmed effects are kept separate from the development screen.

G.2 All endpoints and numerical reconstruction

Table 11: **BAC1: all six fresh-parent REPORT outcomes.** Lower loss is better. The last column is the single registered primary contrast. Its positive sign favors preserved ancestry assignment.

Parent	AdamW	Canonical	Preserved	Reassigned	Reassigned − preserved
175	3.621990	3.619787	3.619809	3.622949	+0.003140
176	3.620917	3.618756	3.618828	3.621942	+0.003114
177	3.623219	3.621145	3.621162	3.624290	+0.003129
178	3.619480	3.616984	3.616968	3.620896	+0.003928
179	3.619918	3.617933	3.617825	3.621081	+0.003256
180	3.618388	3.616282	3.616278	3.619512	+0.003234

The primary mean is $+0.003300288692$, sample SD 0.000313229385, and paired 95% interval $[0.002971574612, 0.003629002772]$. All six primary signs are positive. Mean REPORT losses are 3.620652203 (AdamW), 3.618481194 (canonical BDT), 3.618478053 (preserved), and 3.621778341 (reassigned). Relative to AdamW, preserved and reassigned effects are -0.002174150199 and $+0.001126138493$, respectively. All six preserved arms improve over AdamW and all six reassigned arms worsen it. Preserved-minus-canonical is -0.000003141351 on average. No equivalence interval or superiority claim is attached to this small descriptive difference.

The interval quantifies parent-plus-fixed-draw variation conditional on the chosen recipe and reporting text. It is not a population statement over all possible coordinate charts or architectures. Equal chart multiplicities rule out unequal repetition as the explanation for this contrast, but do not match per-layer norm allocation, residual amplification, or temporal direction consistency separately. Those may participate in the effect rather than being excluded by the design.

Evidence provenance. The saved final analysis is timestamped September 15, 2026, 08:34:56 UTC. Its 24-endpoint cohort lock and source/protocol identifiers were read alongside the completed analysis. The paired differences, interval, and constructive-anchor signs are independently recomputed from the unrounded saved rows. Full source/protocol identifiers and derived inputs are recorded in the retained experiment archive. This is a numerical record reconstruction, not a new tensor-level or full-token-cache audit.

H Attention scope and component localization

The scope studies compare FFN, attention, and combined treatment on S1-B (realization B) and OWN-S1-R1 (realization A). The component test separates the attention projections.

ASF1 reuses both S1-B growth-development parents 159–160. RGS1 independently pretrains development parents 167–168 on OWN-S1-R1 using the same six-to-twelve growth procedure. The four scope arms are AdamW, FFN-only BDT, attention-only BDT, and combined BDT. Each new common prefix accumulates the same five proposal histories: FFN gate/up/down, attention QKV, and

attention output. Post-growth LR 0.0006, cap three, cadence ten, onset/ramp 200/200, and endpoint 3,000 remain fixed. The physical optimizer partition and hypothetical-candidate computations are identical across arms.

Native attention proposals and moment preservation. The implementation retains FFN RoleAdamW and native fused AdamW for attention/rest in every arm. It exposes attention data proposals from the native updated moments at the inherited global Adam age and saves pre-step attention weights. Untreated or inactive attention retains the actual native output exactly; treated nonidentity attention uses the decayed pre-step weight plus its transformed proposal. Native moments are not replaced with differently rounded manual moments. Full-size gates check original Adam/FFN behavior, additional active scopes, repetition, and interruption/resume before scientific training. Merely moving attention into a different optimizer implementation would not be a clean scope test.

Table 12: All completed attention-scope endpoints, final DEV loss. The data realizations are kept separate.

Study	Parent	AdamW	FFN only	Attention only	Combined
ASF1	159	3.661593	3.659418	3.661121	3.658969
ASF1	160	3.664348	3.662232	3.663799	3.661699
RGS1	167	3.662995	3.660907	3.662482	3.660453
RGS1	168	3.661961	3.659816	3.661552	3.659360

For every parent, combined $<$ FFN-only $<$ attention-only $<$ AdamW. Relative enlargement of the small FFN-only improvement is 22.91% on B and 21.48% on A. This ratio describes differences to the baseline, not total loss, accuracy, or speed. The respective loss-scale interactions are $+0.000018947758$ and $+0.000006476417$. A small descriptive interaction is not an equivalence test; there is no matched-full-model-strength control here.

Table 13: **Scope extension: completed development results.** Mean paired final DEV differences, nats/token; two parents per realization. No cross-realization pooled test or confirmatory interval is used.

Contrast	Realization B	Realization A
FFN-only BDT − AdamW	-0.002145	-0.002117
Attention-only BDT − AdamW	-0.000511	-0.000461
Combined BDT − AdamW	-0.002637	-0.002571
Combined BDT − FFN-only BDT	-0.000492	-0.000455
Scope interaction	$+0.000019$	$+0.000006$

H.1 Localizing the two attention contributions

ACL1 reuses the RGS1 parent families 167–168. FFN BDT is active in every arm; QKV and output tempering are two binary factors. The current combined policy and FFN-only anchors are rerun within the new cohort, not supplied as historical endpoints. All eight final losses are shown below.

Table 14: Attention component localization on two reused A-development parents (ACL1). There is no AdamW-only cell in this factorial.

Parent	FFN only	FFN + QKV	FFN + output	Combined
167	3.660907	3.660680	3.660651	3.660453
168	3.659816	3.659605	3.659524	3.659360

Both attention groups contribute: mean QKV-minus-FFN is -0.000218803063 and output-minus-FFN is -0.000274162740. Combined-minus-FFN is -0.000454577617, exactly reproducing the RGS1 anchors. Adding output to FFN+QKV improves by 0.000235774554; adding QKV to FFN+output improves by 0.000180414878. The conditional interaction is $+0.000038388185$, a small subadditive effect. Neither reduced-scope variant beats combined on either parent, so the predefined better-narrower-policy candidate flag is false. This closes the bounded search. It is not a reversal of the attention extension and does not supply a new larger gain.

Audit and inference. The preserved scope audit checks sixteen original endpoints and four prefix records. The component audit checks eight endpoints and their new/original prefix records. Both verify source/protocol identities, fixed budgets, final DEV reductions, complete paired batch/RNG histories, first-active gradients/proposals, checkpoint metadata, and logged attention activity. Treated combined arms have 5,600 nonidentity attention projection-steps; single-group component arms have 2,800. The final numerical checks reproduce the published differences. Neither audit rereads all checkpoint tensors or the full token cache. These are development studies; reuse of parents and repetition of anchors does not create fresh independent evidence for a selected policy.

I COMBINED-POLICY CONFIRMATION AFTER SEPARATE LEARNING-RATE SELECTION

BCR1 primary comparison. Select an LR separately for AdamW and the fixed combined policy using the same three-LR DEV grid. On fresh parents 169–174, compare their final update-3,000 REPORT losses as combined BDT minus AdamW; negative favors combined BDT. The mean of six paired differences and one two-sided 95% Student-t interval are primary. Scope/component ablations belong to development, not this confirmation.

BCR1 uses ordinary Dense growth on OWN-S1-R1. It freezes the previously selected combined FFN+QKV+output cap-three policy rather than choosing a later scope winner. Each method receives the same three-LR development grid on parents 167–168. The selector minimizes mean final DEV independently for each method, with the lower LR resolving an exact tie.

Table 15: **BCR1 development selection.** Mean final DEV over parents 167–168. Both methods independently select the interior LR 0.0006.

Peak LR	AdamW	Combined BDT
0.0003	3.676574122	3.674154664
0.0006	3.662478170	3.659906826
0.0012	3.665787191	3.664929842

Six new parents 169–174 are pretrained for 7,630 updates at LR 0.0008 and stacked to twelve layers, copying weights and moments while advancing global data indexing. Each selected recipe then runs the fixed 3,000-update continuation, including a 200-update inactive prefix. The design permits different prefixes if selected LRs differ. Both selected 0.0006 here, and the recorded initial states and full 2,800-update histories agree within each parent. FFN treatment preserves the inherited RoleAdamW calculation; attention proposals are exposed from the updated native Adam moments. Untreated or inactive attention retains the native output, avoiding a scope-dependent replacement of moment arithmetic.

Table 16: **All BCR1 fresh-parent REPORT outcomes.** One final endpoint per method and parent. The six paired differences determine the primary interval.

Parent	AdamW	Combined BDT	Combined minus AdamW
169	3.622774	3.620134	-0.002639
170	3.621137	3.618391	-0.002746
171	3.625380	3.622639	-0.002741
172	3.622921	3.620224	-0.002697
173	3.620337	3.617695	-0.002642
174	3.618094	3.615733	-0.002362

The twelve training endpoints were locked before REPORT evaluation. The fixed REPORT slice has offset 2,500,000 and scores 1,048,576 targets. Selection was sealed at 00:14:19 CDT on September 12, the complete training cohort at 11:22:26, and the final analysis at 11:25:50. The mean paired difference is -0.002637862849, sample SD 0.000142931637, and the single prespecified two-sided 95% Student-t interval is $[-0.002787860402, -0.002487865295]$. All six signs favor combined BDT. These intervals quantify parent variation conditional on the fixed recipes and reporting text. They do not include another hyperparameter search or a new data distribution.

The final record audit recovered twelve training results, twelve final-checkpoint metadata files, twelve REPORT records, protocol, selection and cohort locks, and analysis. The recorded result hashes, contract hashes, checkpoint/report bindings, source/cache/recipe identities, within-parent states and histories, and the ordering of REPORT after locking pass all 102 checked assertions. No fresh parent-pretraining replay, checkpoint-tensor reload, or full source/job-generation audit is claimed.

Reading the source records. The frozen source also retains inherited descriptive hypothesis, secondary, and maximum-trajectory fields. The boxed definition above states the competitiveness primary used by the selection lock, actual twelve-arm execution, and final analysis; inherited fields do not define additional executed arms or primary tests. Frozen source records are preserved in the experiment archive.

BCR1 establishes the selected combined recipe against the selected AdamW recipe under this bounded search and horizon. It contains neither an FFN-only nor a matched-scrambled branch, so it does not separately confirm the attention increment or the correspondence mechanism. It also does not establish superiority over tuned GS or measured end-to-end acceleration.

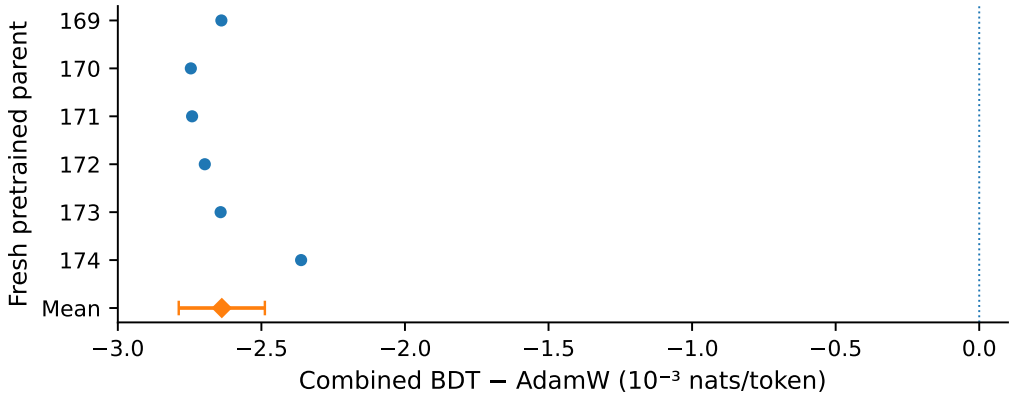

Figure 8: **The combined policy improves quality after separate LR selection.** Each point is one fresh parent's final REPORT difference, combined BDT minus AdamW; negative favors BDT. The bottom diamond marks the six-parent mean, -0.002638, with its single paired 95% interval $[-0.002788, -0.002488]$. The horizontal axis includes the zero-effect reference.

J Initialization and directional evidence

J.1 Absolute ordinary-Dense factorial

This factorial compares independently initialized FFNs with FFNs initialized by copying. Within each initialization condition, AdamW and BDT begin from the same saved prefix.

Table 17: Original absolute Dense endpoints, six seeds, cap two and LR 0.0004.

Seed	Independent Adam	Independent BDT	Cloned Adam	Cloned BDT
113	4.120471	4.119970	4.179681	4.170874
114	4.122206	4.122256	4.174794	4.165005
115	4.142924	4.142301	4.191666	4.183263
116	4.130482	4.129731	4.188116	4.179163
117	4.110355	4.110049	4.170490	4.160850
118	4.139297	4.139386	4.190667	4.181367

All 24 original endpoints use the same model geometry, evaluation convention, logical-batch definition, training horizon, and token realization. The two initialization conditions have different parent states; within each initialization condition, the two policies share a saved prefix. Source differences

implement copying FFN gate/up/down values after initialization, without tying storage or copying attention.

Results and interpretation. The directly registered initialization interaction is -0.008808561 $[-0.009668117, -0.007949004]$. The retrospectively recovered absolute comparisons are descriptive: cloned Adam minus independent Adam is $+0.054946634$ $[+0.050042771, +0.059850497]$; cloned BDT minus independent Adam is $+0.045797708$ $[+0.041005420, +0.050589996]$. Their intervals are unadjusted. The ratio of the mean repair to the mean cloning penalty is 16.65%, not a causal-mediation fraction or general percentage of cloning harm explained.

J.2 Reindexing and norm/direction controls

The reindexing study uses S1-B, seeds 119–120, ordinary cloned Dense, cap two, 3,000 updates, and a common trained parent per seed. Both Adam moments and all three FFN projections are transported consistently for the reindexed-model arms. Known-coordinate transport is an implementation positive control: it reconstructs the intended chart's transformation rather than learning an alignment.

Table 18: Reindexing endpoints. P denotes the function-preserving reindexed chart; transported BDT uses the known inverse permutation.

Seed	Adam	BDT	P-Adam	P-BDT	Transported
119	4.189630	4.180767	4.189590	4.189202	4.181304
120	4.171616	4.161734	4.171269	4.171505	4.161789

The norm/direction grafts use seeds 121–122. If U_r is the raw signed pre-decay proposal and V_r its hypothetical BDT version, the norm graft is $U_r\|V_r\|/\|U_r\|$ and the direction graft is $V_r\|U_r\|/\|V_r\|$. These do not generally preserve the canonical common-depth component.

Table 19: Per-layer component-control endpoints.

Seed	Adam	Full BDT	Norm-only	Direction-only
121	4.194160	4.185669	4.193979	4.186382
122	4.181429	4.172283	4.181119	4.172562

On OWN-S1-R1, seeds 135–136 test matched-angle random and reflected directions. Reflection reverses the BDT component orthogonal to raw Adam while retaining the parallel component. The random construction uses a fixed skew-symmetric depth map to obtain an alternative orthogonal direction, with no validation feedback. Both constructions match the hypothetical canonical candidate at their own branch states.

Table 20: Matched-direction endpoints on the distinct OWN-S1-R1 realization.

Seed	Adam	BDT	Matched random	Reflected
135	4.168868	4.160054	4.168958	4.188397
136	4.172715	4.163636	4.173698	4.192985

All eight original endpoint files and their recorded source hashes were recovered. Two-seed controls are descriptive. The projections and trajectory checkpoints do not enlarge the independent sample count. They motivate specificity of direction, but do not establish that CMS1's control has matched every possible causal attribute of an optimizer update.

K COMPARATIVE INITIALIZATION FACTORIAL

The comparative initialization experiment (ODC) measures how BDT and a modified Gradient Smoothing rule respond to independent versus cloned FFN initialization.

The modified Gradient Smoothing (GS) diagnostic in this appendix acts on FFN proposals only. It is delayed to phase 200 and uses the same activation ramp as BDT. Let U_r and V_r be the raw and smoothed FFN proposals at depth r, and let α_t be the smoothing coefficient derived from ramp a_t:

$$V_r = (1 - \alpha_t)U_r + \frac{\alpha_t}{2}\left(U_{\max(r-1,0)} + U_{\min(r+1,R-1)}\right), \qquad \alpha_t = 0.1a_t. \tag{12}$$

ODC uses ordinary Dense12 with 123,551,232 parameters, context 512, FFN width 2,048, peak LR 0.0004, and 3,000 updates. AdamW and strict-execution settings follow Appendix A, with the LR and policy onset specified here. Realization A is OWN-S1-R1; realization B is S1-B. For each seed the full ordinary-Dense model is initialized once for the independent condition. The cloned condition copies the first FFN's gate/up/down values to other depths with independent storage; non-FFN initialization and post-initialization RNG are matched across the two conditions. Each condition then has its own shared update-200 parent.

The three policies apply AdamW, BDT cap two, and Equation (12). They collect the same kinds of shadow statistics at their own states, applying only the assigned proposal. GS is delayed and ramped with BDT, rather than applied from update one. BDT and GS preserve neither identical realized angles nor identical layerwise magnitude allocations. The fixed LR was inherited, not selected independently for this factorial.

Table 21: All 36 ODC final DEV endpoints. A: four seeds on OWN-S1-R1. B: two replication seeds on S1-B. The accounts/caches are analyzed separately.

Cohort	Seed	Initialization	AdamW	BDT	Modified GS
A	153	independent	4.141945	4.141989	4.149876
A	153	cloned	4.196799	4.188030	4.206669
A	154	independent	4.142536	4.142542	4.149841
A	154	cloned	4.197722	4.188653	4.207753
A	155	independent	4.145564	4.145309	4.152966
A	155	cloned	4.206077	4.197479	4.216867
A	156	independent	4.143478	4.143398	4.151027
A	156	cloned	4.206134	4.196908	4.215844
B	157	independent	4.140115	4.139397	4.147436
B	157	cloned	4.191345	4.181956	4.201603
B	158	independent	4.142191	4.141301	4.149207
B	158	cloned	4.200443	4.191806	4.210245

Table 22: ODC treatment effects averaged within each separately analyzed cohort.

Cohort	Initialization	BDT − Adam	Modified GS − Adam
A (4 seeds)	Independent	−0.000071	+0.007547
A (4 seeds)	Cloned	−0.008916	+0.010100
B (2 seeds)	Independent	−0.000804	+0.007169
B (2 seeds)	Cloned	−0.009013	+0.010031

Initialization interactions. For each seed and method, the initialization interaction is its treatment-minus-AdamW loss difference under cloned initialization minus that difference under independent initialization. The four seed-level interactions on realization A are

$$\text{BDT: } (-0.008813698, -0.009075148, -0.008342717, -0.009146277),$$
$$\text{Modified GS: } (+0.001938436, +0.002726529, +0.003387632, +0.002160614).$$

Their means and ordinary paired 95% intervals are -0.008844460 $[-0.009423312, -0.008265608]$ and $+0.002553303$ $[+0.001522737, +0.003583869]$. Subtracting the modified-GS interaction from the BDT interaction within each seed gives $(-0.010752134, -0.011801677, -0.011730349,$

-0.011306891), averaging -0.011397763 [$-0.012165767, -0.010629758$]. These are descriptive development intervals, not additional confirmatory CMS1 contrasts.

On realization B, the BDT and GS mean interactions are -0.008208783 and $+0.002862091$, giving a between-method interaction difference $J = -0.011070875$ [$-0.017905604, -0.004236145$]. The small replication cohort is not combined with A to obtain a narrower interval. Absolute A cloning penalties under Adam average 0.058302274; cloned BDT retains 0.049386495 of that gap to independent Adam. The corresponding B gaps are 0.054740571 and 0.045727982.

The complete record audit checks original result hashes, contract/configuration/cache identities, checkpoint metadata, all within-parent batch/RNG histories and first-active-step gradients/raw proposals. One A lane migrated to another H200 after interruption. It passed a new target-specific engineering gate before its two remaining continuations. Resume visits to completed jobs produced empty child logs, but completed result hashes and the final 24-cell account analysis establish completion. No scientific endpoint was replaced because of the migration. Full checkpoint payloads were not reloaded in the final readout audit.

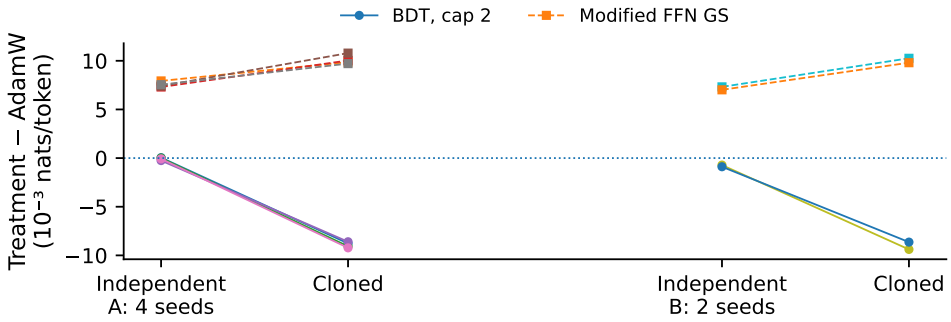

Figure 9: **The two tested rules respond differently to initialization (ODC).** Each segment joins the independent- and cloned-FFN treatment effects for one seed. Both separately analyzed realizations reproduce the larger BDT benefit and larger modified-GS penalty under cloning. Scope and onset/ramp match; realized strength does not. Absolute cells and paired interaction estimates appear in Appendix K.

L Equal-search comparison and its execution distinction

This comparison gives AdamW, BDT, and standalone Gradient Smoothing equal learning-rate search budgets. Its execution protocol precedes the strict shared-state comparison, so the results retain their original statistical interpretation.

Each of AdamW, cap-three BDT, and Standard-Proj GS receives ten full-budget learning rates on two development seeds, 127–128, totaling sixty trajectories. The search is a bounded sequence of bracket extensions, with the common eventual set

$$\{0.0001, 0.0002, 0.0003, 0.0004, 0.0006, 0.0008, 0.0010, 0.0012, 0.0016, 0.0024\}.$$

The original search records define the grid. Each method selects the same 0.0012 recipe before evaluating six other seeds. No best-checkpoint selection is used. BDT–Adam on the selected DEV recipe is approximately $+0.000278$. The favorable final reporting result is not merely a favorable development mean reproduced on the same seeds.

GS applies the local window with interior weights $(0.05, 0.90, 0.05)$ and boundary weights $(0.95, 0.05)$ to the specified FFN and attention projections from update one. The completed ODC FFN-only matched-onset diagnostic in Appendix K is a different specified policy. Its results are not evidence of superiority over this Standard-Proj comparator.

Table 23: Original six-seed equal-search REPORT endpoints. Independent executions under the preceding runtime, not CMS1 shared-parent strict-execution replications.

Seed	AdamW	BDT	GS	BDT − Adam	BDT − GS
129	3.894772	3.890467	3.891650	−0.004305	−0.001183
130	3.903886	3.897904	3.901945	−0.005982	−0.004041
131	3.902869	3.900963	3.898163	−0.001906	+0.002799
132	3.904191	3.903162	3.901478	−0.001029	+0.001684
133	3.894396	3.891056	3.894522	−0.003339	−0.003466
134	3.903549	3.901253	3.898155	−0.002296	+0.003098

The paired BDT–Adam difference is -0.0031429874 $[-0.0050296579, -0.0012563169]$. The BDT–GS difference is -0.0001848070 $[-0.0034970905, +0.0031274765]$, with three favorable signs. Both are ordinary unadjusted intervals under that original protocol, not multiplicity-adjusted CMS1 intervals.

Execution distinction. Matching initial model/RNG fingerprints did not make the preceding Adam and BDT-observer prefixes identical. Their update-200 DEV differences have mixed signs (Table 24). A later six-prefix identity replay also found differences among Adam-only repetitions despite matching batches and RNG, with the first differences present in recorded gradients before the first optimizer step. The strict-execution gate addresses repeatability prospectively for CMS1. The earlier differences are not a measured null distribution for the final 3,000-update REPORT endpoint and do not justify subtracting a noise correction from it.

Table 24: Pre-activation DEV difference, cap-three execution minus Adam execution, in the preceding fair cohort. These are not BDT treatment effects.

129	130	131	132	133	134
+0.007474	−0.030046	+0.020991	−0.030586	+0.025687	−0.006723

M ACHIEVED GEOMETRY AND EXPLORATORY DIRECTION DIAGNOSTICS

The first set of diagnostics checks the quantities explicitly matched by the control. The second describes layerwise and temporal properties that remain free to vary.

The September 12 extraction recovers all six CMS1 and six GCT2 matched-control accumulators. Each study records 50,400 active projection transformations and no numerical-zero event. Table 25 gives their all-recorded-step maxima. Those maxima are not quantiles estimated from the sparse curves. GCT2's maximum residual amplification is 128.6800, larger than CMS1's 34.3817. Amplification is the scaling needed to match the canonical angular departure from a smaller scrambled orthogonal residual. It is not a multiplier of the total applied step norm. Layerwise allocation, the raw alternative spectrum, and temporal direction consistency are not held fixed by the geometric match.

Table 25: **Achieved matching diagnostics.** Maxima over recorded applied-control accumulators, six trajectories and 50,400 projection transformations per study. GCT2 is the growth confirmation in Section 3.1. These are proposal-space errors, not loss errors.

Recorded maximum	CMS1	GCT2
Relative centered-norm error	5.11×10^{-11}	4.28×10^{-11}
Normalized common-component error	7.36×10^{-9}	7.35×10^{-9}
Angle error (radians)	5.52×10^{-11}	4.30×10^{-11}
Normalized displacement error	2.03×10^{-8}	1.90×10^{-8}
Residual amplification	34.38	128.68

Exploratory applied geometry. A second extraction summarizes actual applied chart diagnostics for the completed BAA1 parent 160, rather than all hypothetical candidates in the shadow bank. Each

row below uses 312 stored projection records for amplification and 285 valid consecutive-direction comparisons for temporal cosine. The cosine refers to the rotational component orthogonal to centered raw Adam. Values across steps and projections are repeated measurements within one parent, not additional experimental units.

Table 26: **Exploratory BAA1 applied-direction diagnostics, parent 160.** Medians and 10th–90th percentiles of sparse stored records. These are descriptive distributions, not confidence intervals.

Draw	Assignment	Amplification median	Amplification 10–90%	Temporal median
0	Preserved	1.0337	[1.0290, 1.0406]	0.89770
0	Reassigned	29.7585	[4.1550, 81.5642]	0.89687
1	Preserved	1.0338	[1.0289, 1.0407]	0.89761
1	Reassigned	30.0666	[4.1463, 83.3457]	0.89685

The temporal-cosine medians are close, while amplification distributions differ strongly. Recorded centered-layer norm ratios also have wider central ranges under reassignment: approximately [0.8375, 1.1622] and [0.8341, 1.1634], versus [0.9856, 1.0129] and [0.9857, 1.0130] for preserving assignments. These attributes can participate in the effect of changing correspondence. The distributions do not isolate mediation or prove that all temporal behavior is equivalent.

The prespecified structural score is the mean raw-proposal cosine within true copied pairs minus that over unrelated pairs, summarized across FFN groups at each recorded step. All four applied chart branches begin at the same 0.380159 score at update 201. At update 3,000 it is approximately 0.30540 in the two preserving branches and 0.32793 in the reassigned branches. Thus the better endpoint is not identified simply by a larger contemporaneous value of this score across divergent trajectories. Training-age trends and policy effects must be distinguished. These observations are exploratory rather than an independently tested schedule rule.

Unavailable older temporal and layer-allocation series remain missing; hashes do not reconstruct absent tensors. These are recorded diagnostics, not a tensor-level reproduction.

N Full-cost scratch references and implementation timing

The cost evidence has two parts. Logical matrix-multiplication accounting defines the scratch-training references. Measured implementation timings then distinguish optimizer-only cost from complete training-step cost.

N.1 Core-matmul accounting and token budgets

For sequence length L, model width d, FFN width m, vocabulary V, and depth D, let F_D denote the declared forward principal-matmul approximation per sequence:

$$F_D = D(4Ld^2 + 3Ldm + 2L^2d) + LdV. \tag{13}$$

The three SwiGLU matrices, attention input/output projections and full-square attention products are included, as is the tied vocabulary output projection. Let C_D denote the corresponding training cost per update. The forward-plus-two-backward approximation gives $C_D = 3bF_D$ for effective batch $b = 64$. Norms, activations/softmax, optimizer transforms, validation and I/O are excluded. This is an explicit logical model of the non-checkpointed training path, not a device FLOP counter. It excludes the additional forward recomputation required by the activation-checkpointed stack in Appendix S.

At $L = 512, d = 768, m = 2048, V = 50257$, the recorded six-layer parent and twelve-layer target costs are $8{,}432{,}841{,}129{,}984$ and $13{,}071{,}405{,}809{,}664$ MACs/update. The full growth path is $1.0355679525076992 \times 10^{17}$ core MACs; 7,922 target-model updates use $1.03551676824158208 \times 10^{17}$. Flooring to an integer gives the scratch comparator a marginally smaller budget (less than 0.005% difference), not extra compute. The original accounting also retains a triangular-attention sensitivity in the evidence. The full-square convention was fixed for the primary reference.

GCT1's parent costs 250,019,840 token exposures and the post-growth phase 98,304,000, totaling 348,323,840 per complete path. Core-budget scratch uses 259,588,096 exposures; equal-token scratch

uses 348,323,840. Parent cost is charged once in full for every recommended growth path, never divided over the experimental branches that happen to share it. The total scratch campaign has four separately trained trajectories: two seeds at each horizon, or 37,104 updates.

N.2 Scratch execution and completed results

GSR1's independently initialized scratch AdamW runs use peak LR 0.0008, warmup 200 and cosine decay to 0.1 of peak, with all other Dense settings matched. Each horizon has its own schedule from the start. Seed 159 runs core then token; seed 160 runs token then core. Both begin with Adam age zero. The final four source records have the registered horizons/ages, finite final DEV values, and matching terminal curve/checkpoint metadata. The collector also checks cache/runtime identity and the recorded training-batch sequence against the full growth path. Every result is retained; there is no interval computed from four unpaired runs or selection of the better budget.

The two grown-Adam minus core-scratch differences are -0.003198349848 and -0.002390487120; grown-BDT minus core-scratch gives -0.005372751504 and -0.004506407306. Mean grown-BDT minus token-scratch is $+0.070696207695$, with values $+0.069795401767$ and $+0.071597013623$. These are fixed-recipe, two-seed DEV comparisons. Both growth and scratch hyperparameters retain their documented selection histories. The scratch LR has not had equal new tuning opportunity. A favorable core-budget reference does not certify the best full-cost recommendation.

Table 27: **Completed full-cost development references (GSR1).** Final DEV losses. Both growth paths include the small parent's full training cost. The fixed-LR references are not equally tuned recommendations. Core-matmul matching is not measured wall-clock matching.

Complete path	Seed 159	Seed 160	Mean
Scratch AdamW, core budget	3.664791	3.666739	3.665765
Growth + AdamW	3.661593	3.664348	3.662970
Growth + BDT	3.659418	3.662232	3.660825
Scratch AdamW, token budget	3.589623	3.590635	3.590129

N.3 Full-step and optimizer-only timing (OCB1)

All 36 registered blocks finish on one H200 NVL: twelve main full-step, twelve main optimizer-only, six audit full-step and six audit optimizer-only. The runtime records PyTorch 2.11.0+cu128, CUDA 12.8, cuDNN 91900, NumPy 2.4.6, four CPU threads and the strict settings used in the new studies. A separate untimed profile identifies FlashAttention operations. Reference/candidate quality is not selected with this engineering state. It is fixed at update 500. Warmup and restoration precede every timed window, so each method starts from the same recorded state, input ring and gradient history for its depth.

The main conditions are native fused AdamW, split-FFN AdamW, canonical BDT, split-all-projection AdamW, all-projection GS, and Dense6 split-FFN AdamW. The GS formula is extracted from the archived implementation with alpha 0.1. The all-projection Adam condition isolates the cost of exposing attention proposals from the smoothing operation itself. Twenty-step full windows include input transfer; hundred-call optimizer windows use fixed clipped gradients. The profiler and audit hashes are excluded from both primary timer families. Mean and median timings are not mixed when forming the reported paired ratios.

Table 28: Complete OCB1 timing summary. Absolute entries are median milliseconds per measured call; full-step and optimizer-only windows are distinct measurements.

Implementation	Full step (ms)	Optimizer only (ms)
Native fused AdamW, Dense12	478.330	2.273
FFN-split AdamW, Dense12	483.839	6.246
Canonical BDT, Dense12	526.479	47.175
All-projection split AdamW, Dense12	484.399	8.650
All-projection GS, Dense12	495.000	20.591
FFN-split AdamW, Dense6	341.793	3.468

An identical split-Adam sentinel brackets every block. All complete blocks enter the primary summary. The predefined stable subset is secondary. All pass the 10% drift criterion, with maximum drifts 1.43% (main full step), 9.34% (main optimizer), 0.68% (audit full step), and 7.02% (audit optimizer). Stable contention could still affect the absolute timings. The timing repetitions are not training seeds, so their IQRs are not seed-level confidence intervals.

Memory and instrumentation. Full-step peak allocated memory is 10.714956 GiB for split Adam and 10.714963 GiB for canonical BDT; isolated optimizer peaks are 2.207445 and 2.752380 GiB. Equal full-step peaks therefore do not imply zero extra optimizer memory. Audit-Adam, audit-BDT and audit-scrambled full-step median ratios to their bracketing split-Adam sentinels are 1.409002, 1.414743 and 1.419120. Their optimizer-only ratios are 31.0171, 30.8649 and 31.1665. These large instrumentation ratios are kept separate from the deployable canonical transformation.

Timing verification. The final audit reads all 36 original blocks (252 measurements including sentinels), checks counterbalanced orders and windows, refresh cadence, state/input/runtime/GPU/source identity and repeat final model/Adam hashes, and independently reproduces the published statistics. Five timing/collector sources are freshly hash-checked; full checkpoint payloads and token arrays are not reloaded. No fastest-block or fastest-GPU selection is performed. At an illustrative constant 1.0893 step-cost ratio (not a ratio measured uniformly across configurations), a same-model method needs about 8.2% fewer updates merely to break even on step time. That algebra is not a measured result for the multistage growth path.

N.4 Bounded diagnostics-only cost reduction

BCO1 compares split AdamW, lean canonical FFN BDT, and one diagnostics-lite candidate using twelve counterbalanced full-step blocks and twelve optimizer-only blocks from the fixed engineering state. The candidate omits two optional reporting reductions while retaining the update arithmetic and safety fields. It is distinct from removing the expensive scrambled shadow calculations, which OCB1 had already excluded from canonical cost.

All twelve full-step blocks pass the sentinel criterion; eleven optimizer-only blocks do. The full-step candidate/reference median is 0.987825 (IQR 0.985053–0.990577), with the candidate faster in ten of twelve blocks. This is a 1.22% reduction in step time, below the prespecified 2% advancement screen. Contemporaneous reference/split and candidate/split medians are 1.092967 and 1.079899. The optimizer-only candidate/reference median is 0.854251. Its larger relative reduction is not a full-training speedup. The failed advancement and half-overhead targets remain recorded. This reference-preserving engineering candidate is not substituted into GCT2, and neither its timings nor OCB1's original FFN timings are measurements of combined attention BDT.

N.5 Combined-policy replay and shared-resource timing

CPC1 completes both reused development parents 167–168, eight cost blocks per execution lane, and four lean update-200-to-3,000 replays. Each replay reproduces its original BCR1 development endpoint exactly, including the recorded model and named-Adam signatures; combined BDT also reproduces its covariance signatures. The replay differences are -0.002542080358 and -0.002600608394 nats/token relative to AdamW. These are implementation reproductions of development trajectories, not additional fresh-parent confirmation.

Timing depends strongly on the shared-resource window. Active continuation combined/AdamW elapsed ratios are 1.249451 for parent 167 and 0.564519 for parent 168; reversing the method order does not remove the external-load difference. The later replay is slower in both lanes. Active elapsed includes periodic evaluation and checkpoint work but excludes parent acquisition, the shared prefix, admission waiting, and engineering gates. The recorded brackets for the prespecified DEV target 3.70 are updates 2,250–2,500 for all four replays, with opposite wall-clock orderings across lanes; that evaluation cadence does not resolve exact crossing times.

In the fixed-state full-step cost windows, the primary combined/partitioned-AdamW sentinel ratios are 0.946118 on the less stable lane and 1.058430 on the more stable lane. The latter passes the sentinel-drift screen in all eight blocks, but does not establish a universal 5.84% overhead. A later tightly interleaved CPC2 arrangement gives 1.620651 on its completed default-lane readout,

with an identical-method control of 0.999741 and all five necessary stability checks passed. That materially different measurement is retained rather than replaced by the more favorable window. It is a configuration-specific observation from the retained default-lane readout. No intrinsic overhead or portable time-to-quality gain is inferred from these shared-resource timings.

O SCALE, HORIZON, AND EARLIER APPLICATION EVIDENCE

Fixed-recipe confirmations. Cap-two BDT on six S1 seeds 91–96 gives mean BDT–Adam -0.008393 $[-0.009081, -0.007705]$. A bounded two-seed cap/ramp development study selects cap three; six new seeds 105–110 give cap-three–Adam -0.009915 $[-0.010756, -0.009074]$ and cap-three–cap-two -0.001626 $[-0.002081, -0.001171]$. These are inherited-recipe studies with update-200 shared prefixes, not extra replications of CMS1's selected-LR comparison.

Actual larger-parameter BDT transfer. Four 501,904,152-parameter S1 trajectories, seeds 123–126, apply the actual cap-three policy at the inherited recipe. Their paired differences are retained below. The mean is -0.012030567 $[-0.012563004, -0.011498130]$. All four favor BDT. Each path still uses only 98.304M exposures. This is distinct from an earlier 502M static spectral experiment and from a mature 12B-token scaffold experiment. Neither of those is substituted for mature BDT evidence.

Table 29: Per-seed actual 501.9M BDT transfer, inherited recipe.

Seed	BDT − Adam
123	-0.012126
124	-0.012451
125	-0.011866
126	-0.011679

Longer-budget development. A twelve-trajectory, two-seed development grid on OWN-S1-R1 compares cap-three BDT and Adam for 15,000 updates (491.52M exposures), using LRs $\{0.0004, 0.0008, 0.0016\}$. Both completed selectors choose the interior 0.0008. At that LR, Adam/BDT losses are 3.499444/3.497594 on seed 141 and 3.498036/3.497245 on seed 142, with mean BDT–Adam -0.001320773. This is selection and evaluation on the same two seeds, not fresh confirmation. Its predefined 0.005 practical advancement magnitude is not reached. The schedules begin at the longer horizon from the start. They are not extensions of cooled 3,000-update schedules.

Earlier copied-depth development screen. The growth screen trains two independent six-layer Dense parents (81,074,688 parameters) for 7,630 updates at peak LR 0.0008, using 250,019,840 exposures each. All six blocks are then copied into the twelve-layer order $0, 1, 2, 3, 4, 5, 0, 1, 2, 3, 4, 5$. Weights, first/second Adam moments, and age 7,630 are mapped into distinct storage. Stacking is not function-preserving: the parent-to-grown loss jump is evaluated and recorded. The common grown state initializes each continuation. A new LR phase and a new depth-covariance history begin. Data indexing advances rather than restarting from the parent's first logical batches.

The three-LR grid $\{0.0002, 0.0006, 0.0012\}$ selects 0.0006 for both methods after 3,000 further updates. The selected endpoints are 3.662921751/3.660730850 (Adam/BDT, seed 139) and 3.660211295/3.657929435 (seed 140), averaging -0.002236380. The original 0.005 advancement screen fails. This result is a modest within-growth-pipeline effect. No end-to-end advantage over compute-matched scratch training is established. GCT1 and GSR1 are separate, later fixed-recipe studies on S1-B. Their results do not change this recorded decision on OWN-S1-R1.

Other completed development observations. A two-seed GS scope/normalization grid selects LR 0.0012 for all tested methods. Attention-only GS is favorable versus Adam by -0.003879 on average and both signs; FFN-only GS is unfavorable by $+0.008754$. It has no BDT arm and no independent confirmation. A separate two-seed late-strengthening screen finds fixed cap three better than the selected late $3 \to 4$ schedule on both seeds, by 0.000841 on average. These observations limit simple claims that coupling more tensors or increasing tempering late must help. Neither is a new central performance result.

P CONTEXT: DEPTH ROLES, ACCESS, AND HORIZON

These earlier studies concern functional depth roles, parameterization access, and horizon-dependent update geometry. They provide context for the later correspondence experiments and retain their original models and evaluation budgets.

P.1 FUNCTIONAL-ROLE USE AND EXACT REPRESENTATION

Two archived 124M checkpoints trained for 2.5B tokens give 264/264 harmful nonidentity private-FFN substitutions, including all 44 adjacent substitutions. Retaining ten of eleven joint centered-role directions preserves 94.23%/94.18% of weight variance but increases loss by 0.4049/0.3788; full reconstruction recovers loss within 1.2×10^{-5}. This establishes checkpoint sensitivity, not a minimum useful role dimension or a mediator of CMS1's gain. The archived-manuscript source status is preserved in the evidence map.

Let W stack the flattened role parameters, w_0 be a base parameter vector, A contain role coefficients, and Z contain basis parameter vectors. For $W = \mathbf{1}w_0^\top + AZ$, with $A \in \mathbb{R}^{R\times K}$ and $Z \in \mathbb{R}^{K\times P}$, arbitrary role matrices are representable at $K = R - 1$: choose $w_0 = W_0$, $Z_k = W_{k+1} - W_0$, $A_0 = 0$, and $A_r = e_{r-1}$. Here e_{r-1} is a standard coordinate vector, subscripts on W, A, Z select rows, and A and Z are factors in a full-tensor role parameterization, distinct from the pair-average component used in the main text. This is not low matrix-rank adaptation. The centered differential is

$$H\,dW = H\,dA\,Z + H\,A\,dZ.$$

The Z-update image has dimension $P\,\mathrm{rank}(HA)$. At $HA = 0$, the joint centered tangent has dimension $(R-1)\,\mathrm{rank}(Z)$: the tensor product of the centered role space and the row space of Z. At $A = 0$, $\nabla_Z \mathcal{L} = A^\top \nabla_W \mathcal{L} = 0$ for loss depending on the factors through W. These are local, not persistent rank restrictions. The function-matched Restricted/Open/Explicit-Cloned bridge gives three 502M mean losses 3.9700/3.9034/3.9005. A separate probe reaches full accessibility by update five while loss differences persist.

P.2 STATIC SPECTRAL PREFERENCE DEPENDS ON HORIZON

The static experiment intervenes only during updates 1–200, then releases all arms to Adam. Its frame comes from eight task-gradient calibration batches or a seeded random frame, with shared spectrum and an external same-seed centered-norm schedule. These differ from BDT's online raw-proposal statistics. A six-seed five-dose test uses mode-energy/gain rank correlations $\rho \in \{-1, -0.5, 0, 0.5, 1\}$. Mean loss–dose slope changes from -0.02229 at update 200 to $+0.00321$ at 1,000. The paired change is positive on all six seeds, $+0.02549$ $[+0.01584, +0.03514]$. The late slope alone is unresolved. A six-seed hard-switch schedule gives -0.00035 $[-0.00750, +0.00680]$. This is horizon sensitivity of that intervention, not a derivation of online BDT.

P.3 IMMEDIATE UTILITY DEPENDS ON CONCURRENT UPDATES

Two cloned-Dense Adam trajectories (135–136) supply states at updates 200, 400, 1,000 and 2,000. One clipped gradient constructs every counterfactual; evaluation uses 2,048 held-out tokens at context 256. Let θ be the current parameters and ℓ the held-out loss. With alternative role increments u, v (including role decay) and identical non-role increment r, define the finite interaction $\mathcal{I}$ by

$$\mathcal{I} = [\ell(\theta + v + r) - \ell(\theta + u + r)] - [\ell(\theta + v) - \ell(\theta + u)].$$

For an exactly quadratic objective with Hessian $\mathcal{H}$, it equals $r^\top \mathcal{H}(v - u)$. All six post-ramp states have a less favorable held-out first-order role term and lower role-only curvature under BDT; role-only effects are mixed, but complete steps favor BDT at all six. Reflection wins locally at seed 136/state 400 but loses after training; local preference need not predict endpoints.

Table 30: Local-audit BDT minus Adam, in units of 10^{-6}. Curvature is the radius-one second difference. Eight states are repeated measurements from two seeds.

Seed	State	Linear	Curvature	Role only	Full step
135	200	-3.24	$+0.24$	-3.10	-2.38
135	400	$+174.16$	-394.11	-20.03	-167.37
135	1000	$+96.86$	-149.73	$+22.65$	-41.25
135	2000	$+14.55$	-24.32	$+2.38$	-9.06
136	200	-0.86	$+3.58$	$+0.95$	$+2.38$
136	400	$+186.87$	-407.70	-12.16	-154.02
136	1000	$+82.91$	-150.92	$+8.34$	-61.51
136	2000	$+9.18$	-19.79	-0.24	-9.78

Q ANGULAR-MATCHING ROBUSTNESS: COMPLETE ARM1 DIAGNOSTIC

Q.1 FIXED COMPARISON AND INTERPRETATION

This diagnostic asks whether the ancestry contrast persists when the canonical-angle constraint is released. The common component and pooled centered norm remain controlled.

ARM1 reuses the S1-B growth-development parents 159–160 and the fixed draw-zero balanced ancestry maps. The five arms are AdamW, preserved-matched, reassigned-matched, preserved-unforced, and reassigned-unforced. Each starts at its shared post-growth update-200 state and finishes at update 3,000, with peak LR 0.0006 and cap three. Weights, Adam moments and age, data indexing, random state, and the registered chart histories are retained. The unit remains a reused development parent, not an arm or a chart draw. Final DEV evaluation is used; there is no REPORT evaluation or inferential confidence interval.

The matched arms use the branch-own-state construction in Equation (7). The unforced arms transport the alternative-chart candidate back into the canonical chart, retain its centered direction, restore the canonical pooled centered norm, and restore the canonical common component, without imposing the canonical angle. Releasing that constraint permits different angular departures and subsequent strength histories. It is a robustness diagnostic for the control construction, not a second fully strength-matched comparison.

Contrasts. The fixed primary contrast is unforced reassigned-minus-preserved final DEV loss within each parent. Secondary contrasts are the matched gap and the difference between matched and unforced gaps. All ten endpoints are retained. No intermediate checkpoint or favorable subset replaces the fixed endpoint.

Table 31: All final ARM1 endpoints at update 3,000. P and R denote preserved and reassigned assignments. Values are rounded for display. The source extract retains the recorded precision.

Parent	AdamW	P matched	R matched	P unforced	R unforced
159	3.661593	3.659485	3.662580	3.659511	3.661697
160	3.664348	3.662230	3.665426	3.662305	3.664326

Table 32: ARM1 paired contrasts, in nats/token. Interaction is the matched gap minus the unforced gap. These are descriptive quantities from two reused parents.

Parent	Unforced R − P	Matched R − P	Interaction
159	$+0.002185732$	$+0.003095249$	$+0.000909517$
160	$+0.002020992$	$+0.003196053$	$+0.001175061$
Mean	$+0.002103362$	$+0.003145651$	$+0.001042289$

On both parents, the unforced preserving assignment retains an advantage over its unforced reassigned comparator and over AdamW. Preserved-unforced minus preserved-matched is $+0.000025929883$ for parent 159 and $+0.000074384734$ for parent 160. In contrast, reassigned-unforced minus

reassigned-matched is -0.000883586705 and -0.001100676134. The reassigned unforced endpoints lie $+0.000103808939$ and -0.000022511929 from AdamW. Numerical proximity is descriptive, not formal equivalence.

The matched construction enlarges the observed separation, but is not required for its sign in these two parents. Neither the interaction nor the ratio of the two gaps estimates a fraction of final-loss mediation. Endpoint losses alone do not identify the distribution of update angles, layerwise allocation, or temporal direction changes. The result complements BAC1's fresh-parent matched-assignment confirmation without increasing that confirmation's sample size.

Q.2 Sparse recorded applied geometry and learning curves

We separately extract the `arm1_applied` entries from the four completed unforced result files. The other chart diagnostics in those files describe hypothetical angle-matched alternatives and are not treated as the applied geometry. Each arm records the three FFN projections (gate, up, down) at phases 201, 202, 400, 1,000, 2,000, and 3,000: eighteen applied records per arm. The post-ramp summary below uses phases 400, 1,000, 2,000, and 3,000 only, giving twelve records per arm. These are repeated observations within two reused parents, not twelve or seventy-two independent replications.

Let α_t be the recorded centered angle of the actual unforced proposal and θ_t the hypothetical canonical angle at that same branch state. We summarize the pointwise ratio α_t/θ_t. Because both proposals retain the common component and centered norm, the corresponding normalized displacement ratio is $\sin(\alpha_t/2)/\sin(\theta_t/2)$. This is displacement from AdamW, not the ratio of total update norms. Different branches need not have the same θ_t or state.

Table 33: **Actually applied unforced geometry after the ramp.** Median and full observed range over four saved phases times three projection groups per arm. All ratios compare with the canonical candidate at the same branch state. Ranges are descriptive, not uncertainty intervals.

Parent	Assignment	Angle ratio median	Observed range	Displacement ratio median
159	Preserved	0.95615	[0.95413, 0.95950]	0.95625
159	Reassigned	0.03266	[0.01225, 0.05772]	0.03271
160	Preserved	0.95626	[0.95476, 0.95933]	0.95635
160	Reassigned	0.03049	[0.00969, 0.05800]	0.03053

Preservation's unforced angular departure is close to its own-state canonical departure in these saved records. Reassignment has a much smaller departure after the ramp. The largest recorded relative centered-norm error across the post-ramp entries is below 2.5×10^{-11}. The norm restoration therefore does not explain these angular differences in the recorded proposal geometry. The four saved DEV evaluations also retain a positive unforced reassigned-minus-preserved gap on each parent. This is consistent with a constructive preserving trajectory and a less perturbative unforced reassignment. It does not identify a causal fraction of the endpoint gap, establish between-branch strength equivalence, or reconstruct unrecorded intermediate geometry.

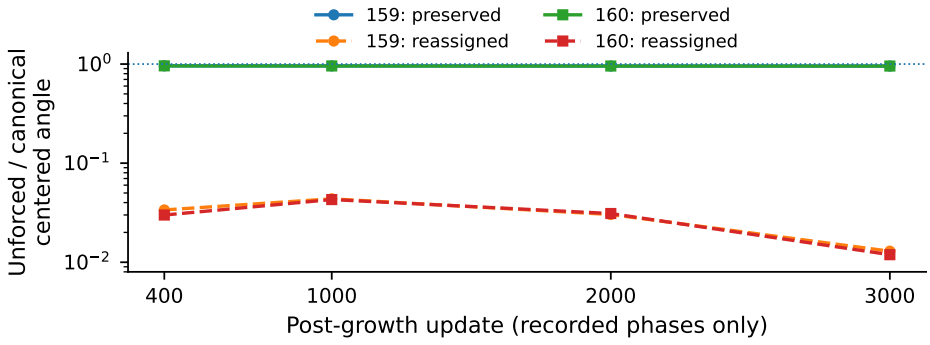

Figure 10: **Unforced angular departure relative to the own-state canonical candidate.** Each point is the median of the three recorded FFN projection ratios at that phase, shown separately for each parent and assignment. The vertical axis is logarithmic. One is the canonical-angle reference. Lines connect the four observed post-ramp phases for legibility, not measurements at intervening updates. Table 33 summarizes all twelve pointwise ratios per arm.

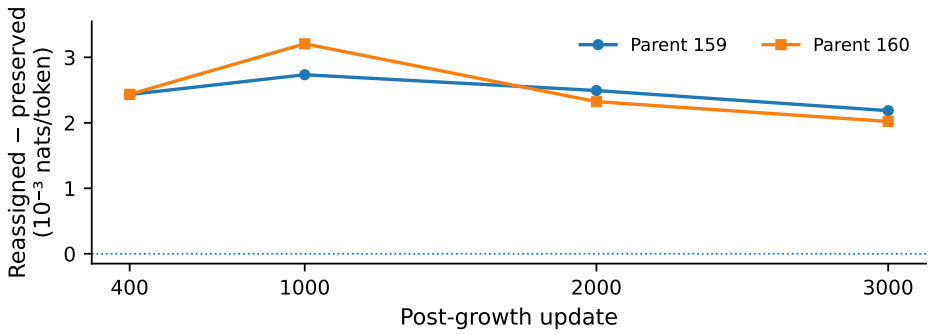

Figure 11: **The unforced preservation advantage at the recorded evaluations.** Each point subtracts the preserved DEV loss from the reassigned DEV loss for the same parent at the same update. Positive favors preservation. These are four saved evaluations from each pair of completed trajectories, not a new schedule experiment or a basis for replacing the registered final endpoint.

Extraction boundary. The original source extraction retained all seventy-two applied-geometry records and the sixteen saved DEV evaluations, with result-file hashes and observation times. The plotting input supplied with this revision contains the exact forty-eight post-ramp records and those sixteen DEV evaluations. The original summaries are retained. This is a re-summary of saved records, not new training, checkpoint evaluation, or tensor reproduction. The earlier reference implementation checked the algebra on small float64 CPU fixtures, separately from the frozen production optimizer.

Q.3 Completion and recovery record

All five parent-160 arms completed before recovery of parent 159. The latter's preserved-unforced trajectory was interrupted with an existing update-2377 reservation. A separately recorded recovery replayed 2,176 already charged updates from the common start, verified exact retained states at phases 1,000 and 2,000, and consumed the stranded reservation once before continuing. It then completed the previously unstarted reassigned-unforced arm. Replay work and three replay evaluations are separate from the original scientific allowance. No completed parent-160 endpoint was replaced.

The complete analysis and success record were written at 2026-09-18 05:43:56 UTC. The frozen scientific source identifier is

`6b4477b9efadaa932b579eb3d3c8fd7064a0`
`b24783517a33a119d94b3431952c`

The success receipt identifies analysis SHA256

```
52890eeced9937041f11eaf640083e278
2137c6caa7ae3aaf6029705c0c3f049
```

The revised numerical tables are recomputed from the completed recorded endpoints. This manuscript revision does not claim a new training execution, an independent reload of every checkpoint tensor, or a new full-source audit.

R ANCESTRY-SECTOR OPERATION CONTRAST

The operation study separates the pairing of layers from the amplification or suppression of their update differences. The original design also includes optimizer anchors, whose outcomes are reported below.

ASQ1 tests a fixed seven-arm development design on reused OWN-S1-R1 parents 167–168. Every arm continues its original BCR1 LR-0.0006 shared update-200 state to update 3,000. Fourteen 2,800-update continuations total 39,200 scientific updates, excluding engineering validation. Each final DEV score covers 1,048,576 targets. No REPORT evaluation is used.

Sector rule. For a fixed pairing, decompose the full centered FFN proposal as $U_c = A + D$, where A contains the pair-average component and D the within-pair difference. Construct $A + \rho_t D$, recenter, restore the original pooled centered norm across *all twelve layers*, and restore the all-depth common component. The relative weight ramps from identity over updates 201–400 to either $\rho = 0.5$ or $\rho = 2$. True pairs are $(j, j + 6)$. The fixed false pairing uses $(j, 6 + (j + 1) \bmod 6)$. These operations treat FFNs only; attention uses AdamW. The combined-BDT anchor additionally treats attention.

The relative weight is applied before pooled normalization. Thus $\rho = 2$ does not exactly double the final absolute difference component, and changing ρ also changes the normalized pair-average component. No pair is normalized separately. True and false rules do not match their realized angles or individual layer norms.

Table 34: All ASQ1 final DEV endpoints. Values are retained at the nine-decimal display precision of the final source readout. The final column is the reported descriptive mean difference, not an interval over fourteen independent models.

Policy	Parent 167	Parent 168	Mean − combined
AdamW	3.662995486	3.661960855	+0.002571344
FFN-only BDT	3.660906579	3.659816228	+0.000454578
Combined BDT	3.660453405	3.659360247	0
True pairs, $\rho = 2$	3.660508789	3.659418248	+0.000056692
False pairs, $\rho = 2$	3.665048059	3.663987856	+0.004611132
True pairs, $\rho = 0.5$	3.668734543	3.667599138	+0.008260014
False pairs, $\rho = 0.5$	3.664176226	3.663045883	+0.003704228

Fixed gain decision. Advancement requires at least 0.001 nats/token mean improvement over combined BDT and a strict improvement on both parents. Selection between the two true-pair weights occurs only after the full cohort. The better choice, $\rho = 2$, remains worse than combined by $+0.000055383891$ and $+0.000058000907$. The mean is $+0.000056692399$. The advancement flag is false. It improves on AdamW by 0.002514651977 and on FFN-only BDT by 0.000397885218 on average, but numerical proximity to combined BDT is not equivalence or evidence of the same mechanism.

Pairing-by-operation crossover. The same-scale false-minus-true contrasts are $+0.004539269954$ and $+0.004569608718$ at $\rho = 2$, and -0.004558317363 and -0.004553254694 at $\rho = 0.5$. The two-parent means are $+0.004554439336$ and -0.004555786029, respectively. Their difference is $+0.009110225365$. True ancestry is helpful relative to false pairing under difference amplification and harmful under difference suppression. Both false-pair policies and true-half are worse than AdamW.

This is a secondary structural observation from the fixed development grid, not fresh confirmation of a theory chosen afterward.

The true-double policy outperforms combined BDT at every stored evaluation from update 400 through 2,750, then finishes slightly worse at the registered endpoint. This horizon-dependent ordering does not show that continued application erases an advantage, or that switching would retain it. No earlier endpoint replaces the final comparison.

Record verification. The recorded final audit passed 162 source/identity/arithmetic checks, including all fourteen endpoint bindings and their histories. It did not freshly reload every checkpoint or token-cache array. Functional specialization is an explanatory hypothesis, not directly measured by these update manipulations.

S EXECUTION COMPATIBILITY AND QUALITY COMPOSABILITY

S.1 ENGINEERING SCOPE: CMP1

CMP1 is a completed full-size integration study with fifteen cells across three optimizer conditions and five execution stacks. It checks the existing combined implementation with BF16, actual FlashAttention forward/backward operations, non-reentrant activation checkpointing, gradient accumulation, and ordinary weight export. All seven summary check groups passed. The final ledger records 542/542 logical updates and no pending debit; engineering and instrumented timing are not independent learning-quality replications.

The observed combined-BDT peak allocated memory was 13.109 GiB under math/BF16, 10.708 GiB under Flash/BF16, and 4.460 GiB under the combined memory-saving stack. These belong to the CMP1 measured configurations, not automatically to the complete QCP1 training paths. The study does not establish compatibility with a different base optimizer, distributed sharding, sparse inference, or untested compiler paths.

S.2 FOUR-PARENT QUALITY COMPARISON: QCP1

The quality comparison tests whether BDT's gain is retained under the fast and lower-memory execution stacks. It complements the preceding integration checks with complete training trajectories.

QCP1 trains fresh six-layer parents 932101–932104 for 7,630 updates at peak LR 0.0008, then performs moment-preserving six-to-twelve growth. Each parent supplies four fixed 3,000-update continuations: native fused AdamW or the existing combined FFN/QKV/output BDT recipe, each under the fast and lower-memory stacks. The post-growth LR is 0.0006 for both methods, inherited from the earlier selection lock. Method/stack order is counterbalanced across parents.

The fast stack uses forced FlashAttention, BF16, microbatch 16, accumulation four, and no activation checkpointing. The lower-memory stack keeps that precision/backend, adds non-reentrant checkpointing, and uses microbatch eight with accumulation eight. Both use effective batch 64 and context 512. All sixteen final endpoints are locked before REPORT, and all use the same forced-Flash, microbatch-16 evaluation path. The inferential sample is four parents, not sixteen independently trained families.

Table 35: All QCP1 final REPORT losses. Native denotes native fused AdamW; BDT denotes the existing combined recipe. The table uses a common evaluator across both training stacks.

Parent	Native, fast	BDT, fast	Native, memory	BDT, memory
932101	3.621676	3.618846	3.621777	3.618993
932102	3.620697	3.618280	3.620814	3.618272
932103	3.623290	3.620628	3.623487	3.620850
932104	3.622680	3.619919	3.622861	3.620133

For parent s and stack k, define $\delta_{s,k} = L_{s,\mathrm{BDT},k} - L_{s,\mathrm{native},k}$. The interaction is $\delta_{s,\mathrm{memory}} - \delta_{s,\mathrm{fast}}$. Report ordinary paired 95% intervals and Bonferroni-adjusted intervals for the three-contrast family. The latter use $t_{3,\,1-0.05/6}$ for nominal simultaneous 95% coverage under the paired-t assumptions.

Table 36: QCP1 final paired contrasts. Ordinary and family-adjusted intervals describe variation over four pretrained parents conditional on the fixed recipes and evaluation text.

Contrast	Mean	Ordinary 95% interval	Adjusted interval
Fast	-0.002667612	$[-0.002955021, -0.002380204]$	$[-0.003106220, -0.002229005]$
Memory	-0.002672927	$[-0.002842214, -0.002503640]$	$[-0.002931271, -0.002414583]$
Interaction	-0.000005315	$[-0.000133092, +0.000122463]$	$[-0.000200313, +0.000189684]$

Both stack-specific effects favor BDT on all four parents and retain negative adjusted intervals. The interaction has mixed signs and an interval crossing zero. This establishes retention of the quality gain under the tested memory-saving features, not extra quality synergy or formal equivalence between stack effects. Native AdamW versus the partitioned BDT implementation is a selected-recipe comparison; not every implementation distinction is separately isolated.

Quality and time are separate. All sixteen trajectories first cross the preselected DEV target 3.70 at the recorded update-2,500 evaluation, with the same $[2250, 2500]$ crossing bracket. The cadence does not resolve earlier exact crossing or token-to-target savings. Shared instrumented timings are not uncontended production benchmarks.

Completion and accounting. The final success record is dated 2026-09-18 01:39:50 UTC. The original scientific ledger reconciles 30,520 parent updates and 48,000 continuation updates, plus 39 engineering updates, eight engineering forward calls, and 228 evaluation calls scoring 466,944 sequences. A forced-relogin recovery adds 986 separately recorded replay updates and four evaluations (8,192 sequences); exact retained step-500 model, optimizer, and random-state agreement was checked. These replay costs are not included in the 78,520 original scientific-update total. Completed outcomes were preserved, and the final ledgers have no pending debit.

The final analysis SHA256 checked against the success receipt is

`e2f3dda7e2d64702e745f98cd665e29ca5a`

`fce36a40fe9d2b4cc7652ea7e9949`

The means and intervals reproduce from the retained endpoints. This is record-level verification, not a new checkpoint evaluation or independent end-to-end reproduction.

T LONG-HORIZON ANCESTRY CONFIRMATION

T.1 FIXED DESIGN AND INDEPENDENT UNITS

The long-horizon study tests the ancestry contrast and the improvement over AdamW at a separately fixed, longer endpoint. The two comparisons retain separate inferential conclusions.

LH1 uses four independently pretrained six-layer Dense parents, 946201–946204, with 81,074,688 parameters each. Parent training lasts 7,630 updates at peak LR 0.0008. Whole-block copying forms the independent-storage twelve-layer order $[0, 1, 2, 3, 4, 5, 0, 1, 2, 3, 4, 5]$, with 123,551,232 parameters. Named Adam moments and global age are transferred; model and random-state provenance is retained.

The post-growth schedule is fixed at 15,000 updates from the start: peak LR 0.0006, 200-update warmup, cosine decay to 0.1 of peak. It is not an extension of a cooled 3,000-update trajectory. A shared 200-update AdamW prefix accumulates canonical and chart histories. Each parent then supplies AdamW, preserved-draw-zero, and reassigned-draw-zero continuations. The FFN-only cap-three treatment uses the inherited onset, 200-update ramp, cadence ten, and EMA coefficient 0.9. Preserved and reassigned proposals match their own-state common-depth component, pooled centered norm, and canonical angular departure. There is no additional canonical-BDT or attention-treatment science arm in this study.

Data are the full, hash-bound S1-B realization. Context 512, microbatch 16, and accumulation four give 32,768 targets per update and 491.52M exposures over the post-growth path. Other AdamW and

precision settings follow Appendix A. All twelve final endpoints are locked before REPORT, which scores 1,048,576 targets at validation-array offset 2,500,000.

The two prespecified contrasts are reassigned-minus-preserved (predicted positive) and preserved-minus-AdamW (predicted negative). Each has four paired parent-level differences. Individual 97.5% Student-t intervals with three degrees of freedom implement the two-comparison Bonferroni adjustment for nominal simultaneous 95% coverage. The cohort is not pooled with BAC1, and neither retained intermediate checkpoints nor reuse of these parents in later studies increases n.

T.2 Completed REPORT outcomes

All twelve arms reach the fixed endpoint. Both contrasts have their predicted sign on every parent (Table 37). The ancestry-preservation contrast excludes zero under the adjusted interval; constructive superiority to AdamW does not (Table 38). The latter's ordinary 95% interval excludes zero, but the prespecified two-contrast adjustment determines the stated conclusion.

Table 37: **All LH1 endpoints at post-growth update 15,000.** REPORT nats/token. P denotes preserved, R reassigned. Four parents. No selected checkpoint or excluded endpoint.

Parent	AdamW	P	R	P − AdamW	R − P
946201	3.404692	3.404546	3.404747	−0.000146	+0.000201
946202	3.406946	3.406457	3.406762	−0.000489	+0.000305
946203	3.406948	3.406575	3.406915	−0.000373	+0.000340
946204	3.405261	3.405036	3.405389	−0.000225	+0.000354

Table 38: **Distinct longer-horizon conclusions.** The adjusted 97.5% intervals correspond to the two prespecified contrasts. Means and intervals are recomputed from the unrounded REPORT records.

Contrast	Mean	Adjusted 97.5% interval
Preserved − AdamW	−0.000308	[−0.000627, +0.000011]
Reassigned − preserved	+0.000300	[+0.000156, +0.000444]

LH1 establishes persistence of the relational contrast at the tested longer endpoint. It does not establish an asymptotic loss floor, a universal convergence-rate change, or equal effect sizes across horizons. In particular, comparing its effect magnitude with BAC1 is not a controlled estimate of temporal attenuation: the cohorts and prespecified schedules differ.

T.3 Evidence and verification scope

The completed record contains twelve REPORT receipts, an endpoint-cohort lock, and a final analysis linked by source and checkpoint identities. The recorded result/contract/latest links, REPORT identities, and analysis hash were checked, and all 56 frozen source files matched their manifest. Means, sample standard deviations, standard errors, and ordinary and adjusted intervals recompute from the twelve unrounded scores.

The completion audit checked checkpoint presence and stored identity links, rather than rehashing all checkpoint payloads or performing a new tensor-level evaluation. The original finalizer required payload hashes before committing the analysis. The accompanying numerical file retains the exact REPORT values, and the verification script reproduces the two registered contrasts.

U Construction, pair-average correction, and treatment-off persistence

These interventions examine three parts of the correspondence effect: construction of the operator, the correction it applies, and the state that remains after treatment is removed. Crossed surrounding trajectories connect the correction and persistence tests.

U.1 Scope, source states, and statistical units

EAC1, DNT1, DNT2, and TOC1 use the same two original growth-development parents, 159 and 160, on S1-B. CTA1 supplies TOC1's source states. They do not add independent parents across studies. The ordinary 123,551,232-parameter Dense model, effective batch 64, context 512, optimizer, and copying construction follow Appendices A and D. The post-growth peak LR is 0.0006. Its original cosine schedule ends at 3,000 and is not shortened to the diagnostic endpoint. BDT uses the inherited FFN-only cap-three recipe.

DNT1, DNT2, and CTA1 continue phase 200 to 1,200; EAC1 continues to 3,000; TOC1 applies 800 more updates to the stated CTA1 phase-1,200 states. Each fixes its endpoint before evaluation. Final DEV scores 2,048 records and 1,048,576 targets from offset zero after the study's all-endpoint barrier, with no intermediate DEV selection or new REPORT. Contrasts are descriptive, without population intervals, equivalence tests, or mediation fractions. Different horizons and surroundings prevent pooling absolute losses or treating gains as additive parts of one effect.

Shared verification scope. The September 22 audit matched source/protocol identities and SUCCESS-to-analysis hashes, and compared retrieved endpoint metadata with the saved locks. Individual DEV receipts additionally matched DNT1 and newly trained EAC1 scores. DNT2, CTA1, and TOC1 scores come from their bound final analyses and joint per-parent endpoints. No checkpoint payloads were rehashed or tensors reloaded, and no training or evaluation was rerun. The companion evidence file retains full score precision and record identities.

U.2 Estimation/application crossover (EAC1)

The ancestry assignment can enter both the observed depth statistic and the application of the resulting operator. EAC1 crosses these roles using the original preserved and reassigned draw-zero charts. Denote the estimation assignment by $e \in \{\mathrm{P}, \mathrm{R}\}$ and the application assignment by $a \in \{\mathrm{P}, \mathrm{R}\}$. At each branch's current state, the native correspondence bank advances its original canonical and chart histories once. The operator is read from history e, applied to the raw proposal transported through chart a, and mapped back through the inverse of a. The resulting direction is matched to that branch's hypothetical canonical proposal using Equation 7. No second observation is made during application. Both crossed candidates are constructed symmetrically. Only the assigned one is committed.

PP and RR are the original matched preserving and reassigned controls. PR estimates with preservation and applies with reassignment; RP does the reverse. The two new crossed arms per parent contribute 11,200 updates. The six linked AdamW/PP/RR anchors retain their original identities and are not new replications. Four new and six linked endpoints form the ten-cell evaluation barrier. Writing L with a cell subscript for its final loss, the registered descriptive contrasts are $L_{\mathrm{RP}} - L_{\mathrm{PP}}$ and $L_{\mathrm{PR}} - L_{\mathrm{PP}}$.

Table 39: **Complete EAC1 final DEV scores at phase 3,000.** First letter: estimation chart; second: application chart. The last column is mean loss minus same-parent AdamW. A, PP, and RR are linked anchors; PR and RP are new.

Cell	Parent 159	Parent 160	Mean − AdamW
A	3.661593	3.664348	+0.000000
PP	3.659485	3.662230	−0.002113
PR	3.661924	3.664623	+0.000303
RP	3.662649	3.665415	+0.001062
RR	3.662580	3.665426	+0.001033

Both crossed cells are worse than PP on both parents. Their mean penalties are 0.003175 for RP and 0.002416 for PR. PP alone improves on AdamW in this tested factorial. The interaction $L_{\mathrm{PP}} - L_{\mathrm{PR}} - L_{\mathrm{RP}} + L_{\mathrm{RR}}$ is -0.002445 on average. Coherent preservation is therefore useful here, while the coherent reassigned RR condition is not. This is not a claim that arbitrary mismatched estimators must fail or that preserved estimation by itself universally suffices.

U.3 PAIRED MEAN AND BATCH-CHOICE FLUCTUATION (DNT1/DNT2)

The paired-batch construction separates the correction averaged across two batches from the fluctuation associated with selecting one batch. Both components are defined at the same model and optimizer state.

At the same model and optimizer state, evaluate two independently sampled effective batches and an independent fair choice $i \in \{1, 2\}$. Let U_i be the signed raw AdamW FFN proposal for the selected batch and V_i its hypothetical canonical BDT proposal; write $\delta_i = V_i - U_i$ for its correction and $\bar{\delta}$ for the mean correction over the two batches. The four policies apply

arm	FFN increment
A	U_i
F	$U_i + \delta_i$
M	$U_i + \bar{\delta}, \quad \bar{\delta} = (\delta_1 + \delta_2)/2$
N	$U_i + (\delta_i - \bar{\delta})$.

The same two-batch schedule and selection are used across arms. Hypothetical moment and observer changes are restored before the selected transition is committed. Correction mixing uses FP64 before casting to FP32; native A/F outputs are applied directly. Each update commits only the selected batch's Adam transition and applies decay once. The mixed M/N proposals are not renormalized afterward. Thus these policies do not inherit all of the common/norm/angle matching constraints of the earlier direction controls.

Conditional on a fixed pair, the residual averages to zero over the independent choice and the pair average does not depend on that choice. The pair average is still a noisy two-batch quantity, not the population conditional mean. Repeated state-dependent optimization also prevents equating endpoint differences with additive one-step components. The primary descriptive quantity is $L_{\mathrm{N}} - L_{\mathrm{M}}$, interpreted alongside the constructive full-correction gain $L_{\mathrm{A}} - L_{\mathrm{F}}$.

DNT1 lets each arm retain its evolving non-FFN parameters. DNT2 applies the same correction comparison but copies the evolving AdamW donor A's non-FFN *weights* to all recipients after each round. Each recipient still computes its own full-model gradients and keeps its own FFN state, moments, and RNG. This holds surrounding weights to a common AdamW trajectory rather than freezing them. The two studies share the original phase-200 inputs and paired sampler. Each contains 8,000 selected updates and 16,000 full-batch gradient evaluations, with engineering charged separately.

Table 40: **All DNT1/DNT2 final DEV scores at phase 1,200.** Last column: loss minus the same study's AdamW arm, averaged over the two reused parents. A is identical across the two recorded comparisons, not an additional independent anchor.

Study	Arm	Parent 159	Parent 160	Mean − A
DNT1	A	3.845634	3.848148	+0.000000
	F	3.843162	3.845341	−0.002640
	M	3.843220	3.845543	−0.002510
	N	3.845507	3.847887	−0.000194
DNT2	A	3.845634	3.848148	+0.000000
	F	3.842400	3.845136	−0.003123
	M	3.842649	3.845375	−0.002879
	N	3.845378	3.847928	−0.000238

Both studies and both parents order F < M < N < A. The mean $L_{\mathrm{N}} - L_{\mathrm{M}}$ differences are 0.002316 in DNT1 and 0.002641 in DNT2. The average-correction effect therefore remains constructive under common evolving surrounding weights. This finding separates applied corrections, not the signal/noise information used to estimate their operator. A transformation inferred from fluctuation covariance can still change the mean update; DNT does not establish that fluctuation-derived observer statistics are unimportant. Nor does comparing these gains identify a percentage of the original ancestry-assignment effect that is mediated by the pair average.

U.4 Crossed surrounding trajectories and the treatment-off source

Here the letter subscripts identify the FFN policy and the source of the surrounding non-FFN trajectory.

CTA1 supplies the source states for TOC1. A quartet combines FFN AdamW (A) or BDT (B) with a non-FFN weight trajectory from the coherent AdamW or coherent BDT donor. The first subscript identifies the FFN policy. The second identifies the donor. All four full-model updates are computed before copying donor non-FFN weights, leaving recipient FFNs, moments, and RNG in their own state. The source horizon is phase 200–1,200 under the original 3,000-step LR schedule.

Table 41: **CTA1 final DEV source-state scores.** These phase-1,200 outcomes identify the states reused by TOC1. They are not added to the later treatment-off replication count.

Cell	Parent 159	Parent 160
$\mathrm{A_A}$	3.843395	3.847049
$\mathrm{A_B}$	3.843721	3.847584
$\mathrm{B_A}$	3.840305	3.844032
$\mathrm{B_B}$	3.840495	3.844372

The FFN BDT gain is positive under either donor: 0.003054 under AdamW surroundings and 0.003219 under BDT surroundings on average. The native-pairing interaction is about 0.000165. This is a conditional trajectory comparison, not a Hessian-block attribution or a complete explanation of the correspondence contrast. The relevant treatment-off sources are specifically $\mathrm{A_A}$ and $\mathrm{B_A}$. The latter is not the native $\mathrm{B_B}$ state.

U.5 Weights versus initial Adam moments after removing treatment (TOC1)

After removing BDT, this experiment separates inherited FFN weights from inherited Adam moments. Its letter pairs identify these two state origins, rather than the policy/donor assignment in the preceding experiment.

At CTA1 phase 1,200, form four descendants by crossing FFN weights from $\mathrm{A_A}$ or $\mathrm{B_A}$ with the named FFN Adam first/second moments from either source. The first letter of AA/AB/BA/BB denotes FFN weight origin and the second denotes moment origin. Non-FFN weights, their initial moments, the common RNG state, and unused observer history begin from $\mathrm{A_A}$. All policies then use ordinary AdamW, with BDT disabled, from phase 1,201 through 2,000. After each round, recipients receive the evolving coherent AA donor's non-FFN weights. The original LR schedule and optimizer age continue. The unused BDT history is not applied.

Table 42: **Complete TOC1 final DEV scores after 800 treatment-off updates.** Weight-origin advantages compare AA with BA and AB with BB. Initial moment-origin effects compare AA with AB and BA with BB.

Cell	Parent 159	Parent 160
AA	3.736868	3.738950
AB	3.736859	3.738988
BA	3.735679	3.737857
BB	3.735697	3.737864

Let $L_{\mathrm{AA}}, L_{\mathrm{AB}}, L_{\mathrm{BA}}, L_{\mathrm{BB}}$ be the final losses of the four weight/moment-origin cells above. Their primary weight-origin contrast G_W is

$$G_W = \tfrac{1}{2}\big[(L_{\mathrm{AA}} - L_{\mathrm{BA}}) + (L_{\mathrm{AB}} - L_{\mathrm{BB}})\big].$$

It is 0.001176 for parent 159 and 0.001108 for parent 160, averaging 0.001142. Averaged over parents, the weight gains are 0.001141 under A-origin moments and 0.001143 under B-origin moments. Moment-origin gains are -0.000014 at A-origin weights and -0.000012 at B-origin weights. The mean pairing interaction is $+0.000002$. All eight endpoints precede evaluation; 6,400 new AdamW updates and gradient evaluations are charged, separately from CTA1.

These comparisons locate a persistent difference in the tested FFN weight states relative to the choice of *initial* moment history under this continuation. They do not show that moments are always

irrelevant, that the trajectories occupy permanently distinct basins, or that phase 1,200 is an optimal stopping time for BDT. Together with EAC1 and DNT2, they constrain operator construction, applied correction, and retained state effects, respectively. A further prediction across a new intervention would need to be specified and tested separately.

V Do signal and noise observers produce distinct directions?

V.1 Question and fixed construction

The observer-component study (OSC1) asks whether signal- and noise-derived statistics supply different useful directions at matched proposal strength. It separates the statistics used to *estimate* the operator. Appendix U instead separates components of the *applied correction*. Both growth-development parents, 159–160, are reused rather than fresh confirmation units.

Native BDT, paired-total, signal-matched, and noise-matched restore the original phase-200 model, Adam, and random states, then continue to phase 3,000. They retain peak LR 0.0006, cap three, BF16, batch 64, context 512, and FFN-only treatment. The selected-batch stream is shared and arm order reverses between parents. All eight endpoints precede the 1,048,576-target DEV evaluation at offset zero; there is no intermediate DEV, REPORT, or policy selection.

What "signal" and "noise" mean here. Let X and Y be two independently sampled, same-state hypothetical Adam proposals after centering, projection into the eleven-dimensional depth basis, and individual Frobenius normalization. These are not raw gradients. Let T be the paired total second moment, N the fluctuation contribution, and S the cross-proposal signal contribution:

$$T = \tfrac{1}{2}(XX^\top + YY^\top), \qquad N = \tfrac{1}{2}(X-Y)(X-Y)^\top, \tag{14}$$

$$S = \tfrac{1}{2}(XY^\top + YX^\top), \qquad T = S + N. \tag{15}$$

Conditional on a fixed state and independent, identically distributed X, Y, the expectation of S is $\mathbb{E}[X]\mathbb{E}[X]^\top$, and that of N is the corresponding matrix fluctuation second moment. Normalization and Adam's inherited moments are part of this definition. The online histories mix successive states. They are not population moments at one fixed state.

Paired histories start from 32 independent calibration pairs at the common phase-200 state. Calibration changes no model, Adam, or random state. Every tenth training update supplies a same-prestate shadow proposal alongside the selected one. Only the selected Adam transition and weight decay are committed.

Observer update. Current observations enter *after* application, with EMA coefficient 0.9. Gram products and EMA histories use FP64. The signal history remains signed. Its negative eigenvalues are clipped only when constructing the positive operator, not when updating the history. The same cap, warmup, ramp, and normalized spectral gain formula are used for each component.

Paired-total applies the total-derived proposal. The two component arms construct their own signal/noise candidate and match its common component, centered norm, and angle to the hypothetical paired-total proposal at the *same branch state*, using Equation (7). Native BDT retains its inherited observer. Thus component-versus-total tests change the candidate direction's statistical source while retaining a total-derived angle; native-versus-paired comparisons additionally change calibration, lag, and numerical estimation. A component arm is not a standalone signal-only or noise-only optimizer.

V.2 Complete fixed-endpoint outcomes

Signal-matched is slightly worse than paired-total on both parents; noise-matched splits signs and its mean difference is near zero. The entire four-arm loss spread is 0.00003380 on parent 159 and 0.00001990 on parent 160. These small contrasts alone do not determine whether the two statistical sources contain interchangeable information. The achieved directional manipulation is essential to that interpretation.

Table 43: **All eight observer-component DEV losses.** Lower is better. The primary comparisons are each matched component minus paired-total. Values are rounded for display; full-precision receipts are supplied.

Policy	Parent 159	Parent 160
Native BDT	3.65832105	3.66298019
Paired-total	3.65829457	3.66297498
Signal-matched	3.65832837	3.66298886
Noise-matched	3.65829818	3.66296896

Table 44: **Prespecified component contrasts.** Units are 10^{-6} nats/token. Both parent-level differences and their descriptive mean are shown. No population interval or equivalence inference is used.

Component − paired-total	Parent 159	Parent 160	Mean
Signal-matched	$+33.80$	$+13.88$	$+23.84$
Noise-matched	$+3.61$	-6.02	-1.21

V.3 Retrospective check of the directional manipulation

Let U be the raw AdamW proposal and $u = HU$ its centered component at the recorded branch state. Let V_k and V_T be the matched component and paired-total proposals, respectively. Their unit AdamW-orthogonal residual directions are d_k and d_T, with k indexing the signal or noise component. Write $c = \langle d_k, d_T \rangle_F$ for their cosine and θ for the total proposal's departure from u. These compare orthogonal residual directions rather than full updates dominated by their shared Adam component. Under exact common/norm/angle matching,

$$\frac{\|V_k - V_T\|_F}{\|u\|_F} = \sin\theta\sqrt{2(1-c)}, \tag{16}$$

$$\frac{\|V_k - V_T\|_F}{\|V_T - U\|_F} = \cos(\theta/2)\sqrt{2(1-c)}. \tag{17}$$

Both identities follow from Equation (7) and $\|d_k - d_T\|_F^2 = 2(1-c)$. Distances use these ideal-geometry formulas event by event before averaging; applied FP32 tensors are not reloaded.

The audit covers phases $400, 425, \ldots, 3000$: 105 records per parent/projection, hence 630 per component across two parents and three FFN projections. These are descriptive repeated measurements, not independent units. Each total reference uses its component branch's own state, not another trajectory's tensors.

Table 45: **The component interventions scarcely separate the direction.** Unweighted means over 630 logged post-ramp events per component. The bracketed cosine ranges are observed minima and maxima, not confidence intervals. Distance rows use Equations (16)–(17), expressed as percentages.

Diagnostic	Signal-matched	Noise-matched
Residual-direction cosine, mean	0.999997	0.999946
Observed cosine range	[0.999983, 1.000000]	[0.999735, 0.999993]
Distance / total correction (%)	0.227	0.990
Distance / raw centered norm (%)	0.0373	0.1626
Matching amplification, mean	0.998798	1.005237
Largest negative spectral mass	0	0

At every sampled phase, signal/noise operators have rank eleven and zero negative spectral mass: PSD clipping is inactive. Matching amplification is near one. Thus the close losses accompany near-coincident residual directions, not evidence that substantially different directions have equal utility.

A sufficient condition for identical operators. The gains depend on $\lambda_i/\lambda_{\max}$. Let C_S, C_N, C_T be the signal, noise, and total observer matrices. If $C_S = aG$ and $C_N = bG$ for nonzero positive-semidefinite G and $a, b > 0$, then $C_T = (a+b)G$ and all three induce the same spectral operator.

The logged cosines do not establish this matrix proportionality or explain the near-collinearity; direct same-state operator measurements are needed.

Audit and scope. The final SUCCESS, analysis, endpoint seal, and all eight DEV receipts have matching hash and identity bindings. All six contrasts recompute exactly. The saved ledger reports 22,400 proved model updates. Interrupted work and replay remain separately recorded in the supplied evidence. The two retained logs support a retrospective manipulation check, not fresh checkpoint-tensor reproduction. These results motivate separating upstream statistical similarity from similarity introduced by the transform before another information-source intervention. They do not justify selecting a component from the tiny loss difference.

W ADJACENT-COPY ANCESTRY TRANSFER AND STRUCTURAL REFINEMENT

W.1 SHARED FAMILIES, COPYING LAYOUT, AND EVALUATION

The adjacent-copy studies use the four pretrained families 946201–946204 from the longer-horizon study (Appendix T), but copy each six-layer parent's blocks into adjacent descendants $(2j, 2j+1)$, $j = 0, \ldots, 5$. The target is the same 123,551,232-parameter Dense model. The parent is pretrained for 7,630 updates; weights and Adam moments are copied, Adam age continues, and a new 3,000-update post-growth schedule uses peak learning rate 0.0006, cap three, and a 200-update inactive prefix followed by a 200-update ramp. All branches restore the registered model, moments, random state, and proposal histories. These studies reuse four families. Neither a new copying layout nor additional branches increases the independent sample count.

The ancestry-assignment test (AGT1) fixes one chart draw per parent, comparing AdamW with ancestry-preserving and reassigned charts. It retains the chart-multiset and own-state common/norm/angle construction of the balanced comparison, now using adjacent ancestry. The fixed-rule transfer (FRT1) adds one continuation per parent using the true-pair amplification rule, without consulting BDT for its applied FFN proposal. Non-FFN parameters retain AdamW. Its four endpoints and all twelve ancestry-test endpoints are locked before the prescribed sequential REPORT evaluations.

Table 46: **Adjacent-copy REPORT outcomes.** Each row is one reused parent. The fixed-rule column is the additional FRT1 continuation. The first three columns are AGT1. All reach post-growth phase 3,000.

Parent	AdamW	Preserved	Reassigned	Fixed rule
946201	3.63141313	3.62540965	3.63373289	3.62582429
946202	3.63244062	3.62601174	3.63447030	3.62668452
946203	3.63383235	3.62781345	3.63587407	3.62829225
946204	3.63484978	3.62874205	3.63682014	3.62930057

Table 47: **Registered adjacent-copy comparisons.** Paired Student-t intervals use four parent differences and three degrees of freedom. The first two are AGT1's adjusted 97.5% intervals. The third is FRT1's single primary 95% interval. All three criteria are met.

Contrast	Mean	Interval	Level
Reassigned − preserved	$+0.00823013$	$[+0.00782539, +0.00863487]$	97.5%
Preserved − AdamW	-0.00613975	$[-0.00655354, -0.00572596]$	97.5%
Fixed rule − AdamW	-0.00560856	$[-0.00576866, -0.00544847]$	95%

Reassignment is worse than preservation on every parent, and preservation improves on AdamW on every parent. The fixed rule also improves on AdamW throughout, while trailing the preserved-chart transformation by 0.00053119 on average. This supports ancestry sensitivity and fixed-rule usefulness in the specified adjacent-copy setting. Because the horizon and construction differ from LH1, these comparisons do not isolate a causal effect of layout or estimate a decay rate between studies.

W.2 THE FIXED RULE AND DIFFERENCE-ONLY REFINEMENT

Let U be the raw proposal, $U_m = (I - H)U$ its common component, and J the matrix that exchanges descendants within each adjacent pair. Here H is the all-depth centering matrix from Equation 1 and I is the depth identity. Define the pair-average and pair-difference projectors P_A, P_D and their projected components A, D by

$$P_A = \tfrac{1}{2}(I + J)H, \qquad P_D = \tfrac{1}{2}(I - J)H, \qquad A = P_A U, \quad D = P_D U, \quad A + D = HU. \tag{18}$$

With activation ramp $a_t = \mathrm{clip}((t - 200)/200, 0, 1)$ and relative difference weight $\rho_t = 2^{a_t}$, the fixed FFN proposal $\mathcal{F}_t(U)$ is

$$\mathcal{F}_t(U) = U_m + \frac{\|HU\|_F}{\sqrt{\|A\|_F^2 + \rho_t^2 \|D\|_F^2}}(A + \rho_t D). \tag{19}$$

A zero-centered proposal is unchanged. The rule preserves the all-depth common component and pooled centered norm. It does not separately match angular departure or individual layer norms.

W.3 REFINEMENT FACTORIAL AND COMPLETE OUTCOMES

ART1 retains the fixed amplification in all four factorial cells. Difference refinement replaces D by a cap-three, difference-sector-conditioned direction $\widehat{D}$ rescaled to $\|D\|_F$, while leaving A unconditioned. Its EMA uses the full centered norm, decay 0.9, and cadence ten. The difference covariance is initialized by restricting the warmed phase-200 canonical statistic to the adjacent-difference basis. The resulting structured FFN proposal $\mathcal{S}_t(U)$ is

$$\mathcal{S}_t(U) = U_m + \frac{\|HU\|_F}{\sqrt{\|A\|_F^2 + \rho_t^2 \|D\|_F^2}}(A + \rho_t \widehat{D}), \qquad \|\widehat{D}\|_F = \|D\|_F. \tag{20}$$

Within-sector conditioning therefore changes orientation while retaining the fixed rule's sector allocation at that state. The attention refinement applies cap-three BDT to QKV and output proposals. Attention observer histories are warmed by replaying the original phase-0–200 AdamW prefix, with exact agreement required for the recorded weights, moments, random state, original histories, and batch sequence. These reconstruction gradients are accounted separately, not counted as new parent prefixes.

Table 48: **Meaning of ART1's six cells.** "Attention-refined" is not an attention-only optimizer: its FFNs retain fixed ancestry amplification. Combined BDT is an additional reference, not a factorial cell.

Cell	FFN proposal	Attention
AdamW	Raw AdamW	AdamW
Fixed (H)	$\mathcal{F}_t(U)$	AdamW
Difference-refined (D)	$\mathcal{S}_t(U)$	AdamW
Attention-refined (HA)	$\mathcal{F}_t(U)$	BDT
Assembled (DA)	$\mathcal{S}_t(U)$	BDT
Combined	Full-depth BDT	BDT

Sixteen new 2,800-update continuations and eight linked AdamW/fixed-rule endpoints form the 24-cell cohort. All are sealed, and AGT1/FRT1 reporting is completed, before ART1 REPORT. The two primaries are assembled-minus-fixed and assembled-minus-combined, each tested with an adjusted 97.5% paired interval; joint success requires both upper endpoints below zero.

Table 49: **Complete ART1 REPORT scores.** Columns follow Table 48. The independent unit remains the parent.

Parent	AdamW	Fixed	Difference	Attention	Assembled	Combined
946201	3.63141313	3.62582429	3.62568665	3.62468247	3.62466877	3.62413008
946202	3.63244062	3.62668452	3.62647061	3.62571845	3.62559070	3.62500163
946203	3.63383235	3.62829225	3.62813171	3.62726878	3.62711731	3.62670819
946204	3.63484978	3.62930057	3.62913438	3.62795999	3.62793025	3.62743120

Table 50: **ART1 comparisons.** Only the first two are registered primaries; remaining 95% intervals are descriptive. Differences are in nats/token.

Contrast	Mean	Interval	Level
Assembled − fixed	-0.00119865	$[-0.00144833, -0.00094897]$	97.5%
Assembled − combined	$+0.00050898$	$[+0.00035010, +0.00066787]$	97.5%
Difference-refined − fixed	-0.00016957	$[-0.00022054, -0.00011859]$	95%
Attention-refined − fixed	-0.00111798	$[-0.00138126, -0.00085470]$	95%
Assembled − difference-refined	-0.00102908	$[-0.00124105, -0.00081711]$	95%
Assembled − attention-refined	-0.00008067	$[-0.00019056, +0.00002922]$	95%
Assembled − AdamW	-0.00680721	$[-0.00695785, -0.00665658]$	95%

Assembled improves on the fixed rule but remains worse than combined BDT on every parent, so the joint primary criterion is not met. The extra change from attention-refined to assembled has mean -0.00008067 with a descriptive interval crossing zero. The mean factorial interaction, $L_{\mathrm{DA}} - L_{\mathrm{D}} - L_{\mathrm{HA}} + L_{\mathrm{H}}$, is $+0.00008890$. No additive-gain claim is inferred. These outcomes motivate testing what the structured reduction omits rather than treating numerical proximity as equivalence.

X Prospective transfer of a sector-budget ordering

X.1 Development observations and the fixed prediction

The budget-transfer study (BCT1) was specified after the complete four-parent allocation/orientation development audit (AOT1 DEV, September 22) and before its own training or outcomes. Development crossed full-BDT versus structured sector norms with the two constructions' within-sector directions. Relative to combined BDT, mean DEV differences were $+0.00047838$ for assembled, -0.00001545 for full budgets with structured directions, and $+0.00051403$ for structured budgets with full directions. These development observations generated the hypothesis. AOT1's registered REPORT analysis and the prospective BCT1 evaluation remain separate results.

The fixed prediction is that *full-BDT sector budgets outperform structured sector budgets when both use a new, common direction-construction rule*. The four adjacent-copy families and the phase-200 initial states are unchanged. Two new arms per family continue to phase 3,000, giving eight new endpoints; combined and assembled are eight linked ART1 anchors. A 16-cell seal and prior ART1 REPORT closure precede the primary evaluation. No smoothing coefficient, endpoint, or contrast is selected from BCT1 outcomes.

Table 51: **Development data used to formulate the prediction.** Completed AOT1 DEV, not its registered REPORT analysis. Full/structured identify budget and within-sector-direction sources, in that order. Every row is one of the same four reused families. The values reproduce the inherited September 22 audit.

Parent	Full/full	Full/structured	Structured/full	Structured/structured
946201	3.664070215	3.664024493	3.664523583	3.664520828
946202	3.665533092	3.665522838	3.666088711	3.666091770
946203	3.665256867	3.665240949	3.665742422	3.665649764
946204	3.664749162	3.664759265	3.665310748	3.665260488

X.2 Direction construction and own-state budget exchange

Use the projectors P_A, P_D in Equation 18. For a raw proposal U, write $U_m = (I - H)U$. Let K_t be the depthwise smoothing matrix, $G_t(U)$ its output, and α_t its mixing coefficient. The matrix L_{path} is the twelve-node path Laplacian, and I the identity. Window smoothing mixes neighboring layers' proposals across depth. It supplies the within-sector directions, while the selected budget source supplies their magnitudes:

$$G_t(U) = K_t U, \qquad K_t = I - \tfrac{1}{2}\alpha_t L_{\text{path}}, \qquad \alpha_t = 0.1 a_t, \tag{21}$$

Interior rows have weights $(\alpha_t/2, 1 - \alpha_t, \alpha_t/2)$. At the first and last layers, the layer's own proposal replaces the missing neighbor. This is the Window Smoothing construction of Meng et al. (2026), delayed and ramped here. No claim that smoothing universally suppresses the ancestry-difference energy is required.

At its own model, moments, gradient, and observer state, each branch computes the full FFN BDT candidate $\mathcal{B}_t(U)$ and the structured candidate $\mathcal{S}_t(U)$. For update rule $X \in \{\mathcal{B}_t, \mathcal{S}_t\}$ evaluated at that state, let $\beta_A^X = \|P_A X(U)\|_F$ and $\beta_D^X = \|P_D X(U)\|_F$ be its sector norms. The hybrid proposal V_t^X uses these norms with the normalized smoothed sectors:

$$V_t^X = U_m + \beta_A^X \frac{P_A G_t(U)}{\|P_A G_t(U)\|_F} + \beta_D^X \frac{P_D G_t(U)}{\|P_D G_t(U)\|_F}. \tag{22}$$

The two policies share a direction-construction rule, not identical tensors along their diverging trajectories. Both compute both candidate budget sources using one gradient and one committed Adam transition; decoupled weight decay is applied once. Attention receives the same inherited BDT rule. All non-target state follows the original continuation recipe.

The raw common component and pooled centered norm are preserved, as are the chosen source's sector norms. No post-mix normalization changes that allocation. Angles, individual layer norms, and later trajectories are not matched. A zero target sector budget contributes zero; positive budget with a zero carrier sector is an explicit failure, not a substituted direction. No donor schedule or alternative fallback is introduced.

X.3 Workload and primary decision

The eight continuations use $4 \times 2 \times 2{,}800 = 22{,}400$ effective-batch updates, excluding inherited pretraining and prefixes. BF16, context 512, microbatch 16, four accumulation steps, LR 0.0006, Adam age offset 7,630, and the exact S1-B token arrays are retained. Writing $L(V^X)$ for the final reporting loss of the continuation applying budget-source policy X, the primary is $L(V^{\mathcal{B}}) - L(V^{\mathcal{S}})$. Its paired two-sided 95% Student-t interval over four parent differences supports the prediction when the upper endpoint is below zero. A zero estimated variance requires review. Anchor contrasts are descriptive secondary comparisons. Reused parents and earlier development rule out treating this as fresh-parent confirmation.

X.4 Complete REPORT results and the remaining gap

Table 52: **BCT1 REPORT scores.** Both new arms use GS-derived within-sector directions. Full/structured denote the budget sources; combined and assembled are linked ART1 anchors.

Parent	Full budgets	Structured budgets	Combined	Assembled
946201	3.62421823	3.62483430	3.62413008	3.62466877
946202	3.62511200	3.62586524	3.62500163	3.62559070
946203	3.62678868	3.62736315	3.62670819	3.62711731
946204	3.62753614	3.62806486	3.62743120	3.62793025

The primary differences are -0.00061607, -0.00075324, -0.00057447, and -0.00052872. Their mean is -0.00061813, and the registered interval lies below zero. The full-budget construction still trails combined BDT by 0.00009599 on average. The result supports a specific prospective allocation ordering across an additional direction construction, not arbitrary-direction interchangeability, a complete explanation of the ancestry effect, or equivalence to full BDT. Because the full-budget arm retains BDT's observer and no standalone-GS reference is tested here, it establishes neither a BDT-free method nor a repair of standalone GS.

Table 53: **BCT1 paired differences.** All intervals are 95%, with three degrees of freedom. The first row is the sole registered primary; anchor comparisons are descriptive.

Contrast	Mean	Interval	Level
Full budgets − structured budgets	−0.00061813	[−0.00077229, −0.00046396]	95%
Full budgets − combined	+0.00009599	[+0.00007370, +0.00011827]	95%
Structured budgets − combined	+0.00071412	[+0.00054870, +0.00087953]	95%
Full budgets − assembled	−0.00041300	[−0.00051855, −0.00030744]	95%
Structured budgets − assembled	+0.00020513	[+0.00010032, +0.00030995]	95%

X.5 Evaluation and provenance

Checkpoint bytes and tensor identities were verified locally. Both accounts' hash-bound attestations established each global cohort. REPORT followed AGT1, FRT1, ART1, then BCT1. The frozen evaluator retained BF16 CUDA, context 512, microbatch 16, 2,048 sequences and 1,048,576 targets at offset 2,500,000. Only its cohort-check callback was adapted for distributed verification. The centralized finalizer did not run unmodified.

A full-size inference gate checked independent-loop agreement, reload/repeat, and weight/RNG preservation. Native analyses and all 56 receipts were audited over 40 distinct checkpoint cells. Repeated linked scores and both accounts' final exports agree exactly.

Y Frozen-state observer identifiability

Y.1 Question and diagnostic states

OSC1's near-collinear residual directions (Appendix V) motivate asking where the agreement enters: the proposal representation, spectral transformation, or action on encountered proposals. The observer-identifiability probe (OIP1) uses four fixed states: the original phase-200 GCT1 prefixes and final phase-3,000 OSC1 paired-total endpoints, each for parents 159 and 160. The copying layout is stacked, with ancestry pairs $(0, 6), \ldots, (5, 11)$, not the adjacent layout in Appendices W–X.

For each fixed state, two independently sampled replicates use 24 pairs (48 effective batches) to fit moment matrices and eight further pairs (16 batches) to test their action. All four counterfactual conditions reuse each computed, clipped gradient. Fitting and action roles have separate draw identities, although their sampled token spans may overlap. The total is 512 effective-batch backward calls, with no committed model updates, optimizer transitions, weight decay, or validation evaluations.

Y.2 Separating inherited numerator memory and denominator anisotropy

The counterfactuals vary two features of the Adam proposal independently: inherited first-moment memory in the numerator and coordinate-dependent scaling in the denominator.

Let g be the current gradient, m, v the fixed inherited moments, and $k + 1$ the true next Adam age. Let $d(g)$ be the ordinary next-step coordinatewise denominator and $\bar{d}(g)$ its root-mean-square scalar replacement. Let $n_m(g)$ retain the inherited first moment and $n_0(g)$ omit that contribution from the numerator:

$$d(g) = \sqrt{\frac{\beta_2 v + (1-\beta_2) g^2}{1-\beta_2^{k+1}}} + \epsilon, \qquad \bar{d}(g) = \sqrt{\mathrm{mean}(d(g)^2)}, \tag{23}$$

$$n_m(g) = \frac{\beta_1 m + (1-\beta_1) g}{1-\beta_1^{k+1}}, \qquad n_0(g) = \frac{(1-\beta_1) g}{1-\beta_1^{k+1}}. \tag{24}$$

Here β_1, β_2 are the Adam moment-decay coefficients and ϵ its stabilizer. The scalar $\bar{d}$ pools all depths and coordinates of one projection, including ϵ. With learning rate η, the four hypothetical proposals are $-\eta n_m/d$, $-\eta n_0/d$, $-\eta n_m/\bar{d}$, and $-\eta n_0/\bar{d}$. All retain the same second-moment calculation and true-age bias correction. Only the numerator calculation removes inherited m. No stored moment is overwritten. The final-state probe holds the learning rate at its schedule endpoint. This positive scalar cancels from normalized geometry.

The counterfactual representation is used both to fit an operator and to probe its action. OIP1 does not isolate a changed observer acting on an unchanged native Adam proposal, and it does not compare momentum-reset training trajectories.

Y.3 Fitting and action measurements

Each proposal is centered, expressed in the same orthonormal depth basis, and normalized separately. For two such normalized proposals X, Y, let T denote their paired total moment, N the fluctuation contribution, and S the cross-proposal signal contribution:

$$\begin{aligned} T &= \tfrac{1}{2}(XX^\top + YY^\top), \qquad N = \tfrac{1}{2}(X-Y)(X-Y)^\top, \\ S &= \tfrac{1}{2}(XY^\top + YX^\top), \qquad T = S + N. \end{aligned} \tag{25}$$

The unweighted mean of 24 pair moments is fitted at each fixed state. This is not OSC1's historical moving average. No old paired-observer state is imported. Signal matrices remain signed until operator construction. All components use full cap three, including at phase 200, where this is deliberately a full-strength counterfactual rather than the inactive training-prefix operator.

Matrix-shape distance is $\|S/\|S\|_F - N/\|N\|_F\|_F$. Operator comparisons likewise remove positive scale. On separate action draws, the probe records Adam-orthogonal residual cosines, matching amplification, and matched-proposal displacement relative to the total correction. Undefined zero/rank cases are recorded, not replaced with arbitrary directions. Reported means average two sampling replicates and three FFN projections. These repeated measurements remain two training parents.

Y.4 Diagnostics across all frozen states

Table 54: **Observer-identifiability metrics for all four fixed states.** $c_{S,T}$ and $c_{N,T}$ compare Adam-orthogonal residual directions with the condition's own total direction. r_S is the matched signal/total proposal separation divided by the total correction norm, expressed as a percentage. $d_{S,N}$ is the normalized matrix-shape distance. Values average two sampling replicates and three projection groups.

Parent / phase	Numerator / denominator	$c_{S,T}$	$c_{N,T}$	r_S (%)	$d_{S,N}$
159 / 200	m/adaptive	0.999959	0.999266	0.904	0.019350
	0/adaptive	0.903530	0.998368	43.373	0.226022
	m/scalar	0.999981	0.999670	0.612	0.019303
	0/scalar	0.862360	0.987984	51.151	0.479408
160 / 200	m/adaptive	0.999925	0.998602	1.217	0.024408
	0/adaptive	0.925181	0.999195	38.369	0.200704
	m/scalar	0.999884	0.997724	1.488	0.048896
	0/scalar	0.915651	0.996987	40.503	0.325930
159 / 3000	m/adaptive	0.999990	0.999804	0.442	0.007787
	0/adaptive	0.775095	0.999958	66.827	0.352688
	m/scalar	0.999991	0.999828	0.429	0.013063
	0/scalar	0.891804	0.999917	45.912	0.356708
160 / 3000	m/adaptive	0.999992	0.999850	0.398	0.007042
	0/adaptive	0.772221	0.999915	67.277	0.379171
	m/scalar	0.999992	0.999845	0.397	0.014014
	0/scalar	0.814828	0.999455	59.217	0.643246

What changes. At phase 3,000, retaining the adaptive denominator but removing inherited first-moment memory changes the signal/total residual cosine from approximately 0.99999 to 0.77510 and 0.77222. The signal/noise normalized matrix distance rises from 0.00779/0.00704 to 0.35269/0.37917, so the separation is already present before spectral transformation. Mean signal-matched displacement rises from about 0.4% to 67% of the total correction. The phase-200 comparison has the same qualitative memory dependence. Replacing the denominator alone retains much closer signal/total directions.

Interpretation and scope. These fixed-state interventions identify inherited numerator memory as a major contributor to the tested shared geometry. They do not establish which representation yields better training or completely explain OSC1's evolving trajectories. Noise/total directions remain close under the memory-removed conditions. Moreover, final-state between-replicate signal-matrix distance under the adaptive denominator rises from about 0.00088/0.00069 to 0.13264/0.12752 after memory removal. Greater directional separation therefore accompanies a less stable fitted signal estimate in these samples. It is not automatically a better observer.

All fitted signal negative-mass fractions in the extracted diagnostics are zero, and all recorded matching directions are defined. Thus active PSD clipping does not explain these particular contrasts. The experiment changes both observer-input and action-probe representations, leaving an observer-only change with native Adam action as a separate, untested intervention.

Verification and data scope. The completed probe record reports exact input/protocol bindings, unchanged protected states, and 512 charged diagnostic backward calls with zero model updates and zero validation evaluations. The record audit checked the final analysis/SUCCESS binding, eight replicate results, eight fitting receipts, and 256 pair-receipt hashes. The accompanying artifact contains selected authenticated state-level metrics, not every raw action record. These summaries retain two training parents; repeated action draws do not increase that parent count or supply a population confidence interval.

Z Interpreting the transferred sector allocation

Z.1 Same-state measurements from completed trajectories

This retrospective analysis uses BCT1's retained diagnostics to characterize the allocation described in Section 5.4. It was not used to specify the prospective prediction. Each of four reused parents supplies two completed branches. At every retained phase, each branch computes the full-BDT and structured candidates from the same raw proposal and that branch's respective candidate histories. Comparing these two candidates isolates their allocation difference at that state; comparing the two branches' applied updates would also mix in trajectory differences.

Records cover phases 201, 210, 250, 300, 310, 400, 420, 1,000, 2,000, and 3,000 for gate, up, and down projections: 240 parent/branch/phase/projection records. Repeated phases, projections, and the second branch are not independent training units. Every plot retains all ten sampled phases; connecting segments are visual guides, not measurements between them.

Let $\mathcal{B}$ and $\mathcal{S}$ denote the full-BDT and structured candidates at the same state. Their logged sector norms are b_A^X, b_D^X for $X \in \{\mathcal{B}, \mathcal{S}\}$. The difference-sector energy fraction q_D^X is

$$q_D^X = \frac{(b_D^X)^2}{(b_A^X)^2 + (b_D^X)^2}. \tag{26}$$

This is a fraction of *centered proposal energy*, not a fraction of all parameters, tokens, or total update energy including U_m. The two candidates preserve the same pooled centered norm at a given state, so their fractions directly describe its allocation between the two ancestry sectors.

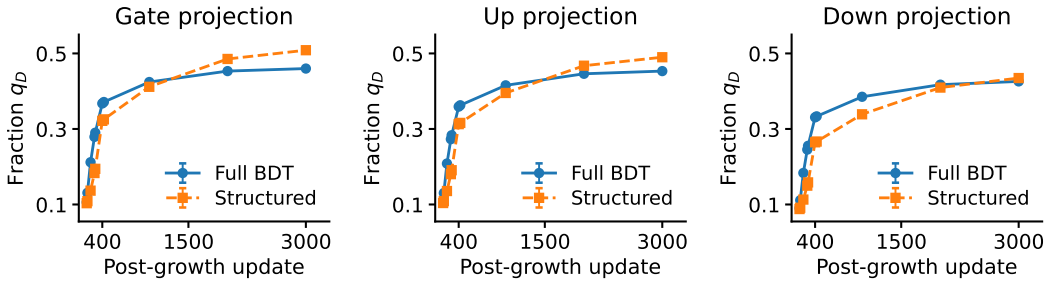

Figure 12: **Candidate allocations on full-budget branch states.** The same comparison as main Figure 4. Points are parent means; bars show the range of the four individual parent values.

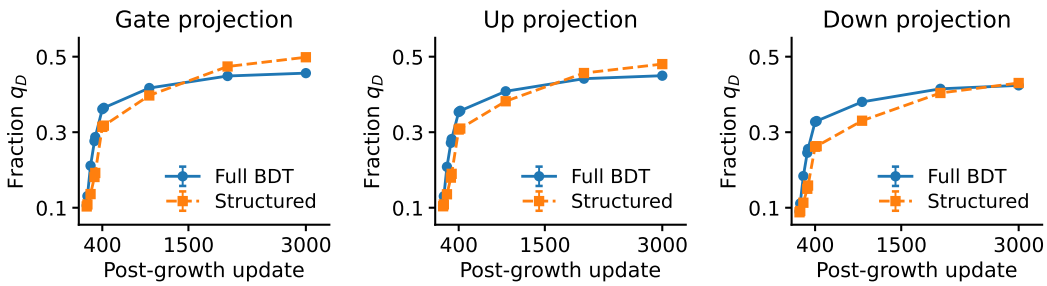

Figure 13: **Candidate allocations on structured-budget branch states.** Both candidates are recomputed along the structured-budget branches. The ordering is again full above structured at phase 400 and below at phase 3,000 for all twelve parent/projection combinations. Bars are parent ranges, not confidence intervals.

The same endpoint ordering reversal appears on each branch-state family, rather than only after choosing the more favorable trajectory. Gate and up already have lower full-BDT fractions by phase 2,000, while down remains higher at that sampled phase. The difference therefore depends on projection as well as training phase.

Z.2 Effective amplification relative to the raw proposal

The sector fraction depends on both the incoming proposal and the applied rule. A norm-equivalent amplification coefficient expresses the rule's relative weighting on the same incoming proposal.

For raw proposal U, let $a = \|P_A U\|_F$ and $d = \|P_D U\|_F$ be the pair-average and pair-difference norms. Their raw difference-energy fraction is $q_{\text{raw}} = d^2/(a^2 + d^2)$. Applying coefficient ρ to the difference sector and restoring pooled norm produces fraction q_ρ:

$$q_\rho = \frac{\rho^2 q_{\text{raw}}}{1 - q_{\text{raw}} + \rho^2 q_{\text{raw}}}. \tag{27}$$

Consequently, a changing q_D can arise even with fixed ρ. The structured construction conditions directions within sectors without changing their relative allocation, so it obeys this relation with $\rho_{\mathcal{S}}(t) = 2^{a_t}$. Let $\widehat{q}_{\text{raw}}$ be the raw fraction inferred from the structured candidate, and let $\rho^{\mathcal{B}}_{\text{eff}}$ be the scalar amplification reproducing BDT's sector-norm ratio on that same proposal:

$$\widehat{q}_{\text{raw}} = \frac{q_D^{\mathcal{S}}}{\rho_{\mathcal{S}}^2(1 - q_D^{\mathcal{S}}) + q_D^{\mathcal{S}}}, \qquad \rho^{\mathcal{B}}_{\text{eff}} = \rho_{\mathcal{S}} \sqrt{\frac{q_D^{\mathcal{B}}(1 - q_D^{\mathcal{S}})}{(1 - q_D^{\mathcal{B}}) q_D^{\mathcal{S}}}}. \tag{28}$$

The second expression is equivalently $(b_D^{\mathcal{B}}/b_A^{\mathcal{B}})/(d/a)$. It is the scalar that matches *sector norms* on the same raw proposal, not BDT's within-sector directions. All retained sectors are nonzero; outside that case an undefined ratio is left undefined rather than assigned a fallback.

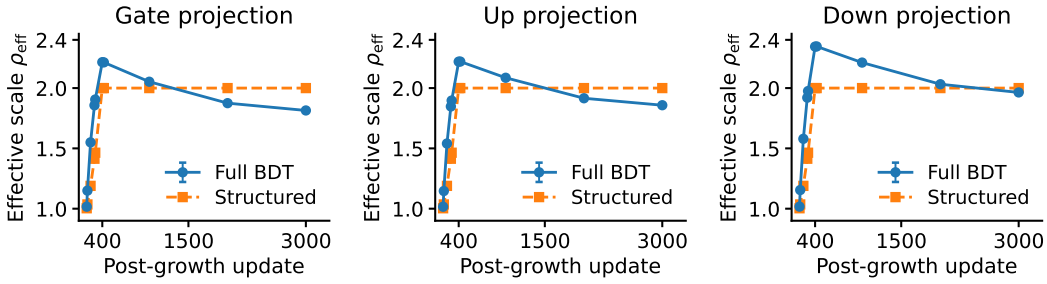

Figure 14: **Effective amplification on full-budget branch states.** Parent-wise ratios are computed before averaging. The structured coefficient reaches two at phase 400 and stays there. The full-BDT norm-equivalent coefficient ends lower, with projection-dependent trajectories. Bars show parent ranges.

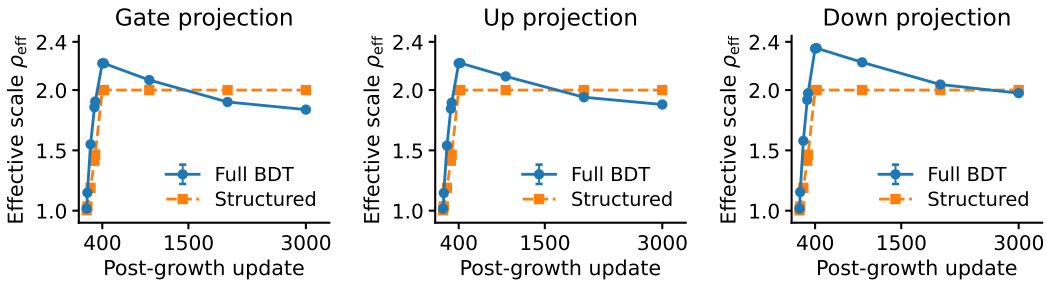

Figure 15: **Effective amplification on structured-budget branch states.** The same qualitative profile appears when both candidates are evaluated along the other branch family. The estimate is algebraically inferred from recorded sector norms. It is not an additional training treatment.

A single constant cannot reproduce all these retained sector ratios on the same inputs. This does not determine which allocation dynamics cause the loss advantage: a different fixed coefficient or a predetermined phase/projection schedule could still obtain similar quality along its own trajectory. No such replacement is evaluated here.

Z.3 Representative values and result context

Table 55 gives means over parents, including the raw fraction inferred from the structured rule. On the full-budget branch states, the mean effective coefficient decreases from 2.214890 to 1.814461 for gate and from 2.343877 to 1.963944 for down between phases 400 and 3,000. Yet the corresponding full-BDT difference fractions increase. Lower *relative amplification* late therefore does not mean a decreasing difference fraction, nor suppression below the raw allocation.

Table 55: **Allocation at the full-ramp and final phases.** Four-parent means, computed from each parent's own-state comparison. $\widehat{q}_{\text{raw}}$ is inferred, not independently sampled. ρ_{eff} is averaged after forming each parent-wise ratio. All 240 records are supplied in the accompanying CSV; six decimals here are display precision.

Phase	Projection	$q_D^{\mathcal{B}}$	$q_D^{\mathcal{S}}$	$\widehat{q}_{\text{raw}}$	$\rho_{\text{eff}}^{\mathcal{B}}$
Full-budget branch states					
400	Gate	0.367547	0.321510	0.105920	2.214890
	Up	0.359106	0.312742	0.102148	2.219334
	Down	0.330796	0.264663	0.082555	2.343877
3000	Gate	0.459944	0.508537	0.205524	1.814461
	Up	0.453228	0.489932	0.193636	1.857940
	Down	0.425881	0.434802	0.161302	1.963944
Structured-budget branch states					
400	Gate	0.361036	0.314066	0.102712	2.221796
	Up	0.353493	0.307060	0.099737	2.221666
	Down	0.327391	0.261546	0.081345	2.344658
3000	Gate	0.456430	0.498530	0.199065	1.838092
	Up	0.449642	0.480356	0.187720	1.880240
	Down	0.424016	0.430275	0.158823	1.974593

Z.4 PROVENANCE AND VERIFICATION

The recorded result hashes agree with the existing BCT1 local exports; result, contract, and checkpoint metadata links and the original source identity were checked. Fractions were recomputed from the saved sector norms and cross-checked against the independently logged structured target and assigned-source fraction. No new gradient, model evaluation, checkpoint-tensor replay, or access to incomplete-study efficacy was performed for this analysis.

The source package includes a checksum-bound nine-decimal plotting extract, all parent-level derived values, an arithmetic checker, and the saved-snapshot audit. Maximum plotting-rounding error is 5.01×10^{-10} or less. The original snapshot and full-precision extract hashes remain recorded separately. The packaged plotting extract is not presented as a fresh tensor-level verification.

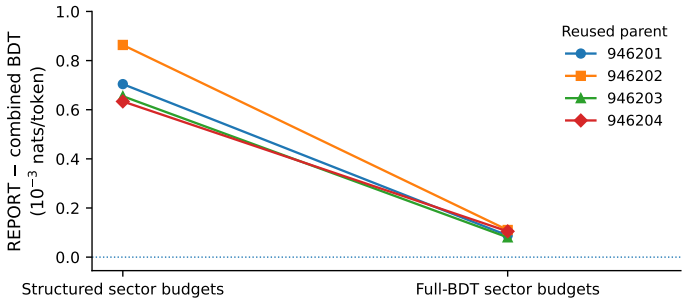

Figure 16: **Prospective outcome versus retrospective profile.** Each line connects one parent's two BCT1 REPORT losses relative to combined BDT. Full budgets win on all four parents, with a residual gap to combined BDT. Appendix X gives all scores.

AA Fresh-parent confirmation of the allocation ordering

AA.1 Independent cohort and fixed comparison

The fresh-parent confirmation (internal study ID BCT2) repeats the budget-source comparison from Appendix X on four independently initialized and pretrained parents, 946301–946304. Each parent has its own seed-specific pretraining stream. The architecture, token cache, copying procedure, optimizer recipe, smoothing construction, and endpoint remain unchanged. There is no parent selection, replacement, or outcome-dependent sample-size extension.

The comparison uses the two hybrid rules in Equation (11). Both obtain their within-sector directions from the same window-smoothed AdamW construction and use the same attention treatment. They differ in whether the sector magnitudes come from the full-BDT or structured candidate at their own evolving states. The six-layer parent trains for 7,630 updates before adjacent six-to-twelve-layer copying. A common 200-update post-growth prefix precedes the two continuations to phase 3,000, following the recipe in Appendix X.

The registered primary is final reporting loss for the full-BDT-budget hybrid minus that for the structured-budget hybrid, paired within each new parent. Negative values favor full-BDT budgets. The four parent-level differences determine a two-sided 95% Student-t interval with three degrees of freedom. The primary criterion requires a negative upper endpoint and nonzero parent-to-parent variance. The four development families from BCT1 do not enter this analysis. The two studies establish distinct stages of evidence and are not pooled into an eight-parent confirmation.

AA.2 Complete outcomes

Both continuations complete on every parent. All eight endpoints are included in the final report. Every parent-level difference favors full-BDT budgets (Table 56).

Table 56: **Fresh-parent allocation confirmation: all eight final scores.** Final reporting loss in nats/token. The last column is the sole registered primary contrast within each parent. Values are rounded to nine decimal places for display.

Parent	Full-BDT budgets	Structured budgets	Full minus structured
946301	3.623607341	3.624165580	−0.000558238
946302	3.624475598	3.625074221	−0.000598622
946303	3.626926653	3.627531044	−0.000604391
946304	3.621706935	3.622321416	−0.000614481

The mean difference is -0.000593933277, with standard error 0.000012341273 and paired 95% interval $[-0.000633208717, -0.000554657837]$. The prespecified criterion is met. The original prospective test has mean -0.00061813 on its four development families (Appendix X). The concordant ordering across the two studies extends the finding to new initializations and pretraining streams within the fixed setting. The corpus, architecture, and hybrid comparison remain unchanged.

AA.3 Completion and verification scope

The good-lane completion record is timestamped September 25, 2026, 09:08:58 UTC. The supplied verification summary records both lane closures and a global seal of all eight endpoints before final reporting. Its earlier live audit checked the eight score-receipt hashes and their checkpoint, runtime, and evaluation bindings.

This revision independently verifies the supplied 2,506-byte `FINAL_REPORT.json` and recomputes its paired differences, mean, standard error, and interval. The report's SHA-256 is

`529b846d764a05147859ab47cd210d07075934246d3ddd9bf15cca2e341a3fd9`

The report retains the seal, protocol, runner, native-source, and receipt hash bindings. This is record-level verification. Checkpoint payloads and full execution sources were not reloaded, and no training or evaluation was rerun. The report and a standalone recomputation script accompany this revision.