SAME FUNCTION, DIFFERENT DYNAMICS: UNDERSTANDING MUON'S RESPONSE TO REPARAMETERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how function-preserving changes in parameter representation affect optimizer updates and subsequent predictions. Focusing on Muon in SwiGLU language models, we develop Libra, a state-resolved framework that connects reciprocal channel scaling to local predictive response. Under consistent optimizer-state transport, update homogeneity explains the response to uniform scaling. After scale compensation, exact-polar derivatives separate column-space motion from singular-frame rotation, while finite Newton–Schulz updates also change spectral amplitudes. We decompose the difference between these derivatives. To connect update changes to predictions, we propagate the original-rule update defect through the full parameter–optimizer state and use Jacobian/Fisher readouts to estimate predictive changes. Experiments on 106.57M-parameter language models find that off-diagonal response contributes 65.3–82.1% of the derivative-difference energy between five-step Newton–Schulz (NS5) and exact-polar updates. This range spans checkpoint groups in the tested random directions. In the larger-dose exploratory cohort, observable projection lowers median relative Kullback–Leibler (KL) error compared with calibrated scalar defect proxies for one-step responses with transported optimizer state. Measurements with retained optimizer state further reveal how memory and update responses reinforce or cancel each other. These measurements characterize local, short-horizon predictive responses under the specified optimizer-history conventions.

1 INTRODUCTION

Different parameter settings can represent the same network function yet lead to different subsequent optimization trajectories. Understanding this dependence matters for optimizer design: rescaling-invariant updates and function-preserving weight balancing seek to control how parameter representation affects learning (Neyshabur et al., 2015; Saul, 2023). Symmetry teleportation uses this freedom to improve optimization (Zhao et al., 2022; 2024). Function-preserving interventions isolate representation dependence by holding the initial predictions fixed. Yet a response to uniform scaling may reflect only update homogeneity shared by different optimizers. Moreover, equal-magnitude update defects can produce different predictive changes. Endpoint KL measures the resulting change without resolving its relation to update direction or optimizer history. A useful diagnostic must distinguish these effects and connect them to predictions.

We develop Libra for this purpose: a state-resolved framework that links representation-dependent update mechanisms to local predictive changes under paired interventions. We focus on Muon in SwiGLU language models. Muon applies a polar or approximate-polar map to a momentum-derived matrix (Jordan et al., 2024; Liu et al., 2025). Reciprocal scaling of the SwiGLU up/down channels changes their norms without changing the block's function (Shirodkar, 2026). Together, these properties let us study how both the idealized update map and its finite Newton–Schulz implementation respond to function-preserving scaling. At a checkpoint with accumulated optimizer history, the intervention must also specify whether that history is retained or transported consistently with the transformed gradients. Following earlier comparisons of momentum conventions (Zhao et al., 2024, Appendix F.1), we make that choice explicit in the update and training responses.

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.

Our analysis first separates shared scale behavior from the heterogeneous matrix response to nonuniform channel scaling. With consistent state transport, Muon's uniform-scale defect follows degree-zero homogeneity, as does Frobenius-normalized momentum. Uniform scaling alone therefore cannot identify a polar-specific mechanism. Learning-rate compensation removes this scale effect. For full-rank inputs, classical polar differentiation (Gawlik & Leok, 2016) then separates the heterogeneous response into column-space motion and singular-frame rotation. Finite Newton–Schulz updates can additionally change spectral amplitudes and reweight the geometric terms. We decompose the derivative difference between exact-polar and finite Newton–Schulz updates to show why matching update norms and uniform-scale behavior does not determine the heterogeneous response.

Even after these matrix mechanisms are separated, non-equivariance alone does not determine the predictive response. Identical commutators can accompany different orders of function change, and optimizer memory can affect later weights while leaving current logits unchanged. To connect update changes to predictions, Libra first relates the compensated derivative to the original-rule update defect. It then adapts tangent sensitivity (Franceschi et al., 2017) to propagate the full parameter–optimizer state response along a fixed control trajectory, with matched checkpoints, data orders, and random streams. This calculation separates the effect of the initial state mismatch from that of update non-equivariance accumulated along the trajectory. Projection through the network Jacobian then yields local logit changes. Predictive Fisher and loss-gradient readouts translate these changes into KL and signed-loss responses, respectively. These readouts retain direction information omitted by defect norms and reveal how memory and update responses reinforce or cancel each other.

Experiments on 106.57M-parameter language models test both the update decomposition and its connection to predictive response. Off-diagonal response accounts for 65.3–82.1% of NS5–polar derivative-difference energy across checkpoint groups in the tested random directions. Under an independent protocol fixed before execution, evaluation on three held-out training seeds gives raw median relative KL errors of 0.07365 at one step and 0.05728 at four steps with transported optimizer state. With matched scalar baselines and resampled probes, projection lowers mean median relative KL error for transported one-step responses across all tested training-seed–budget combinations, although absolute-error results are mixed. In the separate larger-dose exploratory cohort, transported one-step median relative KL error falls from 1.53 for a calibrated commutator proxy to 0.428 for projection. The cohort's retrospective split and distinct metrics are specified in Appendix E.4. Comparisons across history conventions, horizons, and arithmetic precision establish the state dependence and reliable local range of these responses.

Our contributions are twofold. First, a scale-compensated analysis distinguishes shared homogeneity from the heterogeneous mechanisms of exact-polar and finite Newton–Schulz updates, decomposing their derivative difference into normal, amplitude, and off-diagonal components. Second, a state-resolved intervention framework connects these mechanisms to local logit, KL, and signed-loss responses through the original-rule defect, full-state propagation, and observable projection. The framework makes both the history convention and control trajectory explicit. Matrix studies, independent multi-seed evaluation, and probe and numerical checks provide evidence for these contributions.

2 Preliminaries and Problem Setup

Function-preserving scaling. For a bias-free SwiGLU block,

$$f_{\mathrm{MLP}}(x) = W_D[\mathrm{SiLU}(W_G x) \odot (W_U x)], \quad G_s(W_U, W_D) = (\Lambda W_U, W_D \Lambda^{-1}), \tag{1}$$

where $\Lambda = \mathrm{diag}(e^{s_i})$. The factors cancel for every input; the gate is unchanged. Gradients transform as $g'_U = \Lambda^{-1} g_U$ and $g'_D = g_D \Lambda$. This symmetry is described in DDC (Shirodkar, 2026). The norm-ratio coordinate $\theta_i = \log \|W_{U,i:}\|_2 - \log \|W_{D,:i}\|_2$ changes by $2s_i$.

Optimizer-state conventions. Let $S = (W, h)$ denote the parameters and all continuous optimizer buffers. The transported action ρ_s applies G_s to weights and the corresponding covector map to gradient-like state. For Muon momentum,

$$B'_U = \Lambda^{-1} B_U, \qquad B'_D = B_D \Lambda. \tag{2}$$

Adam first moments receive these factors and second moments their squares. A transported pulse is $I_s^{\mathrm{cov}} S = \rho_s S$; a retained-state pulse is $I_s^{\mathrm{ret}}(W, h) = (G_s W, h)$. Both preserve the current function,

but generally lie on different extended-state orbits. Resetting buffers defines a third treatment. We compare retention and transport under the same continuation.

Paired intervention response. We condition on step maps Φ_t with a fixed data sequence, random stream, schedule, precision, and preprocessing. For an observable Y, the paired estimand is

$$\Delta_{s,T}^{I,\Phi} Y = Y(\Phi_{T-1} \circ \cdots \circ \Phi_0(I_s S_0)) - Y(\Phi_{T-1} \circ \cdots \circ \Phi_0(I_0 S_0)). \tag{3}$$

This conditional response fixes both the intervention and the implemented continuation. Normalization floors, skipped zero-norm channels, clipped doses, and gradient preprocessing belong to Φ_t. Transport makes the input state representation-consistent; the update map determines whether subsequent dynamics preserve that relation.

3 MUON: SEPARATING UPDATE SCALE FROM MATRIX SHAPE

Muon applies a polar or approximate-polar map to its momentum-derived matrix (Jordan et al., 2024; Liu et al., 2025). We use the recurrence

$$B^+ = \beta B + (1-\beta) g, \qquad Z = (1-\beta) g + \beta B^+. \tag{4}$$

Without noncovariant preprocessing, transported state gives $Z'_U = \Lambda^{-1} Z_U$ and $Z'_D = Z_D \Lambda$. The update map must still convert those covectors into parameter increments.

3.1 CALIBRATING THE UNIFORM-SCALE RESPONSE

Proposition 3.1 (Homogeneous update and compensating rate)**.** *Suppose $\mathcal{A}(aZ) = a^h \mathcal{A}(Z)$ for $a > 0$ and $\mathcal{A}(Z) \neq 0$. Under uniform scaling $c = e^s$, the symmetric per-side commutator is*

$$\kappa_h(s) = \frac{2 \left\| \mathcal{A}(c^{-1}Z) - c\mathcal{A}(Z) \right\|_F}{\left\| \mathcal{A}(c^{-1}Z) \right\|_F + c \left\| \mathcal{A}(Z) \right\|_F} = 2 \tanh\left(\frac{|(h+1)s|}{2} \right). \tag{5}$$

It is canceled by $\eta'_U = c^{h+1} \eta_U$ and $\eta'_D = c^{-(h+1)} \eta_D$.

Exact polar and Frobenius-normalized momentum both have $h = 0$; ordinary gradient/momentum updates have $h = 1$. With covariant recurrences, an invariant objective, consistently rescaled schedules, and exact arithmetic, the compensated trajectories remain related by G_s at every step. Decay must preserve its actual multiplier; global clipping, fixed ϵ, and normalization floors can invalidate the assumptions. Compensation defines a calibration control with a modified learning rule. Proofs and the Adam-specific extension are in Appendix A.

3.2 THE COMPENSATED HETEROGENEOUS POLAR RESPONSE

After removing the uniform-scale effect, we examine nonuniform channel scaling. For full-column-rank $Z \in \mathbb{R}^{m \times n}$, $m \geq n$, write $Z = QH$, $Q^\top Q = I$, $H \succ 0$. Let $\mathcal{P}(Z) = Q$ and q be a channel direction. For $D = \operatorname{diag}(q)$ set

$$L_D(Z) = \frac{d}{ds} \mathcal{P}(e^{-sD} Z) \Big|_{s=0}, \qquad C = Q^\top D Q. \tag{6}$$

Up-side row-rate compensation left-multiplies the update by ηe^{sD}. Pullback by e^{-sD} then gives $\eta \mathcal{P}(e^{-sD} Z)$, whose residual derivative is $\eta L_D(Z)$.

Theorem 3.2 (Scale-compensated spectral decomposition)**.** *Let $H = V \operatorname{diag}(\sigma_i) V^\top$ and $\widetilde{C} = V^\top C V$. Then*

$$L_D = -(I - QQ^\top) DQ + Q\Omega, \tag{7}$$

$$H\Omega + \Omega H = HC - CH, \qquad \Omega^\top = -\Omega, \tag{8}$$

$$\|L_D\|_F^2 = \underbrace{\left\| (I - QQ^\top) DQ \right\|_F^2}_{\textit{column-space motion}} + \underbrace{\sum_{i,j} \left(\frac{\sigma_i - \sigma_j}{\sigma_i + \sigma_j} \right)^2 |\widetilde{C}_{ij}|^2}_{\textit{singular-frame rotation}}. \tag{9}$$

The residual is zero exactly when $(I - QQ^\top)DQ = 0$ and $HC = CH$. Moreover,

$$\|L_D\|_F \le \|DQ\|_F, \quad L_{D+aI}(Z) = L_D(Z), \quad L_D(\alpha Z) = L_D(Z) \quad (\alpha > 0). \tag{10}$$

For a wide down projection, transpose the matrix and reverse the sign of D; the energy is unchanged.

The derivative specializes classical polar calculus (Gawlik & Leok, 2016). Uniform scaling leaves no residual after compensation. Equal singular values remove the rotation term, but a tall matrix can retain column-space motion. The bounded spectral ratios characterize this full-rank multiplicative perturbation. A five-step Newton–Schulz (NS5) map has a different spectral response, derived next.

3.3 Finite Newton–Schulz response beyond scale

Write the singular value decomposition as $Z = U \operatorname{diag}(\sigma_i) V^\top$. Above the normalization floor, the five-step polynomial map has output $\mathcal{A}_5(Z) = U \operatorname{diag}(y_i) V^\top$, where $y_i = \gamma p_5(x_i)$ and $x_i = \sigma_i / \|Z\|_F$. Here p_5 is the polynomial after five NS iterations, and γ is a fixed update-scale factor (Jordan et al., 2024). With $C = U^\top DU$ and $\mu = \sum_j x_j^2 C_{jj}$, the path $Z(s) = e^{-sD} Z$ gives

$$\dot{y}_i = -\gamma x_i p_5'(x_i)(C_{ii} - \mu). \tag{11}$$

Uniform scaling makes this amplitude response zero. Heterogeneous scaling can change both the output amplitudes and the singular vectors, whereas exact polar has constant output singular values.

For positive, distinct singular values, write the finite-NS derivative as $(N + UB)V^\top$: N is normal to the column space, $B_{ii} = \dot{y}_i$, and the off-diagonal entries of B describe the vector response. Freeze $\alpha = \|\mathcal{A}_5(Z)\|_F / \|Q\|_F$ at zero dose, with $Q = UV^\top$, and denote the matched-polar components by $N^{\mathrm{P}}, B^{\mathrm{P}}$, where $B^{\mathrm{P}}_{ii} = 0$. Then

$$\Delta \dot{A} = D\mathcal{A}_5(Z)[-DZ] - \alpha L_D = [\Delta N + U(B - B^{\mathrm{P}})]V^\top, \tag{12}$$

$$\left\| \Delta \dot{A} \right\|_F^2 = \|\Delta N\|_F^2 + \sum_i |\dot{y}_i|^2 + \sum_{i \ne j} |B_{ij} - B^{\mathrm{P}}_{ij}|^2, \tag{13}$$

where $\Delta N = N - N^{\mathrm{P}}$. These three terms measure normal, amplitude, and off-diagonal contributions to the energy of the derivative difference. Appendix B.1 gives their entries and the normalizer correction. At repeated singular values we use invariant within-block energies; an SVD-free matrix derivative remains well defined at repeated or zero singular values on a fixed normalization branch (Appendix B.2). Section 5.2 measures this decomposition on checkpoint matrices.

4 Libra: State-Resolved Predictive Response

Given a checkpoint with saved optimizer state, a channel direction, a state-handling convention, a fixed control continuation, and a probe set, Libra builds the update response, propagates the full parameter–state tangent, and returns local logit, KL, and signed-loss responses. The matrix analysis explains the update mechanism; the following recurrence carries it through the specified training trajectory.

4.1 Initial state mismatch and ongoing non-equivariance

For a fixed channel direction q and scalar dose s, let K_q be the generator of full-state transport. Pull the branch back using

$$\Phi_t^{(s)} = \rho_{-sq} \circ \Phi_t \circ \rho_{sq}, \qquad \rho_{sq} = e^{sK_q}. \tag{14}$$

For the actual branch $\widehat{S}_t(s)$, define the pulled-back full state $\widetilde{S}_t(s) = \rho_{-sq} \widehat{S}_t(s)$ and its dose derivative $r_t = \partial_s \widetilde{S}_t(s)\Big|_{s=0}$, including weights and optimizer buffers. The transported branch starts at S_0 in these coordinates; the retained branch starts at $\rho_{-sq} I^{\mathrm{ret}}_{sq} S_0$.

Theorem 4.1 (State-resolved finite-horizon response)**.** *Assume the conditioned maps are continuously differentiable near the fixed finite control trajectory $S_{t+1} = \Phi_t(S_t)$. With $A_t = D\Phi_t(S_t)$ and*

$b_t(q) = A_t K_q S_t - K_q S_{t+1}$,

$$r_{t+1} = A_t r_t + b_t(q), \tag{15}$$

$$r_T = \underbrace{A_{T-1} \cdots A_0 r_0}_{\textit{initial state mismatch}} + \underbrace{\sum_{k=0}^{T-1} A_{T-1} \cdots A_{k+1} b_k(q)}_{\textit{ongoing non-equivariance}}. \tag{16}$$

Here $r_0 = 0$ for transport and $r_0 = (K_q^{\mathrm{ret}} - K_q)S_0$ for retention, where K_q^{ret} acts only on weights. The first-order response of a smooth gauge-invariant f is $Df(S_T)r_T$.

The state convention determines the initial term, and the weight–buffer blocks of A_t allow optimizer history to affect later predictions. The recurrence specializes tangent sensitivity (Franceschi et al., 2017); its forcing sum telescopes to the chain-rule response (Appendix A.8). Differentiability is required along the specified control trajectory, with implementation branch conventions given in Appendix E.7.

4.2 Connecting the compensated derivative to the implemented rule

To use the matrix derivative in the state recurrence, restore the increment's transformation under the original learning rule. Let $u(S)$ include the full parameter increment and its learning rates, so $W^+ = W - u(S)$, and define its pulled-back defect

$$\delta_s = G_{-sq} u(\rho_{sq} S) - u(S). \tag{17}$$

Proposition 4.2 (From shape response to state forcing)**.** *For a transformed up/down pair without additional noncovariant preprocessing, let $\mathcal{A}_U, \mathcal{A}_D$ include fixed matrix-size and norm-matching multipliers. Define*

$$\mathcal{L}_U = \partial_s \mathcal{A}_U(e^{-sD} Z_U)\big|_0, \qquad \mathcal{L}_D^{\mathrm{down}} = \partial_s \mathcal{A}_D(Z_D e^{sD})\big|_0. \tag{18}$$

For the original scalar rates η_U, η_D, the pulled-back increment defects satisfy

$$\dot{\delta}_U = \eta_U[\mathcal{L}_U - D\mathcal{A}_U(Z_U)], \qquad \dot{\delta}_D = \eta_D[\mathcal{L}_D^{\mathrm{down}} + \mathcal{A}_D(Z_D)D]. \tag{19}$$

With up-side row-rate compensation $\eta_U e^{sD}$ and down-side column-rate compensation $\eta_D e^{-sD}$, they instead satisfy

$$\dot{\delta}_U^{\mathrm{comp}} = \eta_U \mathcal{L}_U, \qquad \dot{\delta}_D^{\mathrm{comp}} = \eta_D \mathcal{L}_D^{\mathrm{down}}, \qquad b_t^W = -\dot{\delta}_t \quad \textit{(original rule)}. \tag{20}$$

The last equality concatenates original-rule defects over parameter blocks and their actual rates. Compensation instead gives weight forcing $-\dot{\delta}_t^{\mathrm{comp}}$, with the dose-dependent rate derivative included.

The down-side derivative applies the tall-matrix formula to $Z_D^\top$ with $-D$ and transposes back. Differentiating the outer pullback factors gives Eq. (19), while compensation cancels those factors. This restores the original-rule forcing from the isolated shape response. Additional preprocessing and buffer updates are differentiated within Φ_t; even a block with zero direct forcing can respond later through A_t.

A direct full-state derivative route. Equivalently, express the tangent in the original state coordinates as $\xi_t = r_t + K_q S_t$. Theorem 4.1 gives

$$\xi_{t+1} = A_t \xi_t, \qquad \xi_0 = r_0 + K_q S_0, \qquad Df(S_T) r_T = Df(S_T)\xi_T. \tag{21}$$

For transport $\xi_0 = K_q S_0$; for retention $\xi_0 = K_q^{\mathrm{ret}} S_0$. A Jacobian–vector product (JVP) computed by automatic differentiation (AD) along each control step therefore computes the same response without explicitly accumulating forcing. We compare the two routes with centered finite differences (FD) of the actual intervened paths, varying derivative step h separately from dose s (Appendix E.5).

4.3 LOGIT, KL, AND SIGNED-LOSS READOUTS

The propagated tangent becomes a predictive response only after projection through the network. If z denotes the probe logits, their local change along q is $sDz(S_T)r_T + O(s^2)$. For vector dose s, let $\mathcal{R}_T(x)$ map dose to first-order logits at position x in the probe set $\mathcal{X}$, and let $p_T(x)$ and $p_T^{(s)}(x)$ be the control and intervened softmax probabilities. A twice continuously differentiable continued-logit family with locally bounded derivatives gives

$$M_T = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathcal{R}_T(x)^\top F_x \mathcal{R}_T(x), \tag{22}$$

$$F_x = \operatorname{diag}(p_T(x)) - p_T(x) p_T(x)^\top,$$

$$\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathrm{KL}(p_T(x) \| p_T^{(s)}(x)) = \tfrac{1}{2} s^\top M_T s + O(\|s\|^3). \tag{23}$$

For scalar dose along q and smooth probe loss ℓ, the signed readout is $\Delta\ell = sD\ell(S_T)r_T + O(s^2)$ and uses the loss gradient independently of KL calibration.

Projection supplies direction information absent from a defect norm. In the transported one-step case, the local KL is $\frac{1}{2}\delta_s^\top H_z \delta_s$, where $H_z = \mathbb{E}_x[J_z^\top F_x J_z]$ and J_z is the logit Jacobian at the control's updated weights. The direction-dependent gain $\delta_s^\top H_z \delta_s / \|\delta_s\|^2$ can vary at fixed norm. Indeed, the scalar model $f(u, d) = ud$ has the same commutator for balanced and imbalanced factors, yet its function change is respectively $O(s^2)$ and generically $O(s)$ (Proposition A.2). Appendix A.6 gives the one-step remainder bound.

Projection also changes how components combine. Define $\langle v_i, v_j \rangle_F = \mathbb{E}_x[v_i(x)^\top F_x v_j(x)]$ for logit tangents propagated from separate state or matrix components. Then

$$\left\| \sum_j v_j \right\|_F^2 = \sum_j \|v_j\|_F^2 + 2 \sum_{i<j} \langle v_i, v_j \rangle_F. \tag{24}$$

Initial-memory and forcing responses can therefore reinforce or cancel. The Frobenius-orthogonal components in Eq. (13) need not remain orthogonal after state propagation and the logit Jacobian, so their matrix-energy shares are not shares of predictive KL.

5 EXPERIMENTS

The experiments follow the connection from heterogeneous update mechanisms to predictive change. We first identify the NS5–polar difference, then ask how observable projection reweights its components, whether the local response predicts finite interventions, and what information direction and optimizer history add.

5.1 EXPERIMENTAL SETUP

The main evaluations use 106.57M-parameter Muon checkpoints with saved optimizer history. Paired continuations share data, random streams, and arithmetic. Mechanism and local-response evaluations use FP64; numerical controls also cover FP32, BF16, and FP16. We measure full-vocabulary KL and signed next-token loss on fixed probes, with training seeds as the replication unit. The independent confirmation fixes its calibration/evaluation split before execution; the larger-dose exploratory cohort uses a retrospective split and distinct error metrics. Appendix E specifies both protocols, metric definitions, and reporting conventions.

The core experiments use a bias-free 106,568,640-parameter decoder with ten layers, width 576, SwiGLU width 2048, twelve attention heads, GPT-2 vocabulary size 50257, and zero dropout. The data are the fixed GPT-2-tokenized FineWeb-Edu 128M-token shard, with batch size eight and sequence length 256. Checkpoints preserve optimizer history and the data cursor. Seeds 200–203 have 500-step training prefixes using BF16 autocast and FP32 state. Diagnostic runs use FP64 state and NS5, high-precision RMSNorm, deterministic math attention, and disable TF32.

Muon updates 48,660,480 transformer-block matrix parameters, treating Q/K/V as separate logical matrices. Its base learning rate is 0.02 with the original rectangular-matrix scale, momentum is 0.95, and the Nesterov rule is unchanged. The remaining 57,908,160 embedding, LM-head and gain parameters use auxiliary Adam at rate 0.0003, $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. Diagnostics add no weight decay or clipping. NS5 uses five steps with $(a, b, c) = (3.4445, -4.775, 2.0315)$ and Frobenius normalizer $\max(\|Z\|_F, 10^{-7})$. Runs use NVIDIA H200 hardware, Python 3.10.12 and PyTorch 2.6.0+cu124.

5.2 Heterogeneous scaling reveals finite-NS mechanisms

Uniform-scale controls establish why heterogeneous scaling is needed. With zero-dose matrix update norms matched and matching factors frozen, four update rules on three checkpoints obey the shared homogeneity law: at $|s| = 0.5$, commutators agree with its predictions within 5.33×10^{-15}. Rate compensation reduces the 16-step exact-polar and FP64 NS5 KLs to the numerical floor (Table 6). Removing this common scale effect isolates the heterogeneous shape response.

Table 1: NS5–polar derivative difference and component energy shares (mean $\pm$ sample SD across three training seeds). Each seed uses the median relative difference over 640 matrix–direction cases; shares divide summed component energy by summed difference energy. Matching α is fixed at zero dose.

Step	Rel. difference	Normal	Amplitude	Off-diagonal
500	0.602 ± 0.027	0.19 ± 0.003	0.0951 ± 0.011	0.714 ± 0.012
4500	0.606 ± 0.0073	0.0981 ± 0.0054	0.1 ± 0.016	0.802 ± 0.016
16500	0.758 ± 0.05	0.155 ± 0.017	0.148 ± 0.027	0.696 ± 0.042
50500	0.897 ± 0.066	0.198 ± 0.008	0.137 ± 0.00041	0.664 ± 0.0082

Off-diagonal response dominates the NS5–polar derivative difference in the tested random directions (Table 1). Across seeds 0–2 and four checkpoint stages, it accounts for 65.3–82.1% of summed difference energy, with checkpoint-group median relative differences of 0.5752–0.9728. Thus, finite NS differs from matched polar through its vector response as well as the amplitude term in Eq. (11).

Controlled spectra show that this attribution depends on the input spectrum. With singular vectors and Frobenius norm fixed, the tested distinct exponential spectra emphasize between-block response, whereas flat and gapped spectra concentrate the discrepancy within repeated-value blocks. Only aggregate energy within a repeated block is basis invariant. Appendix E.3 gives the complete matrix, direction, and spectral coverage, layer variation, and derivative checks.

5.3 Matrix energy does not determine observable importance

The matrix decomposition alone cannot determine which NS5–polar differences matter to predictions. We therefore propagate individual forcing components through the same NS5 control states and Jacobians, with the same scalar rates and terminal probes. Four directions unused by calibration test the 500-step checkpoints of seeds 201–203 and the 16500-step checkpoints of seeds 0–2 under both history conventions. The comparison retains every signed Fisher cross term (Appendix F).

Observable projection substantially reweights the one-step forcing components (Figure 1a). At 500 steps, the normal share falls from 21.9% of summed weight-forcing energy after actual learning rates to 0.05% of projected self energy, while the amplitude share rises from 10.1% to 50.8%. For transported one-step responses, omitting off-diagonal forcing yields 28.3–29.5% relative Fisher error across the three early seed medians, compared with 0.89–0.94% for normal forcing. Off-diagonal omission remains consequential later, at 12.2–34.8%. A matched-polar shape surrogate has 37.8–41.9% error early; adding only amplitude correction still leaves 28.4–29.5%. These comparisons identify predictive information that matrix-energy shares and total defect norms leave undetermined.

Finite-dose accuracy has a narrower scope (Figure 1b). At 500 steps, raw transported relative KL errors have pooled medians 0.0592 at $T = 1$ and 0.0433 at $T = 4$ over 72 seed–direction–dose cases after probe averaging. At 16500 steps, the one-step median is 0.4026 and four-step predictions fail at the tested doses. Retained results and surrogate rankings remain mixed, with dose-specific seed

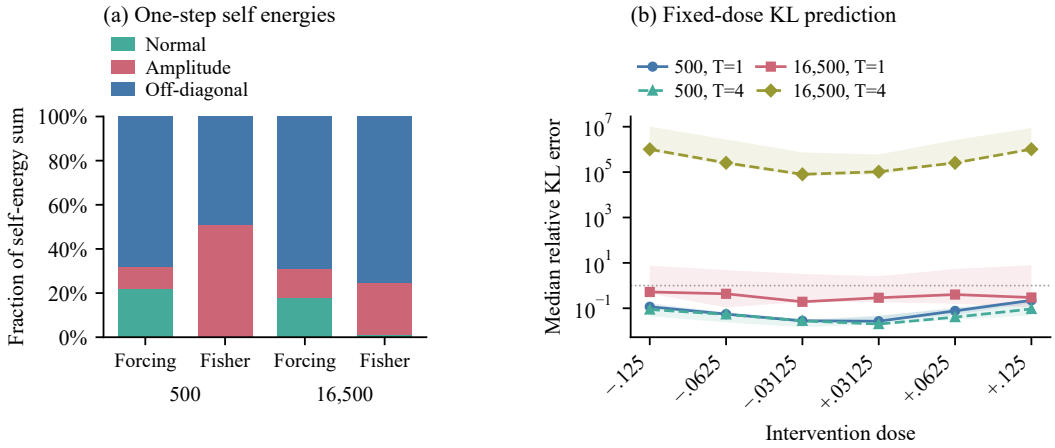


Figure 1: Component propagation and its predictive range. (a) One-step NS5–polar gap forcing after the actual matrix learning rates, and projected Fisher self energies, pooled over three seeds and four directions per checkpoint group. Fractions exclude cross terms and are not KL shares. (b) Full transported prediction at each dose: median of three seed-level median relative KL errors, with their range shaded. Each direction–dose case first averages sixteen probe windows.

variability in Appendix F. Local information loss and finite-dose error are therefore distinct: omitting a component may accidentally improve finite-dose error without recovering its tangent.

5.4 Local responses predict short-horizon functional change

We next test whether the full-state response quantitatively predicts actual finite interventions along the fixed control trajectory. An independent protocol fixes seed 200 for calibration and seeds 201–203 for evaluation before execution, with held-out windows and two pre-fixed directions shared across the splits. Both history conventions are tested at $T = 1, 4$ and $s \in \pm\{0.03125, 0.0625, 0.125\}$ using full-state AD (Eq. (21)). We report mean absolute KL error (MAE), median relative error, and Spearman correlation ρ.

Table 2: Independent confirmation (mean $\pm$ sample SD across three training seeds). Each seed uses 192 window–dose records. Relative error averages seed medians, distinct from the pooled text values; ρ averages within-seed correlations. MAE units are 10^{-7} nats/token. Transported transforms weights and optimizer history consistently; retained transforms weights only. Cal. denotes the locked KL calibration; raw denotes no calibration.

T	State	Mode	Factor	KL MAE	Rel. error	ρ
1	Transported	Cal.	1.046	0.513 ± 0.14	0.0788 ± 0.031	0.97 ± 0.011
1	Transported	Raw	1	0.571 ± 0.11	0.073 ± 0.028	0.97 ± 0.011
1	Retained	Cal.	1.203	1.44 ± 0.12	0.184 ± 0.021	0.955 ± 0.0051
1	Retained	Raw	1	1.83 ± 0.17	0.193 ± 0.019	0.955 ± 0.0051
4	Transported	Cal.	1.023	5.92 ± 1.2	0.0615 ± 0.0086	0.983 ± 0.0043
4	Transported	Raw	1	6.02 ± 1	0.0572 ± 0.0036	0.983 ± 0.0043
4	Retained	Cal.	1.084	10.2 ± 2.2	0.101 ± 0.02	0.98 ± 0.0049
4	Retained	Raw	1	12 ± 1.7	0.0966 ± 0.0089	0.98 ± 0.0049

Raw pooled recordwise median relative KL errors at one and four steps are 0.07365 and 0.05728 for transported state, and 0.1890 and 0.09872 for retained state, respectively. Table 2 reports seed-level means and SDs separately. Locked multiplicative calibration lowers MAE in all four strata, but median relative error can increase and ranking is unchanged. The separate signed-loss readout also has lower MAE than a zero predictor in every stratum, with only a small retained one-step improvement (Table 12). These results support local readouts along the specified continuations.

5.5 Direction information improves relative KL calibration

For transported one-step responses, we test whether projection adds information beyond defect magnitude. Calibrated Jacobian/Fisher predictions are compared with calibrated squared commutator and squared weight-defect norm. Their coefficients are fitted on seed 200 and locked separately. All methods share directions, doses, and five pre-fixed probe draws with nested budgets $B \in \{1, 4, 16, 64\}$, using windows disjoint from calibration and confirmation. Predictions and actual KL are averaged over matched windows before computing errors.

Table 3: Transported state, $T = 1$, $B = 64$ (mean $\pm$ sample SD across three training seeds). Each seed averages five fixed probe draws; each draw uses twelve direction–dose cases after matched-window averaging. Calibration uses seed 200 only. Cal. denotes calibrated predictions.

Metric	Jacobian/Fisher	Cal. κ^2	Cal. $\|\delta_s\|^2$
Rel. error	0.0838 ± 0.038	0.175 ± 0.071	0.182 ± 0.081
MAE (10^{-7})	0.513 ± 0.18	0.564 ± 0.14	0.566 ± 0.15

Jacobian/Fisher has lower mean median relative error than both proxies in all twelve training-seed–budget combinations (Table 3 shows $B = 64$). Absolute-error results are mixed: projection has lower MAE in ten of twelve combinations against the commutator proxy and eleven against squared defect norm, but higher MAE than the commutator for seed 201 at $B = 64$. Direction information therefore helps most consistently in relative KL error. Appendix E.4 reports all budgets, state conventions, horizons, and aggregation metrics; Appendix E.4 gives the separate larger-dose comparison.

5.6 Memory interference and the scope of local prediction

Retained history changes the response through memory self energy and its signed cross term with forcing. In the independent confirmation, these two terms together reduce the mean Fisher response relative to forcing alone. Yet including memory improves one-step finite-dose prediction in 74/96 cases at $s = +0.03125$ and only 11/96 at $s = -0.125$ (Appendix E.4). Memory interference describes the local response; it does not alone determine finite-dose accuracy.

The confirmed predictive range is also limited by horizon, checkpoint, and arithmetic. An independent scan through $T \in \{1, 4, 8, 16, 64\}$ finds agreement between direct AD and explicit forcing plus memory within 4.96×10^{-13}, but actual-path derivatives remain unresolved and finite-dose predictions fail at $T = 16, 64$ (Appendix E.5). Multi-seed confirmation at the 500-step checkpoints therefore covers $T = 1, 4$, and Section 5.3 shows that this range does not transfer unchanged to later checkpoints. Low-precision endpoint KL can also be comparable to a numerical null; controls with different perturbation directions or learning rules cannot be subtracted as additive causal contributions (Appendix E.7). Appendix F.3 reports the diagnostic's runtime and memory cost.

6 Conclusion

Libra connects representation-dependent updates to local predictive change through scale homogeneity, matrix response, full-state propagation, and observable projection. Our scale-compensated analysis distinguishes shared homogeneity from the heterogeneous mechanisms of exact-polar and finite Newton–Schulz updates. The state-resolved framework then connects update defects to logit, KL, and signed-loss responses under specified history conventions and fixed control trajectories. Independent experiments support local prediction at one and four steps on three held-out training seeds in 106.57M-parameter SwiGLU language models. Within this short-horizon scope, Libra explains why update geometry must be interpreted through state dynamics and observable projection. Further discussion of these findings and their scope appears in Appendix G.2.

AI USE STATEMENT

Generative AI tools assisted with language polishing and editing of the manuscript and with figure preparation. The authors are responsible for the technical content, experimental results, references, and the final manuscript.

REFERENCES

Albert Catalan-Tatjer and Jonas Geiping. SIB: Reparameterization of LLMs for better learning-forgetting under SFT. In *ICML Workshop on Weight Space Symmetries*, 2026. URL https://openreview.net/forum?id=p6MoyF2fiq.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018. URL https://arxiv.org/abs/1806.00900.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023. URL https://arxiv.org/abs/2305.07759.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173, 2017. URL https://proceedings.mlr.press/v70/franceschi17a.html.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Evan S. Gawlik and Melvin Leok. Computing the Fréchet derivative of the polar decomposition. *arXiv preprint arXiv:1608.04491*, 2016. URL https://arxiv.org/abs/1608.04491.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Charles Kenney and Alan J. Laub. Polar decomposition and matrix sign function condition estimates. *SIAM Journal on Scientific and Statistical Computing*, 12(3):488–504, 1991. doi: 10.1137/0912027. URL https://doi.org/10.1137/0912027.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2012.04728.

Tim Tsz-Kit Lau and Weijie Su. Symmetry-compatible principle for optimizer design: Embeddings, LM heads, SwiGLU MLPs, and MoE routers. *arXiv preprint arXiv:2605.18106*, 2026. URL https://arxiv.org/abs/2605.18106.

Yibang Li, Bihari Lal Pandey, Ravi Sah, Andi Han, Cyrus Mostajeran, Pratik Jawanpuria, and Bamdev Mishra. Intrinsic Muon: Spectral optimization on Riemannian matrix manifolds. *arXiv preprint arXiv:2605.09238*, 2026. URL https://arxiv.org/abs/2605.09238.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122, 2015. URL https://proceedings.mlr.press/v37/maclaurin15.html.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nati Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper/2015/hash/eaa32c96f620053cf442ad32258076b9-Abstract.html.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Data-dependent path normalization in neural networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06747.

Lawrence K Saul. Weight-balancing fixes and flows for deep learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uaHyXxyp2r.

Tejas Pradeep Shirodkar. Dead-direction conditioners: Gauge-equivariant preconditioning for deep networks. *arXiv preprint arXiv:2606.29176v1*, 2026. URL https://arxiv.org/abs/2606.29176v1.

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. URL https://arxiv.org/abs/2605.11181.

Rodrigo Carmo Terin. Scale redundancy and soft gauge fixing in positively homogeneous neural networks. *arXiv preprint arXiv:2602.14729*, 2026. URL https://arxiv.org/abs/2602.14729.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. *arXiv preprint arXiv:2205.10637*, 2022. URL https://arxiv.org/abs/2205.10637.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=L0r0GphlIL.

The appendix is organized into theoretical foundations, experimental evidence, and related work and discussion. Appendix A provides the proofs and state calculus, Appendix B derives the finite Newton–Schulz response, and Appendix C analyzes observable interference and numerical resolution. The experimental material begins with the visual analyses in Appendix D, followed by the protocols and additional checks in Appendix E and the component-propagation study, direction coverage, and diagnostic costs in Appendix F. Appendix G covers the connections to prior work and discusses the interpretation and scope of the findings.

A Proofs and State Calculus

A.1 Symmetry and the state action

Let $a_i = \|W_{U,i:}\|_2$ and $b_i = \|W_{D,:i}\|_2$ on the nonzero-norm stratum. Then $\theta_i = \log a_i - \log b_i$ changes by $2s_i$. Differentiating function invariance yields the orbit-gradient identity

$$\langle g_{U,i:}, W_{U,i:}\rangle = \langle g_{D,:i}, W_{D,:i}\rangle\,. \tag{25}$$

The contribution of channel i to the block output is $W_{D,:i}\ \mathrm{SiLU}(W_{G,i:}x)(W_{U,i:}x)$. Scaling the up row by λ_i and the down column by λ_i^{-1} cancels exactly for every x. Differentiation of $\mathcal{L}(G_sW) = \mathcal{L}(W)$ gives Eq. (25) and the covector transformation $\nabla\mathcal{L}(G_sW) = G_s^{-\top}\nabla\mathcal{L}(W)$. The action is diagonal and positive, so an elementwise second moment transforms by the square of each covector factor. This specifies a representation of historical gradient-like quantities; the correctness of future recurrences is a further property of the update implementation.

A common decoupled decay multiplier preserves θ_i; compensation must preserve that actual multiplier rather than rescale the decay substep inadvertently.

A.2 Proof of Proposition 3.1

The covector-scaled update is $\mathcal{A}(c^{-1}Z) = c^{-h}\mathcal{A}(Z)$. Canceling its common nonzero norm in the symmetric commutator gives

$$\frac{2|c^{-h} - c|}{c^{-h} + c} = \frac{2|1 - c^{h+1}|}{1 + c^{h+1}} = 2\tanh\left(\frac{|(h+1)s|}{2}\right).$$

For the compensated step, $\eta_U'\mathcal{A}(c^{-1}Z_U) = c^{h+1}\eta_U c^{-h}\mathcal{A}(Z_U) = c\eta_U\mathcal{A}(Z_U)$, and similarly the down update has factor c^{-1}. Thus the weights remain related by G_s. If the full gradient and state recurrences transform covariantly, this relation re-establishes the transformed-gradient relation for the next batch. Induction gives full trajectory conjugacy. Covariant initial state, compatible decay, and all other stated preprocessing assumptions are essential to this induction.

For elementwise Adam at $\epsilon = 0$, transported bias-corrected moments satisfy $m' = t^{-1}m$, $v' = t^{-2}v$ for a positive parameter factor t, so $m'/\sqrt{v'} = m/\sqrt{v}$. Multiplying the coordinate learning rate by t produces the required tangent update. At nonzero ϵ, setting $\epsilon' = t^{-1}\epsilon$ preserves this ratio; retaining a scalar global ϵ generally does not. Exactly zero histories must use a defined, compatible zero-update convention rather than an undefined $0/0$.

A.3 Proof of Theorem 3.2

Let $E = -DZ = -DQH$, and differentiate $Z = QH$ as $E = \dot{Q}H + Q\dot{H}$. Differentiating $Q^\top Q = I$ gives $\Omega = Q^\top\dot{Q}$ skew-symmetric. Multiplication by $I - QQ^\top$ yields

$$(I - QQ^\top)\dot{Q} = (I - QQ^\top)EH^{-1} = -(I - QQ^\top)DQ.$$

Taking the skew part of $Q^\top E = \Omega H + \dot{H}$ yields

$$H\Omega + \Omega H = Q^\top E - E^\top Q = HC - CH.$$

This is the standard Sylvester representation of the polar derivative (Gawlik & Leok, 2016). Since $H \succ 0$, its solution is unique. In the eigenbasis of H,

$$(V^\top\Omega V)_{ij} = \frac{\sigma_i - \sigma_j}{\sigma_i + \sigma_j}\widetilde{C}_{ij}.$$

The normal component $-(I - QQ^\top)DQ$ is Frobenius-orthogonal to $Q\Omega$, and multiplication by Q preserves the Frobenius norm of an $n \times n$ matrix. This proves Eq. (9). Both terms vanish precisely under the two stated conditions. Because the spectral ratios have magnitude at most one,

$$\|L_D\|_F^2 \le \left\|(I - QQ^\top)DQ\right\|_F^2 + \|C\|_F^2 = \|DQ\|_F^2 .$$

Uniform components of D add aI to C and nothing to its commutator with H, and have no normal component. Overall rescaling of Z scales all singular values equally and leaves Q and every ratio unchanged. These observations prove Eq. (10). For the uncompensated derivative, use $DQ = (I - QQ^\top)DQ + QC$ and $\langle C, \Omega\rangle = 0$ to obtain Eq. (26). Applying the same derivation to the transposed down matrix changes the sign, not the energy.

The full-rank condition ensures a differentiable polar factor. At rank-deficient inputs, the finite-NS derivative is instead computed by the matrix recurrence in Appendix B.2.

Uncompensated response. Without the row-rate compensation, the first-order pulled-back up-side defect is $L_D - DQ$. Orthogonality of the symmetric and skew-symmetric components gives

$$\|L_D - DQ\|_F^2 = 4\left\|(I - QQ^\top)DQ\right\|_F^2 + \|C\|_F^2 + \|\Omega\|_F^2 . \tag{26}$$

This separates the tangent-scaling contribution from the polar shape response. Its effect on a network still requires the observable projection.

A.4 Proof of the canonicalized null

Canonicalization uses $s_i^\star = -\theta_i/2$, transports the full state, and equalizes paired norms on the nonzero-norm stratum. Define $\mathcal{C}(S) = \rho_{-\theta(W)/2}S$ on the nonzero-norm stratum and

$$\Psi_t = \mathcal{C} \circ \Phi_t \circ \mathcal{C}. \tag{27}$$

This modified learning rule supplies an exact-arithmetic implementation null.

On the nonzero-norm stratum, $\theta(G_sW) = \theta(W) + 2s$ gives

$$\mathcal{C}(\rho_s S) = \rho_{-\theta(G_sW)/2}\rho_s S = \rho_{-\theta(W)/2}S = \mathcal{C}(S).$$

Therefore $\mathcal{C}\Phi_t\mathcal{C}(\rho_s S) = \mathcal{C}\Phi_t\mathcal{C}(S)$. Repeating this equality gives identical canonicalized states at every step. The statement requires full-state canonicalization both before the first update and at subsequent canonicalization calls. Retained buffers, clipped canonical doses, unmatched stochasticity, and finite arithmetic can each break the equality.

A.5 Proof of Theorem 4.1

For a transported initial pulse, let $\widehat{S}_t(s)$ be the actual branch and set $\widetilde{S}_t(s) = \rho_{-sq}\widehat{S}_t(s)$. The dose is fixed along the continuation, so

$$\widetilde{S}_{t+1}(s) = e^{-sK_q}\Phi_t(e^{sK_q}\widetilde{S}_t(s)), \qquad \widetilde{S}_0(s) = S_0.$$

Differentiate the right-hand side at $s = 0$ with its state argument fixed. The derivative is $-K_q\Phi_t(S_t) + D\Phi_t(S_t)K_qS_t = b_t(q)$. Differentiating through the varying state then gives $r_{t+1} = A_t r_t + b_t(q)$. Repeated substitution proves Eq. (16).

For a retained-state pulse, the same conjugated step maps apply; only the initial condition differs:

$$\widetilde{S}_0(s) = e^{-sK_q}e^{sK_q^{\mathrm{ret}}}S_0, \quad \partial_s\widetilde{S}_0(0) = (K_q^{\mathrm{ret}} - K_q)S_0.$$

For example, in the up-side first moment this initial derivative is $+q_i m_{U,i:}$ and in the down-side first moment it is $-m_{D,:i}q_i$; second moments have twice the corresponding generator. Finally, gauge invariance gives $f(\widehat{S}_T(s)) = f(\widetilde{S}_T(s))$, so its derivative is $Df(S_T)r_T$. This derivation is conditional on one exogenous input sequence. Exchanging the derivative and an expectation over sequences requires separate domination assumptions and is not needed for the paired estimand.

Why state cannot be removed from the calculation. Writing A_t in weight–buffer blocks, a buffer perturbation $r_t = (0, r_t^h)$ produces a weight perturbation $A_t^{Wh}r_t^h$ at the next step. Current outputs ignore h, but later outputs can depend on it. Deleting the optimizer buffers from S generally destroys the Markov description needed by the recurrence.

A.6 One-step projection and the KL expansion

Write J_f for the Jacobian of the observable f. Pulling the transported one-step branch back leaves the base weights unchanged and shifts the next weights by $-\delta_s$. Function invariance therefore gives

$$\Delta f = f(W^+ - \delta_s) - f(W^+). \tag{28}$$

Proposition A.1 (Observable projection). *If J_f is L_J-Lipschitz on the connecting segment,*

$$\Delta f = -J_f(W^+)\delta_s + R_s, \qquad \|R_s\| \le \tfrac{1}{2} L_J \|\delta_s\|^2 . \tag{29}$$

Linearity of G_s and function invariance give Eq. (28). The fundamental theorem of calculus yields

$$f(W^+ - \delta) - f(W^+) = -\int_0^1 J_f(W^+ - t\delta)\delta \, dt.$$

Subtract $-J_f(W^+)\delta$ and apply the Jacobian Lipschitz bound to obtain $\|R\| \le \int_0^1 L_J t \|\delta\|^2 dt = \frac{1}{2} L_J \|\delta\|^2$. For logits z and softmax probabilities p, the Hessian of log-sum-exp is $F = \mathrm{diag}(p) - pp^\top$. Hence

$$\mathrm{KL}\big(\mathrm{softmax}(z) \,\|\, \mathrm{softmax}(z + \Delta z)\big) = \tfrac{1}{2} \Delta z^\top F \Delta z + O(\|\Delta z\|^3).$$

Putting $\Delta z = \mathcal{R}_T s + O(\|s\|^2)$ and averaging over the fixed evaluation set proves Eq. (23). Smoothness through the finite training horizon and locally bounded derivatives are assumed. The KL includes the full vocabulary at each position of the fixed probe set.

A.7 A scalar counterexample

Proposition A.2 (Identical commutator, different function-response order). *For $f(u, d) = ud$, $u, d > 0$, let both scalar polar updates have the same nonzero sign σ, so $u^+ = u - \eta\sigma$ and $d^+ = d - \eta\sigma$. Reciprocal scaling gives the same per-side $\kappa = 2\tanh(|s|/2)$ for every u, d, but*

$$\Delta f = -\eta\sigma[(e^s - 1)u + (e^{-s} - 1)d]. \tag{30}$$

For $u = d = a$, this is $-2\eta\sigma a(\cosh s - 1) = O(s^2)$; for $u \neq d$, it generally has a nonzero $O(s)$ term.

Thus the same commutator can accompany $O(s^4)$ or $O(s^2)$ predictive KL under a nondegenerate logit readout. This is a counterexample to a universal proxy. Appendix A.9 states the corresponding local expansion when the first-order readout vanishes.

The transformed one-step product is $(e^s u - \eta\sigma)(e^{-s} d - \eta\sigma)$, whereas the control product is $(u - \eta\sigma)(d - \eta\sigma)$. Subtraction proves Eq. (30). Equal signs occur, for example, under the squared loss $\frac{1}{2}(ud - y)^2$ when $u, d > 0$ and $ud \neq y$, with zero initial momentum. Since the pulse preserves ud, it preserves both gradient signs.

When $u = d = a$, the linear term cancels and the leading change is $-\eta\sigma a s^2$. For two-class logits $(f, 0)$ with a nondegenerate base probability p, the KL leading term is $\frac{1}{2} p(1 - p)\eta^2 a^2 s^4$. For $u \neq d$, the generic leading KL term is quadratic in s. All cases share the same degree-zero parameter commutator. The shared commutator therefore leaves the order of predictive change undetermined.

A.8 Interpreting the state decomposition

Let $P_{T:k} = A_{T-1} \cdots A_k$ with $P_{T:T} = I$. Substitution of $b_k = A_k K_q S_k - K_q S_{k+1}$ gives the telescoping identity

$$\sum_{k=0}^{T-1} P_{T:k+1} b_k = P_{T:0} K_q S_0 - K_q S_T, \tag{31}$$

$$r_T = P_{T:0}(r_0 + K_q S_0) - K_q S_T.$$

For a gauge-invariant observable, $Df(S_T) K_q S_T = 0$. Hence the observable response is exactly the chain rule applied to the initial extended-state perturbation followed by training. The time-resolved forcing terms locate noncommutation in pulled-back coordinates. The state distinction is explicit: transport and retention give different initial extended-state perturbations despite identical current weights in their treated coordinates.

Table 4: Libra response computation. Every step fixes the state convention and the implemented training rule. The matrix decomposition diagnoses the update mechanism; the full-state calculation propagates its observable effect.

Stage	Computation and output
1. Intervention	Specify q, dose s, retained/transported history, and paired data and arithmetic.
2. Update mechanism	Compute the compensated NS/polar derivative and restore original-rule defects with Eq. (19).
3. State propagation	Along $S_{t+1} = \Phi_t(S_t)$, propagate $\xi_{t+1} = D\Phi_t(S_t)\xi_t$; independently accumulate $r_{t+1} = A_t r_t + b_t$.
4. Predictive readout	Project r_T or ξ_T through the logit Jacobian; use Fisher for KL and the loss gradient for signed loss.
5. Validation	Compare direct/forcing AD, path differences across h, finite doses across s, and matched numerical null controls.

Computational form of the telescoping identity. Adding $K_q S_{t+1}$ to $r_{t+1} = A_t r_t + A_t K_q S_t - K_q S_{t+1}$ gives Eq. (21). Thus the direct route initializes the actual intervention tangent and applies one full-state JVP per control step. The explicit route separately propagates initial memory and adds b_t at each step; it subtracts quantities involving $K_q S_t$ and can exhibit different cancellation. Both routes use the same conditioned Φ_t and gauge-invariant terminal readout. Their comparison tests numerical consistency between two evaluations of the same mathematical response. The response checks also compare to actual paired paths and to a chain of finite-difference JVPs, with all denominators specified in Appendix E.5.

A.9 A vanishing first-order readout is a local degeneracy

Let the scalar-dose continued logits $z_T(s)$ be three times continuously differentiable and suppose $z_T'(0) = 0$ on the probe set. Then $z_T(s) - z_T(0) = \frac{1}{2}s^2 z_T''(0) + O(s^3)$, and

$$\mathbb{E}_x \,\mathrm{KL}(p_T(x) \| p_T^{(s)}(x)) = \frac{s^4}{8}\mathbb{E}_x[z_T''(0,x)^\top F_x z_T''(0,x)] + O(|s|^5). \tag{32}$$

This expansion explains the scalar counterexample: a vanishing first-order logit response can leave a nonzero second-order response at finite dose.

A.10 A dose-composition identity for implementation audits

For the linear extended-state action define the finite defect $\mathcal{D}_s(S) = \Phi^{(s)}(S) - \Phi(S)$. The abelian group law gives the exact identity

$$\mathcal{D}_{s+r}(S) = \rho_{-r}\mathcal{D}_s(\rho_r S) + \mathcal{D}_r(S). \tag{33}$$

Indeed, $\Phi^{(s+r)} = \rho_{-r}\Phi^{(s)}\rho_r$; substitute $\Phi^{(s)} = \Phi + \mathcal{D}_s$ and use linearity of ρ_{-r}. This is a standard conjugation identity specialized to the audit. Comparing a composed pulse with its combined-dose counterpart tests state-map composition without assuming that optimizer effects add linearly at finite dose.

B Finite Newton–Schulz Response

B.1 The heterogeneous response of a finite NS polynomial

The following calculation gives the finite-NS matrix derivative used in the mechanism evaluation, including the normalizer derivative. Its SVD-free form below extends computation to repeated spectra and rank-deficient inputs on a fixed branch.

Proposition B.1 (Finite-polynomial response under a multiplicative intervention)**.** *Let $Z = U\Sigma V^\top \in \mathbb{R}^{m\times n}$, $m \geq n$, have positive, distinct singular values σ_i. Let $r = \|Z\|_F$, and let $n(r) > 0$ be a*

differentiable scalar normalizer near r. Define

$$x_i = \sigma_i/n(r), \qquad y_i = \gamma p_5(x_i), \qquad \mathcal{A}_5(Z) = U\,\mathrm{diag}(y_i)V^\top, \tag{34}$$
$$p_0(x) = x, \qquad p_{k+1}(x) = ap_k(x) + bp_k(x)^3 + cp_k(x)^5, \quad k = 0, \dots, 4.$$

The multiplier γ is fixed along the intervention. For $Z(s) = e^{-sD}Z$, $D = \mathrm{diag}(q)$, put

$$C = U^\top DU, \qquad \mu = \sum_j \frac{\sigma_j^2}{r^2} C_{jj}, \qquad \tau = \frac{rn'(r)}{n(r)}. \tag{35}$$

Then

$$\dot{y}_i = -\gamma x_i p_5'(x_i)(C_{ii} - \tau\mu), \tag{36}$$
$$\dot{\mathcal{A}}_5 = \big[N + UB\big]V^\top, \qquad N = -(I - UU^\top)DU\,\mathrm{diag}(y_i), \tag{37}$$
$$B_{ii} = \dot{y}_i, \qquad B_{ij} = C_{ij}\frac{(\sigma_i^2 + \sigma_j^2)y_j - 2\sigma_i\sigma_j y_i}{\sigma_i^2 - \sigma_j^2} \quad (i \neq j). \tag{38}$$

In particular, there is an orthogonal energy decomposition

$$\left\|\dot{\mathcal{A}}_5\right\|_F^2 = \|N\|_F^2 + \sum_i \dot{y}_i^2 + \sum_{i\neq j} |B_{ij}|^2. \tag{39}$$

Proof. Write $E = \dot{Z} = -DU\Sigma V^\top$. The singular-value differential is $\dot{\sigma}_i = (U^\top EV)_{ii} = -C_{ii}\sigma_i$ and $\dot{r} = \langle Z, E\rangle / r = -r\mu$. Therefore $\dot{x}_i = -x_i(C_{ii} - \tau\mu)$, proving Eq. (36) by the chain rule. For the vector part, $U^\top EV = -C\Sigma$. The left-vector component normal to the column space is $-(I - UU^\top)DU$; multiplying it by the output amplitudes gives N. Let $A = U^\top\dot{U}$ and $G = V^\top\dot{V}$, both skew-symmetric. The off-diagonal part of the differentiated SVD gives

$$A_{ij} = C_{ij}\frac{\sigma_i^2 + \sigma_j^2}{\sigma_i^2 - \sigma_j^2}, \qquad G_{ij} = C_{ij}\frac{2\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}. \tag{40}$$

Differentiating $U\,\mathrm{diag}(y)V^\top$ yields $B = A\,\mathrm{diag}(y) + \mathrm{diag}(\dot{y}) - \mathrm{diag}(y)G$, which gives Eq. (38). Finally, $U^\top N = 0$; inside B, diagonal and off-diagonal entries are Frobenius-orthogonal. This proves Eq. (39). □

Normalizer and implementation scope. The LM NS5 implementation and its FP64 JVP reference use $n(r) = \max(r, 10^{-7})$. Thus $\tau = 1$ for $r > 10^{-7}$, where the main-text formula applies, and $\tau = 0$ for $0 < r < 10^{-7}$; the switching boundary is excluded. An additive normalization constant instead gives $n(r) = r + \epsilon$ and $\tau = r/(r + \epsilon)$. Recomputing norm matching after the pulse would add a $\dot{\gamma}p_5(x_i)$ term and is not the frozen-factor control. For repeated singular values, the displayed SVD fractions should be replaced by a continuous matrix-derivative calculation, not evaluated with a zero denominator. The calculation is evaluated in FP64 on the specified smooth branch.

Connection to the exact polar and the observed discrepancy. For the frozen norm match, $N^{\mathrm{P}} = -\alpha(I - UU^\top)DU$ and $B_{ij}^{\mathrm{P}} = \alpha C_{ij}(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)$ with zero diagonal. Subtracting these components gives Eq. (13), the energy of the derivative difference rather than a difference of energies.

Set all output amplitudes to $y_i = 1$. Their derivatives vanish, and $B_{ij} = C_{ij}(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)$; Eqs. (37)–(39) reduce to the exact-polar formula. A finite polynomial instead has a spectral-amplitude response as well as different weights on the vector responses. Under $D = dI$ and $n(r) = r$, $C = dI$ and $\mu = d$, so every response term vanishes for both maps. The uniform control therefore tests a common invariance, while heterogeneous directions distinguish the two derivatives.

Synthetic verification of the finite-NS derivative. A deterministic float64 calculation checks the full derivative against symmetric differences on twenty 9×5 Gaussian matrices for each of $\epsilon \in \{0, 10^{-7}, 0.2\}$, using $n(r) = r + \epsilon$, $\gamma = 1.3$, five polynomial steps, and $h = 2 \times 10^{-6}$. The coefficients are $(3.4445, -4.7750, 2.0315)$. Maximum relative derivative error is 1.47×10^{-9} and maximum energy residual is 2.85×10^{-14} across these sixty checks. These synthetic checks verify the derivative calculation.

B.2 AN SVD-FREE DERIVATIVE AND THE NS–POLAR DIFFERENCE

Let $E = \dot{Z}$ and $r = \|Z\|_F$. Write the implemented normalizer as $n = \max(r, \varepsilon)$ with $\varepsilon > 0$, and set

$$X_0 = Z/n, \qquad H_0 = E/n - Z\dot{n}/n^2, \qquad \dot{n} = \begin{cases} \langle Z, E\rangle/r, & r > \varepsilon, \\ 0, & r < \varepsilon. \end{cases} \tag{41}$$

Below the floor this formula includes $Z = 0$: normalization is locally constant, so $H_0 = E/\varepsilon$. The branch boundary $r = \varepsilon$ is excluded. With fixed polynomial coefficients and a fixed scale γ, define for $k = 0, \ldots, 4$

$$\begin{aligned} G_k &= X_k^\top X_k, & \dot{G}_k &= H_k^\top X_k + X_k^\top H_k, \\ P_k &= aI + bG_k + cG_k^2, & \dot{P}_k &= b\dot{G}_k + c(\dot{G}_k G_k + G_k \dot{G}_k), \\ X_{k+1} &= X_k P_k, & H_{k+1} &= H_k P_k + X_k \dot{P}_k, \\ \mathcal{A}_5(Z) &= \gamma X_5, & D\mathcal{A}_5(Z)[E] &= \gamma H_5. \end{aligned} \tag{42}$$

These formulas follow directly from the product rule, including both orders in the derivative of G_k^2. They can equivalently be evaluated with a left Gram matrix and transposition, as in the stored implementation. Each finite matrix step is polynomial; the normalization is smooth on either open branch. Consequently the total finite-NS derivative exists at repeated singular values and rank-deficient inputs on these branches. This conclusion concerns the real-valued finite-NS map, evaluated numerically in FP64. With a general differentiable positive normalizer $n(r)$, replace $\dot{n}$ by $n'(r)\langle Z, E\rangle/r$ away from $r = 0$.

For the multiplicative direction use $E = -DZ$. Subtracting the frozen matched polar derivative from Proposition B.1 yields Eq. (12). Since $U^\top \Delta N = 0$ and U, V have orthonormal columns, its squared Frobenius norm is $\|\Delta N\|_F^2 + \left\|B - B^{\mathrm{P}}\right\|_F^2$. The polar diagonal is zero. Splitting the remaining matrix into diagonal and off-diagonal entries proves Eq. (13); differentiating α would invalidate this fixed-comparator identity.

Degenerate spectral blocks. For full-column-rank Z, partition indices into exact equal-singular-value blocks I_a. Compute the total derivative using Eq. (42), subtract the polar derivative, and form $B^\Delta = U^\top \Delta\dot{A} V$ and the normal component $\Delta N = (I - UU^\top)\Delta\dot{A}V$. A basis-invariant split is

$$\left\|\Delta\dot{A}\right\|_F^2 = \|\Delta N\|_F^2 + \sum_a \left\|B^\Delta_{I_a I_a}\right\|_F^2 + \sum_{a \neq b} \left\|B^\Delta_{I_a I_b}\right\|_F^2 . \tag{43}$$

Rotating singular vectors jointly inside any equal-value block applies orthogonal left/right factors to these blocks and preserves their Frobenius norms. Individual diagonal and off-diagonal shares inside a repeated block need not be invariant. The recorded numerical grouping uses $|\sigma_i - \sigma_j| \leq 10^{-10}\sigma_{\max}$; for the flat, gap, and repeated-tail constructions this groups the prescribed repeated values. Threshold grouping of merely nearby values is a numerical reporting convention, not an additional exact symmetry. The finite-NS total derivative also covers rank-deficient inputs; the matched-polar comparison uses full rank.

C OBSERVABLE INTERFERENCE AND NUMERICAL RESOLUTION

C.1 DIRECTION-SENSITIVE PROJECTION AND MEMORY INTERFERENCE

For a fixed probe, $H_f = \mathbb{E}_x[J_f^\top F_x J_f]$ gives

$$\mathrm{KL} = \tfrac{1}{2}\delta^\top H_f \delta + O(\|\delta\|^3) = \tfrac{1}{2}\|\delta\|^2 g(\delta) + O(\|\delta\|^3), \qquad g(\delta) = \frac{\delta^\top H_f \delta}{\|\delta\|^2}. \tag{44}$$

A zero eigenvalue of the response Gram matrix indicates first-order insensitivity on the fixed probe set. Signed loss uses its own gradient readout. A scalar fitted to update norms summarizes this gain on its calibration directions; it need not transfer to another direction or checkpoint. This motivates testing projection against scalar proxies, rather than assuming that good ranking implies magnitude calibration.

Let $a(x)$ and $b(x)$ be the forcing and initial-memory logit tangents, and $\langle a, b\rangle_F = \mathbb{E}_x[a(x)^\top F_x b(x)]$. Expanding the squared Fisher norm gives

$$\mathrm{KL}_{\mathrm{full}} - \mathrm{KL}_{\mathrm{forcing}} = s^2\langle a, b\rangle_F + \tfrac{1}{2}s^2 \|b\|_F^2 + O(|s|^3). \tag{45}$$

The cross term has either sign, so memory need not increase predictive distance. Inclusion of memory is required by the retained-state model even when a finite-dose error metric worsens. The state comparison therefore measures both the signed cross term and the full-versus-forcing prediction errors.

C.2 Derivative resolution and finite-dose remainder

Let $z_T(s)$ be the exact-arithmetic logits under the specified rule, $v_T = z_T'(0)$, $\widehat{v}_{T,h}$ a numerical derivative, and $d_T(s) = z_T(s) - z_T(0)$. Then

$$d_T(s) - s\widehat{v}_{T,h} = \underbrace{d_T(s) - sv_T}_{\text{finite-dose remainder}} + \underbrace{s(v_T - \widehat{v}_{T,h})}_{\text{derivative-estimation error}}. \tag{46}$$

A comparison with a finite-precision branch also includes its arithmetic discrepancy. Small derivative error does not remove the finite-dose remainder, while a large remainder alone cannot diagnose a faulty derivative. A standard finite-difference bound further separates truncation from arithmetic. Suppose z_T has third directional derivative norm bounded by $M_{3,T}$ on $[-h, h]$, and each evaluated endpoint has error at most $\varepsilon_{\pi,T}$. For the symmetric estimate,

$$\|\widehat{v}_{T,h} - v_T\| \le \frac{M_{3,T}}{6}h^2 + \frac{\varepsilon_{\pi,T}}{h}. \tag{47}$$

Taylor's theorem bounds the symmetric-difference truncation; the two endpoint errors contribute at most $2\varepsilon_{\pi,T}/(2h)$. For positive constants independent of h, the upper bound is minimized at $h = (3\varepsilon_{\pi,T}/M_{3,T})^{1/3}$. The response experiments scan step sizes empirically; the bound explains the competing truncation and arithmetic terms but its constants are not fitted. Smaller steps can therefore increase finite-precision error. With $\|z_T''\| \le M_{2,T}$,

$$\|d_T(s) - s\widehat{v}_{T,h}\| \le \tfrac{1}{2}M_{2,T}s^2 + |s|\,\|v_T - \widehat{v}_{T,h}\|. \tag{48}$$

C.3 Amplification and numerical controls

Let the unperturbed reference satisfy $S_{t+1} = \Phi_t(S_t)$, and let the pulled-back perturbed implementation obey

$$\widetilde{S}_{t+1} = \Phi_t^{(s)}(\widetilde{S}_t) + \nu_t, \qquad e_t = \widetilde{S}_t - S_t.$$

The term ν_t represents the numerical discrepancy relative to this exact reference; in a paired finite-precision analysis both branches' injections must be tracked. Suppose $\Phi_t^{(s)}$ is L_t-Lipschitz on the relevant segment. Adding and subtracting $\Phi_t^{(s)}(S_t)$ gives

$$\|e_T\| \le \left(\prod_{j=0}^{T-1} L_j\right)\|e_0\| + \sum_{k=0}^{T-1}\left(\prod_{j=k+1}^{T-1} L_j\right)\big(\|\mathcal{D}_{s,t=k}(S_k)\| + \|\nu_k\|\big). \tag{49}$$

This deterministic bound is generally loose. Its role is logical: exact optimizer non-equivariance and arithmetic error are both injected into a dynamical amplifier, so a large endpoint difference cannot be attributed by comparing it only with the initial function error. A numerical null with the same error norm but a different direction may also amplify differently. The numerical evaluation therefore compares several matched controls with distinct perturbation directions.

D Additional Visualizations of Archived Experiments

The following analyses connect the archived measurements to the two contributions of Libra: the decomposition of heterogeneous update responses and their propagation to predictive observables. They examine layer variation, finite-dose prediction, the information supplied by observable projection, and the role of optimizer history. All figures reuse the reported experiments, with training seeds as the replication unit.

D.1 Layer variation in the finite-NS mechanism

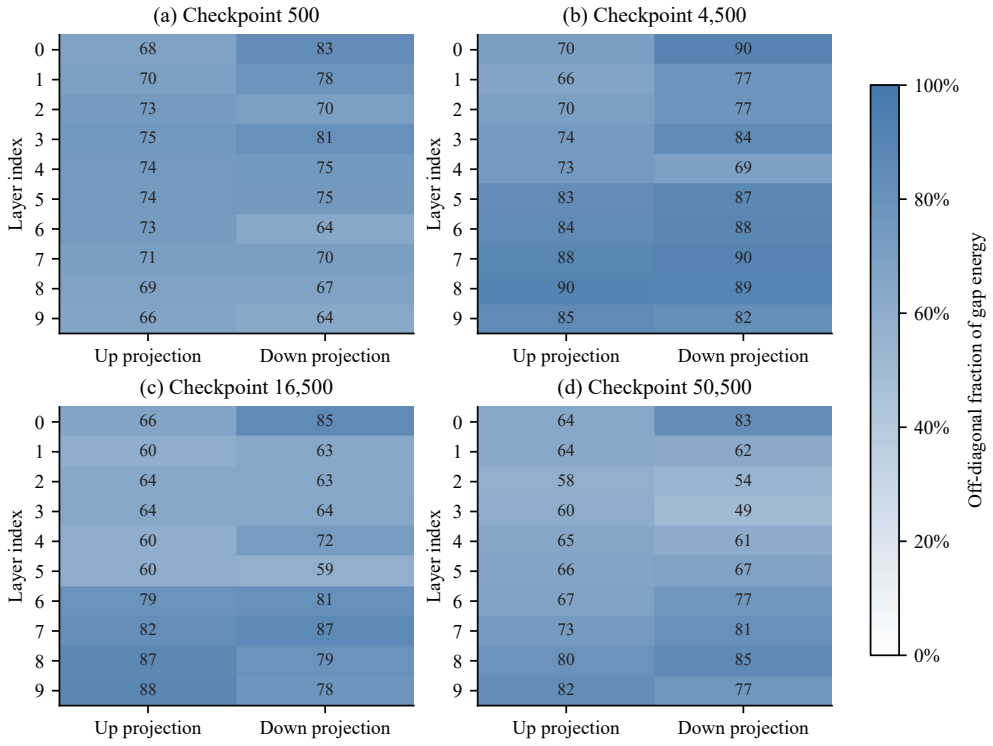


Figure 2: Layer-resolved NS5–polar derivative-difference energy. Each cell divides summed off-diagonal energy by summed total gap energy over three training seeds and 32 random directions for the indicated layer and projection (96 cases per cell). Cell values are in %; all panels share the same color scale. The four checkpoints contain 7,680 cases.

Figure 2 shows that the off-diagonal contribution extends across layers and both projections, with variation across checkpoints. This coverage complements the pooled result in Section 5.2 and supports the need for the off-diagonal term in Eq. (13), beyond a description based only on spectral amplitudes. The fractions characterize matrix derivatives under the tested random directions. They do not measure predictive importance, because state propagation and observable projection can reweight the components and introduce cross terms. Figure 1 supplies the corresponding evidence in prediction space.

D.2 AGREEMENT BETWEEN LOCAL PREDICTIONS AND FINITE INTERVENTIONS

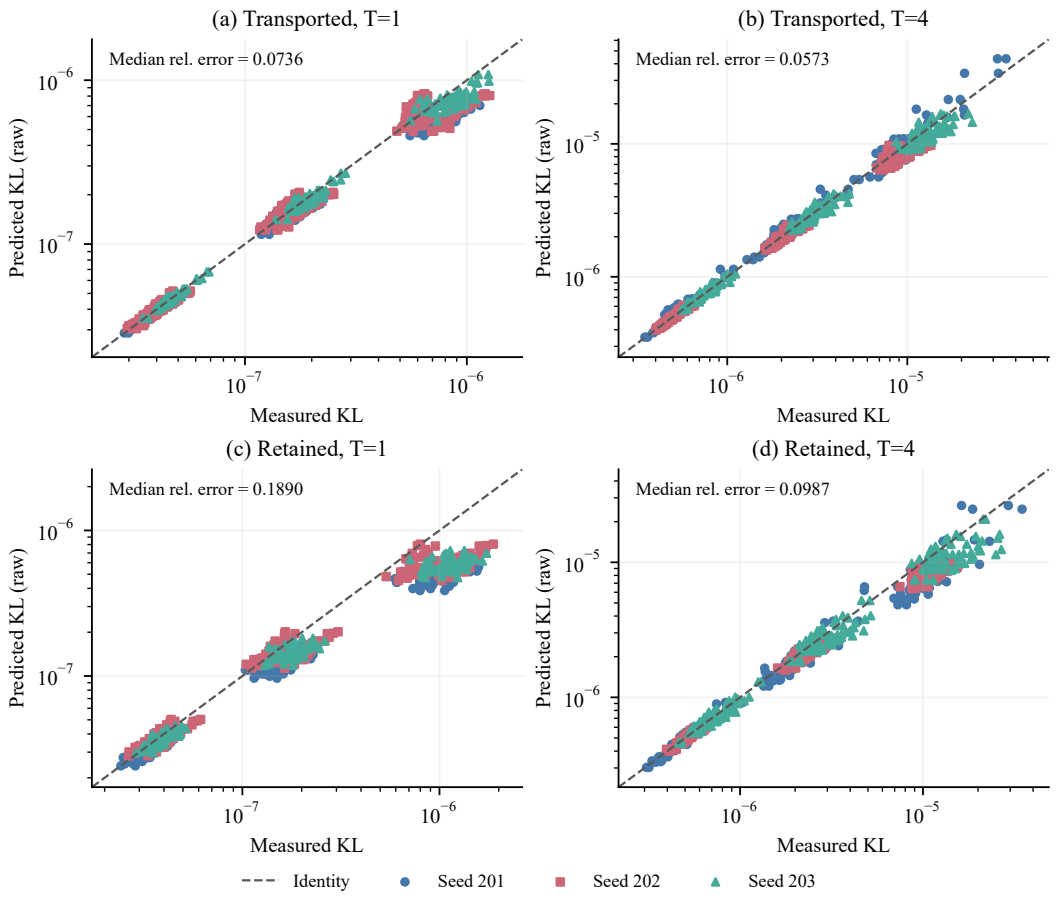


Figure 3: Raw predicted and measured KL in the independent confirmation cohort. Each panel contains 576 records from three held-out training seeds (201–203), two directions, sixteen probe windows, and six signed doses. The dashed line denotes equality. Annotations report pooled recordwise median relative error, rather than a mean of seed-level summaries. Windows and doses within a training seed are dependent observations. All plotted KL values are positive, and no calibration is applied.

Figure 3 shows agreement between the raw local readout and measured KL across the tested intervention magnitudes, with larger deviations under retained history. Transported one-step predictions have a pooled median relative error of 0.07365, compared with 0.1890 under retention. These observations support the connection from full-state propagation in Eq. (21) to the Fisher readout in Eq. (23), under the independent protocol in Section 5.4. The evidence covers the 500-step checkpoints, the specified doses, and horizons $T = 1, 4$. It does not establish accuracy throughout training or at longer horizons, for which Section 5.3 and Appendix E.5 document limitations.

D.3 DIRECTION INFORMATION ACROSS PROBE BUDGETS

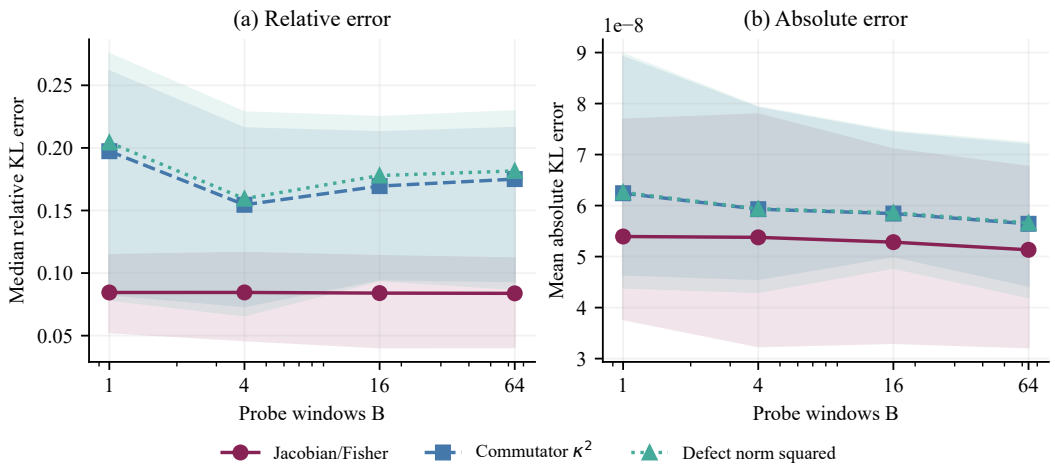


Figure 4: Probe-budget dependence for transported one-step prediction. Each predictor uses its separately locked calibration coefficient. For each seed and budget, each metric is first averaged over five pre-fixed probe draws. Lines average the three seed summaries, and bands span their minimum and maximum. These bands are descriptive ranges, not confidence intervals. Panels report (a) median relative KL error and (b) mean absolute KL error under the same nested probe budgets.

Observable projection has lower mean median relative KL error than both scalar proxies at every tested probe budget (Figure 4a). At $B = 64$, the seed-averaged error is 0.0838 for Jacobian/Fisher, compared with 0.1751 for the calibrated commutator and 0.1816 for the calibrated defect norm. This comparison supports the direction-dependent gain described after Eq. (23): a scalar defect magnitude omits information about output sensitivity along the defect direction. The lower seed-averaged MAE in Figure 4b does not imply an advantage for every seed. In particular, projection has higher MAE than the commutator for seed 201 at $B = 64$, as reported in Appendix E.4. The evidence therefore supports the relative-error comparison more consistently than an unconditional absolute-error advantage.

D.4 Optimizer memory and signed interference

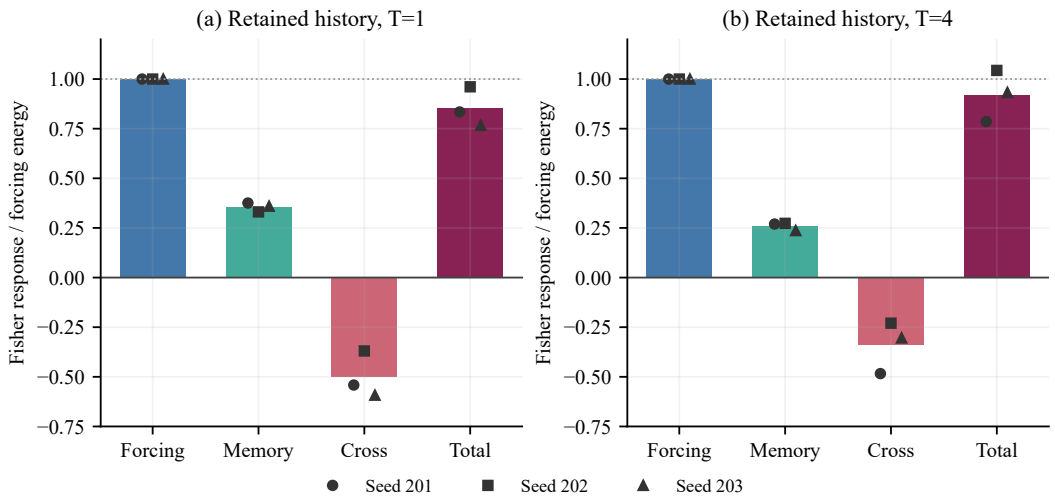


Figure 5: Signed memory interference under retained history. Tangent measurements repeated across doses are deduplicated to give 32 direction–window observations per seed and horizon. Within each seed, each component is averaged over these observations and divided by the mean forcing energy. Bars average the three seed ratios, and symbols identify the seeds. The signed cross term is $2\langle v_{\mathrm{forcing}}, v_{\mathrm{memory}}\rangle_F$; total equals forcing plus memory plus cross.

The negative cross terms in Figure 5 show that the propagated initial-memory response opposes part of the forcing response in Fisher geometry. This behavior explains why the state decomposition in Eq. (16) requires both initial mismatch and ongoing non-equivariance, and why their projected energies must include the cross term in Eq. (24). Negative interference alone does not ensure that the total response is below forcing energy: seed 202 at $T = 4$ has sufficient memory self energy to exceed that reference. Nor does cancellation guarantee a more accurate finite-dose prediction. The dose-dependent comparisons in Appendix E.4 distinguish this local geometric effect from prediction error. The figure thus supports explicit optimizer-state treatment without assigning a uniform benefit to memory retention.

E Core Experimental Protocols and Additional Checks

E.1 Shared model, optimizer, and reporting conventions

Tables report means and sample SDs across training seeds (SD denominator $n-1$), after the within-seed reductions specified in their captions. The two-seed exploratory cohort and repeated timing measurements are labeled separately. The independent confirmation reports mean absolute KL error (MAE), median relative error, and Spearman correlation ρ; its calibration/evaluation split is fixed before execution.

The model, training configuration, and optimizer parameters are specified in Section 5.1. Additional metric definitions follow.

All KLs use natural logarithms and the full vocabulary. Independent-confirmation KL averages all logit positions, including the last; next-token signed loss averages only positions with labels. Its KL MAE is a mean absolute error and relative error is a recordwise median with denominator $\max(|\,\mathrm{KL}\,|, 10^{-20})$. The larger-dose exploratory cohort instead reports median absolute KL error and uses $\max(\mathrm{KL}, 10^{-12})$ for relative error. Results are reported separately by cohort. Directions, windows, doses, and checkpoints index descriptive cases within training seeds. The fixed experiment record retains full-precision CSVs, raw traces, source versions, checkpoint/data hashes, the protocol, and the calibration lock; tables round values for display; full precision is retained in the records.

E.2 Intervention, recurrence, and canonical-state checks

The architectural audit uses a 29.8M bias-free SwiGLU decoder (width 448, ten layers, seven heads, hidden width 1600, byte-level vocabulary) on a fixed TinyStories subset (Eldan & Li, 2023). Correct up/down scaling has zero measured relative output error in float64 and median normalized orbit-gradient residual 9.93×10^{-16}. Applying the same recipe incorrectly to gate/down changes outputs by a median 24.3% at factor two, with residual 0.279. This smaller model checks the symmetry implementation. The same reciprocal cancellation applies when the unchanged gate is GeGLU.

In the populated Muon audit, transport removes the stale-buffer recurrence discrepancy while leaving the update-map defect (Table 5). Immediate FP64 mean relative logit errors are 8.65×10^{-16} for uniform and 8.69×10^{-16} for heterogeneous pulses; isolated FP32 means are $(1.68 \pm 0.18) \times 10^{-7}$ and $(1.68 \pm 0.17) \times 10^{-7}$ over three seeds. The joint commutator uses its original joint normalization and is not directly the per-side quantity in Proposition 3.1.

Table 5: Mean $\pm$ sample SD across three training seeds. Populated-state audit at $s = 0.5$. Zero-momentum rows also transport state. The original joint normalization and implemented arithmetic are retained; these joint values are separate from the per-side scale identity.

Profile	State	Recurrence	Polar joint	NS5 joint
Heterogeneous	Retained	0.358 ± 0.012	0.605 ± 0.00095	0.61 ± 0.0011
Heterogeneous	Transported	$(6.18 \pm 0.016)10^{-8}$	0.722 ± 0.002	0.758 ± 0.0012
Heterogeneous	Zero momentum	$(5.67 \pm 0.022)10^{-8}$	0.727 ± 0.0019	0.773 ± 0.0012
Uniform	Retained	0.33 ± 0.01	0.533 ± 0.00045	0.532 ± 0.0015
Uniform	Transported	$(6.16 \pm 0.02)10^{-8}$	$0.485 \pm 2.9 \times 10^{-14}$	0.486 ± 0.00018
Uniform	Zero momentum	$(5.66 \pm 0.014)10^{-8}$	$0.485 \pm 2.2 \times 10^{-13}$	0.487 ± 0.00026

The uniform-scale experiment crosses doses $\pm\{0.125, 0.25, 0.5\}$ with exact polar, FP64 NS5, normalized momentum, and momentum on three populated checkpoints. We match zero-dose update norms separately for each logical matrix and freeze the factors. The $T = 16$ KLs average 32 fixed contexts; canonical rows use only the positive dose. Exact compensation requires covariant recurrences, compatible decay multipliers, and the stated preprocessing assumptions. For compensated exact polar and FP64 NS5, the maximum relative discrepancy over weights and all buffer families is 5.77×10^{-12}. BF16 NS5 compensation gives mean $\mathrm{KL}_{16} = 3.88 \times 10^{-3}$ as a separate finite-precision control.

Table 6 reports the paired endpoint responses. At $|s| = 0.5$, unchanged-rule commutators agree with the homogeneous predictions within 5.33×10^{-15}. Over 16 steps, rate compensation reduces mean

KL from 4.50×10^{-3} and 4.96×10^{-3} to the numerical floor for exact polar and FP64 NS5. For normalized momentum, KL falls from 0.129 to 5.50×10^{-4}; for ordinary momentum, it falls from 0.607 to 0.012. Each KL compares a branch with its matched control.

The canonical control $\Psi_t = \mathcal{C} \circ \Phi_t \circ \mathcal{C}$, with $\mathcal{C}(S) = \rho_{-\theta(W)/2} S$, identifies gauge-equivalent transported states in exact arithmetic. In floating point, the first mismatch occurs at canonicalization itself ($T = 0$), with relative weight differences of 7.39–7.47×10^{-17}. Copies of an identical canonical full state remain bitwise identical through 64 steps under all four rules. The nonzero momentum residuals in Table 6 are consistent with amplification of this roundoff-scale initial difference.

Table 6: KL entries are mean $\pm$ sample SD across three training seeds. The two dose signs are averaged within each seed before aggregation; canonicalization uses the positive dose. Weight errors retain maxima over the original cases, and floor means all relevant seed-level $|\mathrm{KL}| < 10^{-16}$. $|s| = 0.5$, $T = 16$. Comp. denotes rate compensation, Canon. denotes canonicalization, and W denotes weights.

Rule	Unchanged KL	Comp. W max	Comp. KL	Canon. W max	Canon. KL
Exact polar	0.0045 ± 0.00083	6.83×10^{-14}	floor	2.88×10^{-14}	floor
NS5	0.00496 ± 0.001	4.66×10^{-14}	floor	1.71×10^{-14}	floor
Normalized mom.	0.129 ± 0.021	0.001083	$(5.5 \pm 7.3)10^{-4}$	0.002187	0.00124 ± 0.0021
Momentum	0.607 ± 0.064	0.008159	0.012 ± 0.018	6.07×10^{-4}	$(2.74 \pm 3)10^{-4}$

The canonical audit covers seeds 0–2 and four rules at uniform dose 0.5, tracing $T = 0, \ldots, 64$ with eight validation windows 2048–2055. Control and intervened states are separately canonicalized, and an additional branch copies the exact canonical control state. All twelve jobs first differ at $T = 0$ in weights and Muon buffers, with maximum absolute state difference 2.22×10^{-16}; auxiliary moments agree initially. Identical-state branches remain exactly equal. The archive contains 780 step records and 131820 state-block records, plus common input/label hashes, cursor, and random-state hashes. Table 7 includes the large $T = 64$ momentum outlier. The traces localize the mismatch to the initial canonicalization, with the individual floating-point operation unresolved.

Table 7: Mean $\pm$ sample SD across three training seeds. KL is first averaged over eight windows within seed. First mismatch is an event index; identical-state checks remain counts. The momentum T64 maximum is 853.4, retained in the seed-level records; its large spread is not suppressed. W rel denotes relative weight error, T denotes continuation steps, and First denotes the first mismatch step.

Rule	First	W rel T0	W rel T16	W rel T64	KL T64	Identical
momentum	0	$(7.44 \pm 0.042)10^{-17}$	$(2.4 \pm 3.2)10^{-4}$	$285 \pm 4.9 \times 10^{2}$	0.332 ± 0.17	3/3
normalized	0	$(7.44 \pm 0.042)10^{-17}$	0.000765 ± 0.0012	0.184 ± 0.011	0.206 ± 0.034	3/3
ns5	0	$(7.44 \pm 0.042)10^{-17}$	$(1.46 \pm 0.22)10^{-14}$	$(3.65 \pm 2.7)10^{-6}$	$(1 \pm 1.3)10^{-9}$	3/3
polar	0	$(7.44 \pm 0.042)10^{-17}$	$(2.46 \pm 0.4)10^{-14}$	$(1.15 \pm 1.7)10^{-6}$	$(5.44 \pm 9.3)10^{-11}$	3/3

E.3 Derivative checks and mechanism coverage

The 500-step derivative check uses three fixed zero-mean unit-RMS directions on twenty up/down matrices at each of three 500-step checkpoints (180 cases). At $h = 5 \times 10^{-4}$, exact polar and NS5 agree with their own independent derivatives to the errors in Table 8. The corresponding matched-norm NS5–polar median difference is 0.612, mean exact-polar column-space energy share is 72.1%, and maximum exact-polar energy residual is 2.44×10^{-12}. That polar energy share uses a different denominator from the derivative-difference shares in Table 1.

The real-matrix mechanism evaluation uses total training cursors 500, 4500, 16500, and 50500 for seeds 0–2: twenty matrices per checkpoint, 32 zero-mean unit-RMS random Gaussian directions, and three additional directions per matrix. Table 1 reports 7680 random cases; 240 imbalance, 240 anti-imbalance, and 240 uniform cases give 8400 in total. Imbalance is the centered, unit-RMS channel vector $\log \|W_{U,i:}\| - \log \|W_{D,:i}\|$ and anti-imbalance is its negative. Opposite directions

Table 8: Smooth FP64 checks at $h = 5 \times 10^{-4}$. The two LM rows report mean $\pm$ sample SD across three training seeds; each seed first takes the median derivative error and RMS finite-difference quotient over 60 matrix–direction cases. Synthetic cases remain individual checks. Uniform polar uses absolute error; other rows use relative error.

Condition	Derivative error	RMS quotient norm
Uniform exact polar	1.26×10^{-12}	1.26×10^{-12}
Tall, equal spectrum	7.95×10^{-8}	1.4
Invariant column space	3.41×10^{-9}	0.167
LM exact polar	$(1.95 \pm 0.0041)10^{-7}$	18.3 ± 0.063
LM NS5, own JVP	$(6.42 \pm 0.36)10^{-7}$	20.4 ± 0.057

have opposite derivatives and equal energies, so they are not independent replications. The Frobenius norm-matching coefficient is frozen at zero dose.

Pooling random cases descriptively gives median relative difference 0.6459 and off-diagonal energy share 71.49%, versus 0.6779 and 54.45% for imbalance. Uniform difference energy is at most 3.609×10^{-24}; its relative ratios are not interpreted because the reference derivative is numerically zero. Maximum $\|q\|_\infty$ is 4.605 for random and 6.638 for imbalance directions; the archive records $\exp(0.125\,\|q\|_\infty)$, since equal RMS dose does not equalize maximum channel scaling. Layer/checkpoint variation appears in Table 9.

Table 9: Mean $\pm$ sample SD across three training seeds. Within each seed/layer/checkpoint, the relative derivative difference is first reduced by the median over two matrix sides and 32 random directions (64 cases).

Layer	500	4500	16500	50500
0	0.68 ± 0.014	0.655 ± 0.027	0.691 ± 0.027	0.702 ± 0.039
1	0.558 ± 0.031	0.634 ± 0.013	0.973 ± 0.13	1.14 ± 0.071
2	0.518 ± 0.025	0.647 ± 0.037	1.03 ± 0.074	1.31 ± 0.17
3	0.503 ± 0.017	0.631 ± 0.014	0.988 ± 0.11	1.29 ± 0.14
4	0.538 ± 0.0047	0.623 ± 0.034	0.842 ± 0.064	1.1 ± 0.035
5	0.57 ± 0.0023	0.575 ± 0.019	0.788 ± 0.16	0.952 ± 0.061
6	0.614 ± 0.038	0.563 ± 0.018	0.579 ± 0.031	0.729 ± 0.067
7	0.67 ± 0.017	0.554 ± 0.01	0.536 ± 0.0051	0.607 ± 0.052
8	0.732 ± 0.024	0.562 ± 0.0099	0.524 ± 0.015	0.513 ± 0.0047
9	0.792 ± 0.046	0.67 ± 0.013	0.674 ± 0.024	0.662 ± 0.028

Across the twelve seed/checkpoint groups, median relative derivative differences range from 0.5752 to 0.9728. Off-diagonal response accounts for 65.3–82.1% of summed difference energy, normal motion for 9.30–20.31%, and amplitude response for 8.18–17.84%. These matrix-space measurements show that finite NS differs through its vector response as well as its amplitude channel; downstream KL additionally depends on state propagation and observable projection.

The twelve checkpoint-group mean absolute difference energies range from 96.51 to 130.5 in squared Frobenius units. The maximum NS matrix-recurrence/AD discrepancy on random cases is 5.04×10^{-15}, the centered path-FD error at $h = 10^{-5}$ is at most 2.560×10^{-8}, and maximum relative reconstruction/energy-closure residuals are $2.36 \times 10^{-17}/2.58 \times 10^{-13}$. Reconstruction divides by the difference norm and energy closure by its squared norm, each floored at 10^{-30}. Raw component energies and 960 layer/checkpoint/direction groups remain in the experiment record.

For the controlled scan, layers 0, 5, and 9 contribute both up/down matrices at each seed's 500-step checkpoint. With u_i evenly spaced on $[0, 1]$, the seven spectra are original; flat 1; $\exp(-2u)$, $\exp(-6u)$, $\exp(-12u)$; gap (1 for $u < 0.5$, 0.01 otherwise); and tail ($\exp(-u)$ for $u < 0.9$, 10^{-6} otherwise). Each is rescaled to the original singular-value two-norm with singular vectors fixed. There are $3 \times 6 \times 7 \times 35 = 4410$ cases, including 576 random cases per shape. The numerical repeated-block criterion is $|\sigma_i - \sigma_j| \le 10^{-10}\sigma_{\max}$, as specified in Appendix B.2.

Changing the spectrum changes the source of the discrepancy. Table 10 shows how these controlled spectra change the source of the discrepancy. For distinct exponential spectra, between-block response contributes 95.27% for decay2 and 55.51% for decay12, while normal response rises from 1.83% to 41.9%. Flat and gapped spectra instead concentrate the difference inside repeated-value blocks (100% and 91.85%). Equal input singular values therefore do not eliminate the finite-NS response: heterogeneous scaling can split their output amplitudes. The diagonal/off-diagonal attribution within such a block depends on its basis, whereas the block energy is invariant. The tail spectrum again has substantial between-block energy (88.71%). These controlled results identify how the input spectrum changes the source of the derivative difference. The 4410 scan cases include the same three additional direction types.

Table 10: Mean $\pm$ sample SD across three training seeds. Each seed first aggregates 192 random-direction matrix cases per spectrum. Relative difference is a seed-level median; shares are ratios of energy sums. Within-block energy includes the entire repeated-value block; singleton blocks contribute diagonal entries.

Spectrum	Rel. difference	Normal	Within block	Between blocks
decay12	0.926 ± 0.006	0.419 ± 0.0029	0.0259 ± 0.003	0.555 ± 0.00065
decay2	1.01 ± 0.0086	0.0183 ± 0.00028	0.029 ± 0.0022	0.953 ± 0.0019
decay6	0.676 ± 0.006	0.225 ± 0.0024	0.0342 ± 0.0029	0.741 ± 0.00063
flat	2.22 ± 0.028	$(4.43 \pm 0.027)10^{-26}$	$1 \pm 7.4 \times 10^{-16}$	0 ± 0
gap	1.85 ± 0.014	0.0538 ± 0.00072	0.918 ± 0.0013	0.0278 ± 0.00055
real	0.681 ± 0.012	0.212 ± 0.0045	0.0764 ± 0.028	0.712 ± 0.023
tail	1.26 ± 0.011	0.103 ± 0.00035	0.0104 ± 0.00024	0.887 ± 0.00016

E.4 Predictive-response cohorts, calibration, and memory

The larger-dose FP64 finite-difference cohort uses seed 0, direction 1729, and sixteen windows for development, and seeds 1–2, directions 2718/31415, and sixteen disjoint windows for evaluation. The split was imposed after execution. Each state/horizon has 24 seed–direction–signed-dose cells for $s = \pm\{0.125, 0.25, 0.5\}$; these are not 24 training seeds. Separate nonnegative coefficients map gauge-block κ^2 and squared defect norm to KL using development data only; two calibration-run proxy measurements were reconstructed retrospectively. At $T = 1, 4$, maximum derivative errors against independent paired pulses at $h = 10^{-5}$ are 1.35×10^{-9} and 1.87×10^{-3}; finite-dose errors remain distinct from derivative error.

The larger-dose exploratory cohort compares the Jacobian/Fisher readout with scalar proxies fitted on development data. At $T = 1$, pooled transported-state median relative KL error falls from 1.53 for calibrated κ^2 and 1.18 for calibrated squared defect norm to 0.428 (Table 11). Median absolute error falls from 3.34×10^{-6} to 1.93×10^{-6} relative to κ^2, while rank correlation is lower (0.919 versus 0.959). The readout adds magnitude information, while the calibrated commutator retains the higher rank correlation in this comparison. At $T = 4$, its transported median absolute error is 8.24×10^{-6} versus 2.91×10^{-5} for squared defect norm. This cohort uses a retrospective development/evaluation split and larger doses; its protocol and numerical checks are kept separate from the independent confirmation.

The two pre-fixed channel directions are independently permuted, zero-mean ± 1 directions in each layer and are shared across calibration and evaluation. Full-state AD follows the fixed control trajectory using Eq. (21).

The independent AD confirmation fixes seed 200 and windows 2048–2063 for calibration; seeds 201–203 and windows 4096–4111 for evaluation; and direction seeds 93001/93002 for both splits. It tests both state conventions at $T = 1, 4$ and six doses $\pm\{0.03125, 0.0625, 0.125\}$. Factors are median actual/raw-predicted KL on seed 200, fixed separately by state and horizon. Transported/retained factors are $1.045678/1.203111$ at one step and $1.023103/1.083626$ at four steps. There are 3072 window–dose records including calibration, and 576 evaluation records per state/horizon. Calibration affects KL only.

Table 11: Exploratory cohort: mean ± sample SD across the two evaluation training seeds ($n = 2$). Each seed metric uses two directions and six doses after averaging sixteen held-out windows. The retrospective split and seed-0-only fits are unchanged. The commutator uses only transformed up/down blocks; this differs from the full-weight probe baseline. Relative error is a mean of seed medians with floor 10^{-12}; pooled medians in the text remain separate. Loss MAE units are 10^{-4} nats/token. C/R denote transported/retained optimizer state.

Predictor	State/T	Rel. error	ρ	Loss MAE
Gauge-block κ^2	C/1	1.47 ± 0.088	0.948 ± 0.025	–
Squared defect	C/1	1.11 ± 0.14	0.843 ± 0.025	–
Jacobian/Fisher	C/1	0.421 ± 0.093	0.94 ± 0.03	2.91 ± 0.75
Full state	C/4	0.266 ± 0.035	0.898 ± 0.03	8.77 ± 1.7
Forcing only	R/4	0.426 ± 0.0088	0.898 ± 0.01	4.15 ± 1.4
Forcing + memory	R/4	0.547 ± 0.013	0.912 ± 0.01	3.62 ± 2.2

Pooled recordwise median relative KL errors are $0.07365/0.05728$ for transported state and $0.1890/0.09872$ for retained state at $T = 1/4$, with rank correlations 0.9574–0.9918. Multiplicative factors, fitted on seed 200 and locked before evaluation, lower absolute MAE in all four strata, but do not uniformly lower median relative error. Positive scalar calibration cannot change ranking. Signed-loss AD MAE is lower than the zero-prediction baseline in all four strata, with a small retained one-step improvement (Table 12). The loss readout uses no KL calibration. Probe resampling preserves the transported one-step relative-error advantage across all seed–budget means, but not uniform MAE superiority (Appendix E.4).

Table 12: Mean ± sample SD across three training seeds. Within each seed, signed-loss MAE uses 192 window–dose records. Units are 10^{-4} nats/token. Zero predicts no loss change; KL calibration is not applied. C/R denote transported/retained optimizer state.

T	State	AD MAE	Zero MAE
1	C	0.219 ± 0.019	0.425 ± 0.067
1	R	0.719 ± 0.026	0.756 ± 0.039
4	C	0.653 ± 0.0095	1.55 ± 0.33
4	R	0.751 ± 0.099	1.61 ± 0.22

Memory terms are evaluated at common control probabilities: $F_{\mathrm{total}} = \|a + b\|_F^2$, $F_{\mathrm{forcing}} = \|a\|_F^2$, $F_{\mathrm{memory}} = \|b\|_F^2$, and $F_{\mathrm{cross}} = 2\langle a, b\rangle_F$, with the tangents from Eq. (45). At $T = 1/4$, forcing Fisher means are $8.385 \times 10^{-5}/1.359 \times 10^{-3}$ in both state conventions; transported memory is zero. Fisher entries are deduplicated by job/horizon/window, yielding 96 evaluation records per state/horizon. Table 13 compares raw full predictions $s^2 F_{\mathrm{total}}/2$ to forcing-only predictions $s^2 F_{\mathrm{forcing}}/2$ for every signed dose. The one-step cross term is negative in all 96 windows; at four steps it spans $[-2.923 \times 10^{-3}, 1.077 \times 10^{-4}]$. Because the local quadratic prediction is even in dose while actual finite-dose KL need not be, a signed-dose error difference cannot be assigned to the cross term alone.

Memory self and cross terms alter the local response, while the approximation error also depends on dose. In the exploratory retained $T = 4$ cohort, adding the initial-memory term improves ranking (0.875 to 0.928) and signed-loss MAE (4.15×10^{-4} to 3.62×10^{-4}), but worsens relative KL error (0.432 to 0.556). In the independent confirmation, mean memory self energies at $T = 1/4$ are $2.977 \times 10^{-5}/3.504 \times 10^{-4}$ and signed cross terms are $-4.244 \times 10^{-5}/-4.772 \times 10^{-4}$. Their sums reduce the mean total Fisher response relative to forcing alone. These measurements use 96 distinct seed–direction–window records per state/horizon after removing repeated doses.

The case-level comparison also depends on dose: the full retained predictor improves over forcing alone in 74/96 one-step cases at $s = +0.03125$, but only 11/96 at $s = -0.125$. At $T = 4$, the counts are 71/96 and 75/96 at the two smallest signed doses, versus 39/96 at either largest dose. Table 13 gives the complete signed-dose comparison. The cross terms explain how memory changes the local

response; finite-dose remainders and estimation error determine whether that exact state component improves a particular approximation.

Table 13: Mean ± sample SD across three training seeds. Each seed first averages 32 direction–window errors at each fixed dose. Full-better remains the total count of strictly smaller absolute errors across 96 cases, not a seed-level significance test.

T	s	Full better	Full KL MAE	Forcing KL MAE
1	-0.125	11/96	$(5.81 \pm 1.1)10^{-7}$	$(4.82 \pm 0.43)10^{-7}$
1	-0.0625	16/96	$(4.44 \pm 0.74)10^{-8}$	$(2.47 \pm 0.72)10^{-8}$
1	-0.03125	47/96	$(4.07 \pm 0.41)10^{-9}$	$(4.48 \pm 2.1)10^{-9}$
1	0.03125	74/96	$(2.74 \pm 1.7)10^{-9}$	$(7.65 \pm 2.8)10^{-9}$
1	0.0625	56/96	$(2.89 \pm 0.92)10^{-8}$	$(3.43 \pm 1)10^{-8}$
1	0.125	12/96	$(4.35 \pm 0.049)10^{-7}$	$(3.42 \pm 0.83)10^{-7}$
4	-0.125	39/96	$(3.26 \pm 0.29)10^{-6}$	$(3.04 \pm 0.83)10^{-6}$
4	-0.0625	57/96	$(3.07 \pm 0.93)10^{-7}$	$(4.29 \pm 1.6)10^{-7}$
4	-0.03125	71/96	$(3.43 \pm 1.4)10^{-8}$	$(1.02 \pm 0.54)10^{-7}$
4	0.03125	75/96	$(3.49 \pm 1.5)10^{-8}$	$(1.09 \pm 0.72)10^{-7}$
4	0.0625	56/96	$(2.99 \pm 0.94)10^{-7}$	$(4.92 \pm 2.6)10^{-7}$
4	0.125	39/96	$(3.28 \pm 0.54)10^{-6}$	$(3.45 \pm 0.56)10^{-6}$

The recorded norm gain is $g_{\text{state}} = \|\xi_T\| \,/\, \max(\|\xi_0\|, 10^{-30})$, approximately one in the confirmation. It is not the observable gain $g(\delta)$ in Eq. (44). Complete records distinguish the confirmation and convergence cohorts and do not merge repeated measurements indexed by different h.

Probe-set size and selection robustness. The probe robustness evaluation changes only the terminal probe set. It reuses the same 106.57M model, populated 500-step checkpoints, training batches, FP64 realization, state conventions, horizons $T = 1, 4$, directions 93001/93002, and six signed doses as the independent confirmation. All four previously locked Jacobian/Fisher factors remain unchanged. We fix five probe-sampling seeds, 94001–94005, before evaluation. For each seed, we uniformly sample 64 windows without replacement from the same 25000-window validation split, excluding calibration windows 2048–2063 and the confirmation evaluation windows 4096–4111. The first $B \in \{1, 4, 16, 64\}$ windows define nested budgets within each draw; the five realized draws contain 320 distinct windows. Each window contains 256 tokens, so the budgets contain 256, 1024, 4096, and 16384 tokens. The same subsets are used for every method, training seed, direction, state, and horizon; no subset is selected using outcomes. The confirmation evaluation windows are recomputed solely to verify agreement with the independent confirmation and are excluded from these robustness scores.

For each training seed, state, horizon, budget, sampling seed, direction, and dose, we first average both predicted and actual KL over the identical selected windows. We then compute mean absolute KL error and median relative KL error over the 12 direction–dose cases, with relative-error denominator $\max(|\overline{K}_X|, 10^{-20})$. Tables 14–16 report the mean and sample standard deviation (denominator $3 - 1$) across the three training-seed means of those five draw-level metrics. Within-seed sampling SDs (denominator $5 - 1$) remain in the archived per-seed records. Neither windows nor sampling repetitions increase the number of independent training seeds. Reusing the same propagated state tangent across probe sets changes neither the model trajectory nor the intervention.

Only for transported state at $T = 1$, we compare against the full-weight jointly normalized κ^2 and pulled-back squared weight-defect norm $\|\delta_s\|^2$. Each proxy is measured for the same direction and finite dose as the Fisher prediction. Its coefficient is fitted once on seed 200, using the original 16 calibration windows: first average actual KL within each direction–dose case, then fit $c = \sum_i x_i y_i / \sum_i x_i^2$ with zero intercept. Both fitted coefficients are positive: $c_\kappa = 4.9272620401 \times 10^{-5}$ and $c_\delta = 2.4951256776 \times 10^{-6}$. They are locked before held-out evaluation and never refitted for a budget or subset. Here $\kappa = 2\,\|U_s - G_s U_0\| \,/ (\|U_s\| + \|G_s U_0\|)$ uses the original joint weight-update normalization, and $\delta_s = W_1 - G_{-s} W_{1,s}$; optimizer buffers do not enter the defect norm. These baseline coefficients are fitted and locked separately from the independent Fisher calibration; the larger-dose exploratory cohort uses a retrospective split.

Table 14: Transported. Mean ± sample SD across three training seeds. Each seed first averages five fixed probe draws; each draw computes metrics over twelve direction–dose cases after averaging matched windows. Calibration uses seed 200 only. Within-seed draw SDs remain in the archived records. MAE units are 10^{-7} nats/token. Cal. denotes calibrated predictions.

T	B	Raw MAE	Raw rel.	Cal. MAE	Cal. rel.
1	1	0.599 ± 0.18	0.0775 ± 0.036	0.539 ± 0.2	0.0845 ± 0.031
1	4	0.591 ± 0.19	0.0727 ± 0.036	0.538 ± 0.23	0.0845 ± 0.035
1	16	0.59 ± 0.16	0.0736 ± 0.036	0.528 ± 0.19	0.084 ± 0.038
1	64	0.576 ± 0.15	0.0735 ± 0.038	0.513 ± 0.18	0.0838 ± 0.038
4	1	6.17 ± 0.88	0.0709 ± 0.015	5.95 ± 1.1	0.0745 ± 0.01
4	4	5.1 ± 1.4	0.0586 ± 0.0063	4.92 ± 1.3	0.0639 ± 0.0067
4	16	5.25 ± 1.2	0.0639 ± 0.0013	5.19 ± 1	0.0697 ± 0.0034
4	64	4.66 ± 0.69	0.0523 ± 0.014	4.61 ± 0.29	0.0661 ± 0.014

Table 15: Retained. Mean ± sample SD across three training seeds. Each seed first averages five fixed probe draws; each draw computes metrics over twelve direction–dose cases after averaging matched windows. Calibration uses seed 200 only. Within-seed draw SDs remain in the archived records. MAE units are 10^{-7} nats/token. Cal. denotes calibrated predictions.

T	B	Raw MAE	Raw rel.	Cal. MAE	Cal. rel.
1	1	1.8 ± 0.16	0.193 ± 0.027	1.42 ± 0.11	0.186 ± 0.021
1	4	1.88 ± 0.17	0.19 ± 0.024	1.47 ± 0.11	0.18 ± 0.03
1	16	1.89 ± 0.2	0.186 ± 0.035	1.49 ± 0.14	0.186 ± 0.017
1	64	1.87 ± 0.2	0.185 ± 0.035	1.47 ± 0.14	0.183 ± 0.013
4	1	10.8 ± 2.1	0.0847 ± 0.009	8.89 ± 2.1	0.102 ± 0.017
4	4	10.5 ± 2	0.0795 ± 0.0093	8.45 ± 2.3	0.0944 ± 0.02
4	16	10.4 ± 1.6	0.083 ± 0.0016	8.03 ± 1.7	0.0873 ± 0.019
4	64	10.7 ± 1.9	0.0797 ± 0.004	8.09 ± 1.8	0.0827 ± 0.017

Table 16: Mean ± sample SD across three training seeds. Each seed first averages five fixed probe draws; each draw computes metrics over twelve direction–dose cases after averaging matched windows. Calibration uses seed 200 only. Within-seed draw SDs remain in the archived records. Transported state, $T = 1$. Cal. denotes calibrated predictions.

B	Metric	Jacobian/Fisher	Cal. κ^2	Cal. $\|\delta_s\|^2$
1	Rel.	0.0845 ± 0.031	0.197 ± 0.099	0.204 ± 0.11
1	MAE (10^{-7})	0.539 ± 0.2	0.624 ± 0.23	0.626 ± 0.24
4	Rel.	0.0845 ± 0.035	0.155 ± 0.073	0.159 ± 0.084
4	MAE (10^{-7})	0.538 ± 0.23	0.593 ± 0.18	0.594 ± 0.18
16	Rel.	0.084 ± 0.038	0.169 ± 0.065	0.178 ± 0.073
16	MAE (10^{-7})	0.528 ± 0.19	0.584 ± 0.14	0.586 ± 0.14
64	Rel.	0.0838 ± 0.038	0.175 ± 0.071	0.182 ± 0.081
64	MAE (10^{-7})	0.513 ± 0.18	0.564 ± 0.14	0.566 ± 0.15

Observed sensitivity. Increasing the budget from one to 64 windows lowers the sampling SD of calibrated median relative error in all 12 training-seed–state–horizon strata, but neither the mean error nor its SD decreases monotonically at every intermediate budget. At $B = 64$, the five-draw mean relative errors across seeds are 0.04088–0.1116 (transported, $T = 1$), 0.05184–0.08059 (transported, $T = 4$), 0.1684–0.1926 (retained, $T = 1$), and 0.06816–0.1013 (retained, $T = 4$). Retained one-step error therefore persists at larger probe budgets; increasing the probe set does not remove finite-dose or calibration mismatch.

For transported $T = 1$, calibrated Jacobian/Fisher has lower five-draw mean median relative error than both scalar proxies in all 12 training-seed–budget combinations. The mean relative-error ranges are 0.04073–0.1161 for Jacobian/Fisher, 0.07370–0.2612 for calibrated κ^2, and 0.06624–0.2750 for calibrated squared defect norm. This comparison concerns the reported means, not every individual sampled subset. Absolute MAE is lower in 10 of 12 combinations against κ^2 and 11 of 12 against squared defect norm. In particular, seed 201 at $B = 16, 64$ does not uniformly favor Jacobian/Fisher. The relative-error advantage therefore survives the tested probe-size and selection changes, whereas absolute error depends on the training seed and probe budget.

E.5 Convergence and dose range across all tested horizons

Direct AD and explicit forcing plus memory agree within 4.96×10^{-13} through $T = 64$. Agreement between these two routes does not resolve the actual-path derivative at long horizons.

The derivative-step and dose-scan cohort uses seed 0 at step 500, direction 93001, and validation windows 2048–2051. It scans four derivative steps independently of six evaluation doses and five horizons. Table 17 displays every state/step/horizon path-FD comparison; direct-versus-forcing AD discrepancies are summarized over the repeated step index. Both comparisons divide by $\max(\|r_{\mathrm{direct}}\|, 10^{-30})$. Step halving divides $\|r_{\mathrm{pathFD}}(h/2) - r_{\mathrm{pathFD}}(h)\|$ by $\max(\|r_{\mathrm{pathFD}}(h)\|, 10^{-30})$, and the propagated-FD JVP chain divides by $\max(\|\xi_{\mathrm{AD}}\|, 10^{-30})$. At

Table 17: All 40 state/step/horizon diagnostics on seed 0, checkpoint 500, direction 93001. Four middle columns give actual-path FD versus direct AD relative error. The final column is the maximum direct-versus-forcing AD discrepancy over those step sizes. C/R denote transported/retained state. No case is removed for nonconvergence.

State	T	$h = 10^{-3}$	10^{-4}	10^{-5}	10^{-6}	AD routes
C	1	1.591×10^{-6}	1.591×10^{-8}	7.79×10^{-10}	7.421×10^{-9}	0
C	4	0.04861	0.0136	2.563×10^{-4}	2.583×10^{-6}	5.452×10^{-15}
C	8	0.1253	0.05886	0.006879	1.854×10^{-5}	8.529×10^{-15}
C	16	0.389	0.3401	0.1945	0.07816	2.109×10^{-14}
C	64	1	1	1	1	7.903×10^{-14}
R	1	4.888×10^{-6}	4.888×10^{-8}	1.036×10^{-9}	8.884×10^{-9}	8.281×10^{-15}
R	4	0.0544	0.01446	2.558×10^{-4}	2.58×10^{-6}	7.773×10^{-15}
R	8	0.1194	0.04635	0.00191	6.149×10^{-6}	1.209×10^{-14}
R	16	0.5507	0.5319	0.3963	0.1761	3.784×10^{-14}
R	64	1	1	1	1	4.953×10^{-13}

$h = 10^{-6}$, transported/retained step-halving discrepancies are $1.634 \times 10^{-8}/1.956 \times 10^{-8}$ at $T = 1$, $1.939 \times 10^{-6}/1.934 \times 10^{-6}$ at $T = 4$, and $1.390 \times 10^{-5}/4.612 \times 10^{-6}$ at $T = 8$. They remain $0.06024/0.08491$ at $T = 16$ and $1.940/1.873$ at $T = 64$. The propagated-FD/AD discrepancies at the same step are $0.02199/0.04428$ at $T = 16$ and approximately one at $T = 64$. These unresolved path checks are retained even though the two AD routes agree. In a separate three-seed/two-state step-halving check, median paired-pulse step-halving discrepancies were 0.0593 and 1.83 at $T = 16, 64$, likewise leaving the derivative unresolved.

Table 18 summarizes all six doses without ranking long-horizon estimates as ordinary predictor scores. Each endpoint is the minimum or maximum of the six dose-specific median relative errors over four fixed windows, retaining both favorable and unfavorable doses. At $T = 4$, $s = -0.03125$, errors are $0.006685/0.008062$ for transported/retained state, compared with $0.07423/0.1019$ at $s = -0.125$.

At $T = 64$, actual mean KL over those windows ranges approximately 0.036–0.056, while AD-based local predictions range from 7.941×10^{13} to 4.758×10^{16}. The complete 960 window records retain all 240 state–h–T–dose combinations; the AD predictions are indexed by, but not numerically differentiated with, h.

Table 18: Finite-dose validity over all six signed doses, separate from derivative convergence. Columns are the minimum and maximum dose-specific median relative KL errors for each state at the recorded $h = 10^{-4}$ index. Every horizon is retained; long-horizon values mark failure of the local approximation. C/R denote transported/retained optimizer state.

T	C min	C max	R min	R max
1	0.02158	0.239	0.02303	0.4617
4	0.006685	0.07423	0.008062	0.2456
8	0.02292	0.1331	0.03568	0.1351
16	0.617	8.935	0.5044	5.183
64	5.588×10^{16}	8.483×10^{17}	1.725×10^{15}	2.184×10^{16}

E.6 Later checkpoints test the same core diagnostics

The checkpoint-extension validation covers 720 matrix–direction cases from nine intermediate and three later checkpoints, retaining median polar derivative error 1.95×10^{-7}. Separately, three controls continued on a 750M-token FineWeb-Edu shard reach 260k continuation updates after the 500-step prefix, approximately 5.006 cumulative input tokens per parameter. Weights and optimizer history are preserved at the shard transition and the cursor is reset. The shard has 725595 documents, no complete-document token-hash matches against the earlier shard, and no internal duplicate complete-document hashes in the reported audit; it is another shard of the same dataset configuration.

Repeating the matrix and predictive-response diagnostics at these checkpoints gives accurate exact-polar derivatives, while response sensitivities again lose convergence beyond the shortest horizon. These results extend training-stage coverage within the same model; the available records report these outcomes without a per-seed error breakdown.

E.7 Numerical controls and implementation assumptions

The numerical evaluation runs duplicate, sham, pulse–inverse round trip, three norm-matched random directions, compensation, canonicalization, and unchanged-rule comparisons for 64 steps in each precision stratum. Each nonrandom row in Table 19 averages three paired seeds over 32 fixed contexts; random directions are averaged within seed first. All 108 short continuations have finite updates. BF16/FP16 store parameters and optimizer state in FP32 and use low-precision NS5 with autocast. FP32-state controls also include a high-precision RMSNorm reference. Random perturbations match each stored tensor's round-trip error norm, not the predictive-error distribution.

Table 19: Mean $\pm$ sample SD across three training seeds. Endpoint KL at $T = 64$ is first averaged over probe windows; the random family additionally averages three directions within each seed. Each precision remains separate. Floor cells indicate all three seed means below 10^{-15} in magnitude.

Control	FP64	FP32	BF16	FP16
Duplicate	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$
Zero-dose sham	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$
Pulse–inverse	$(3.17 \pm 3.9)10^{-9}$	0.0168 ± 0.0048	0.0402 ± 0.0017	0.0492 ± 0.0083
Matched random	$(1.14 \pm 1.8)10^{-8}$	0.0175 ± 0.0042	0.0404 ± 0.0021	0.048 ± 0.005
Rate compensated	$(3.03 \pm 5.2)10^{-8}$	0.0169 ± 0.0023	0.0415 ± 0.00042	0.0473 ± 0.0055
Canonicalization	$(4.09 \pm 7.1)10^{-11}$	0.0138 ± 0.0026	0.0397 ± 0.0017	0.051 ± 0.0026
Unchanged rule	0.0458 ± 0.002	0.0451 ± 0.0017	0.0465 ± 0.0027	0.0584 ± 0.0074

Duplicate and zero-dose sham runs agree, while a pulse–inverse round trip changes the numerical path despite preserving the exact function. At $T = 64$, the FP64 round-trip KL is 3.17×10^{-9} versus 0.0458 for the unchanged-rule pulse; in BF16 these values are 0.0402 and 0.0465 (Table 19). Thus a low-precision endpoint difference can be comparable to a numerical null and cannot, by itself, establish the heterogeneous mechanism. The FP64 matrix-derivative checks provide independent evidence for that mechanism.

Compensated and canonicalized branches change the update rule, while random controls match stored-tensor error norms. Their endpoints therefore characterize distinct numerical comparisons; they are not additive components that can be subtracted from the pulse response.

The round-trip initial maximum relative error across weight/buffer families is 3.56×10^{-17} in FP64 and 2.13×10^{-8} with FP32 state. Immediate relative logit errors are 3.86×10^{-16} (FP64), 1.89×10^{-7} (FP32), 1.65×10^{-3} (BF16), and 2.10×10^{-4} (FP16). The FP64 implementation baseline repeats sixteen steps from seed 0 at step 500 and seeds 0–2 at total step 16500: duplicate states and KL agree exactly across all 64 steps, while original-versus-functional weight discrepancy is at most 2.961×10^{-14}. This establishes a reference floor for the implementation comparison.

The exact-polar rank check uses $\sigma_{\min}/(\sigma_{\max} \max(m, n)\epsilon_{\mathrm{FP64}})$; its minimum over the real-matrix groups is 2.209×10^{7}. Minimum normalization-floor distance is 0.001292, and no sampled $\pm h$ pair crosses the floor. Forty-five NS tests at repeated spectra, rank deficiency, and below-floor inputs give maximum recurrence/AD relative discrepancy 1.177×10^{-15}. Twenty-four full-step functional-map checks have maximum relative discrepancy 8.422×10^{-17}.

For auxiliary Adam entries with zero second moment, the implementation assigns derivative zero along the identically zero state and rejects a tangent leaving that stratum. This is a restricted directional convention, not ambient differentiability of $\sqrt{v}$ at zero. The recorded branch checks cover the NS floor and these second-moment tangents, not all operators throughout training. The smooth derivative statements use the high-precision realization of these real-valued maps; quantized execution is assessed by the paired finite-precision controls.

E.8 Synthetic algebra checks

The deterministic NumPy float64 checks use seed 93711, small matrices, and a two-factor scalar model. They complement the proofs and model experiments by testing the algebraic identities in synthetic settings.

Table 20: Deterministic *synthetic* algebra checks in float64. These use small matrices and a scalar model. FD denotes centered finite differences.

Check	**observed error**
Homogeneity identity, maximum absolute error	3.47×10^{-16}
Heterogeneous polar derivative, maximum relative FD error	3.78×10^{-10}
Polar energy decomposition, maximum absolute error	2.66×10^{-15}
Overall-scale invariance of polar response, relative error	1.11×10^{-15}
Uniform polar derivative, norm	1.61×10^{-15}
Finite-horizon covariant response, relative error	8.40×10^{-11}
Finite-horizon retained response, relative error	2.04×10^{-10}
Uniform compensated trajectory, maximum relative error	6.84×10^{-16}
Canonicalized trajectory, maximum relative error	2.43×10^{-17}
Bilinear Taylor remainder identity, absolute error	7.51×10^{-17}
KL quadratic coefficient at dose 10^{-3}, relative error	1.87×10^{-2}

The KL coefficient error decreases across the tested small-dose sequence; unlike the exact algebraic equalities, it is a finite-dose asymptotic approximation. Numerical checks can catch implementation mistakes.

The scalar counterexample also reproduces log–log slopes 2 and 1.03 for function change, and 4 and 2.07 for KL, in balanced and imbalanced cases. These checks support the different local orders in Proposition A.2.

F Component Propagation and Direction Coverage

F.1 Fixed-trajectory component intervention

The component experiment tests whether the finite-NS derivative differences in Section 5.2 survive propagation to the observable. All variants use the same NS5 control states S_t, state Jacobians A_t, training batches, scalar learning rates, and terminal probe windows. For each up/down matrix at each control step, we set

$$\alpha_t = \frac{\|\mathcal{N}_5(Z_t)\|_F}{\|\mathcal{P}(Z_t)\|_F}$$

and hold α_t fixed under the channel perturbation. This is a zero-dose match at each local control point. The reference trajectory remains NS5 throughout.

Let $\Delta_{N,t}, \Delta_{A,t}, \Delta_{O,t}$ denote the normal, amplitude, and off-diagonal components of $L_{\mathrm{NS5},t} - \alpha_t L_{\mathcal{P},t}$. After applying the transposed down-side convention and the actual matrix learning rate, their state forcings have weight blocks $c^W_{j,t} = -\eta_t \Delta_{j,t}$ and zero direct buffer blocks. Each component follows

$$r_{j,0} = 0, \qquad r_{j,t+1} = A_t r_{j,t} + c_{j,t}, \qquad g_j = Dz(S_T) r_{j,T}, \quad j \in \{N, A, O\}. \tag{50}$$

Forcing in a buffer block can be zero even though the propagated component later enters that buffer through A_t. We verify the sum of the analytic original-rule forcing blocks against $A_t K_q S_t - K_q S_{t+1}$ at every step.

For the one-step comparison in Figure 1, the preprojection energies are $\|c^W_{j,0}\|_F^2$, summed over matrices after the actual learning rates. This removes matrix-rate weighting as an explanation of the subsequent change in self-energy fractions under the terminal Jacobian/Fisher map. The unweighted matrix derivative energies are retained separately in the data archive. Multi-step component responses additionally contain state propagation and cross-time interference.

Let g_{tr} be the complete transported logit tangent and $g_{\mathrm{mem}} = g_{\mathrm{ret}} - g_{\mathrm{tr}}$ the propagated initial-memory difference. Define

$$g_B = g_{\mathrm{tr}} - g_N - g_A - g_O.$$

The polar-shape surrogate uses g_B and therefore retains the NS5 outer pullback while replacing its local shape derivative by the norm-matched polar derivative. The polar-plus-amplitude surrogate uses $g_B + g_A$. Component omissions subtract g_N, g_A, or g_O from the complete response. Retained versions add the same g_{mem} to every surrogate, so the matrix comparison does not change the initial-history convention.

We also report a separate original-rule polar surrogate. In the signed tall-matrix convention, it subtracts the propagated forcing $\eta_t D[\mathcal{N}_5(Z_t) - \alpha_t \mathcal{P}(Z_t)]$ from g_B. This changes both terms of the local original-rule bridge. It still uses the NS5 control states and Jacobians, so it is a local forcing comparison rather than the endpoint of a separately trained polar optimizer.

For $i, j \in \{B, N, A, O, \mathrm{mem}\}$, we retain the full signed Gram matrix

$$H_{ij} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} g_i(x)^\top F_x g_j(x).$$

The transported response sums the B, N, A, O block; the retained response sums all five rows and columns. Cross terms are counted twice when the upper triangle is displayed. The fractions $H_{jj} / \sum_{k \in \{N,A,O\}} H_{kk}$ describe the distribution of projected *self energies*; their denominator excludes cross terms, and they are not additive KL contribution fractions.

F.2 Prespecified coverage and error units

The experiment reuses the independent 500-step checkpoints for training seeds 201–203 and the existing 16500-step checkpoints for seeds 0–2. Each checkpoint receives four new balanced Rademacher

channel directions, generated with seeds 95011–95014 and the same per-layer construction as the independent confirmation. These directions were not used by the original calibration. Both state conventions use $T \in \{1, 4\}$, doses $s \in \pm\{0.03125, 0.0625, 0.125\}$, and the sixteen validation windows 8192–8207. The windows are disjoint from the calibration, original confirmation, and previous probe-resampling windows. No coefficient is fitted on the new evaluation outcomes. The two checkpoint groups use different training seeds; their comparison extends coverage but does not isolate training time as a causal factor.

We distinguish local information loss from finite-dose prediction error. For a surrogate logit tangent $\widehat{g}$, the relative Fisher error is

$$E_F(\widehat{g}) = \sqrt{\frac{\mathbb{E}_x[(\widehat{g} - g)^\top F_x(\widehat{g} - g)]}{\mathbb{E}_x[g^\top F_x g]}}. \tag{51}$$

Here g is the complete tangent under the same state convention. The expectation includes all tokens in the sixteen windows. Separately, we average predicted and measured KL over those windows for each seed–direction–dose case, then report median relative error with denominator $\max(|\mathrm{KL}|, 10^{-20})$, MAE, and Spearman correlation. Signed-loss error is compared with a zero predictor. Tables summarize the three training seeds by mean and sample SD after their stated within-seed reductions; four directions and sixteen windows do not constitute additional independent training seeds. Seed-level records remain in the data archive. Recordwise summaries are also retained for comparison with Table 2.

The main protocol checks centered actual-path derivatives at $h = 10^{-4}$ and $h/2$, independently of the intervention doses. Some four-step cases retain appreciable truncation error at these resolutions. A separate diagnostic, fixed after that observation, checks every one of the 24 prespecified checkpoint–direction combinations at $h \in \{10^{-5}, 10^{-6}\}$ on the first four probe windows. The late-stage responses require a further check at $h \in \{10^{-7}, 10^{-8}\}$, applied to all twelve late checkpoint–direction cases. These checks neither change the intervention doses nor select cases by prediction quality. Full records retain every resolution and every failure to resolve the derivative. A seed-200, direction-95000 smoke case is used only for implementation checks and is excluded from evaluation.

Table 21: Mean ± sample SD across three training seeds. Within each seed and horizon, Gram entries are averaged over sixteen windows and four directions before computing each share. All columns are in %. N/A/O divide by the sum of the three self energies; cross divides the signed doubled cross-term sum by that same denominator. These are not additive KL fractions. N, A, and O denote normal, amplitude, and off-diagonal components, respectively.

Step	T	N	A	O	Cross
500	1	0.0503 ± 0.0096	49.9 ± 9.1	50 ± 9	-2.99 ± 8.3
500	4	0.0579 ± 0.0091	42.9 ± 6.7	57.1 ± 6.7	3.45 ± 4.3
16500	1	0.848 ± 0.62	23.4 ± 0.58	75.7 ± 0.87	12.5 ± 7.4
16500	4	0.758 ± 0.46	28.4 ± 23	70.8 ± 23	-8.52 ± 16

Numerical resolution. At $T = 4$ and the 500-step checkpoints, the maximum logit derivative errors at $h = 10^{-6}$ are 1.51×10^{-7}/3.72×10^{-6} for transported/retained state. The later checkpoints are more sensitive: at $h = 10^{-7}$ the corresponding maxima are 0.003045/9.74×10^{-4}, but decreasing the step to 10^{-8} raises them to 0.0256/0.009046. Thus the smallest tested step is not automatically the most accurate reference. These derivative checks use four windows per case and do not validate the much larger intervention doses, whose later four-step predictions fail in Table 23.

F.3 Measured diagnostic cost

The existing timing jobs use an H200 on each of two hosts, the same 106.57M checkpoint, deterministic FP64 arithmetic, synchronized wall time, warm-up, and five repetitions per repeated phase. Their recorded target GPU has no compute process before the benchmark and only the benchmark process at completion. The records include these endpoint occupancy snapshots, but no continuous

Table 22: Mean ± sample SD across three training seeds. Each seed first takes the median relative Fisher error over four directions; each direction averages sixteen-window energies before taking the ratio. Units are %. P is polar shape with the NS5 pullback retained; P+A adds amplitude correction, and P rule replaces both local original-rule terms. All retained variants keep the same memory term. C/R denote transported/retained optimizer state. N, A, and O denote normal, amplitude, and off-diagonal components, respectively.

Step	State	T	P	P+A	No N	No A	No O	P rule
500	C	1	39.6 ± 2.1	29.1 ± 0.6	0.912 ± 0.029	27.9 ± 4.5	29 ± 0.62	42.4 ± 2.7
500	C	4	31.4 ± 0.97	23.9 ± 2.2	0.761 ± 0.044	20.3 ± 1.9	23.7 ± 2	34.1 ± 1.7
500	R	1	42.1 ± 2.6	31.1 ± 1.1	0.979 ± 0.053	29.5 ± 5.2	31 ± 1.1	45.7 ± 2.5
500	R	4	34 ± 3.8	25.5 ± 3.1	0.823 ± 0.043	21.6 ± 3.8	25.4 ± 3	36.9 ± 4.9
16500	C	1	29 ± 13	26.1 ± 12	2.47 ± 0.2	14.2 ± 5.9	25.7 ± 12	32.5 ± 16
16500	C	4	31.6 ± 11	25.6 ± 12	2.96 ± 1.6	20.8 ± 12	23 ± 10	47.8 ± 34
16500	R	1	39 ± 20	34.3 ± 17	3.24 ± 0.46	20 ± 9.7	33.6 ± 17	45 ± 23
16500	R	4	34.9 ± 19	35.5 ± 21	4.08 ± 2.6	39.4 ± 31	34.7 ± 25	53.3 ± 22

Table 23: Mean ± sample SD across three training seeds. Each seed median uses four directions and six doses after averaging sixteen windows. Entries are means of seed medians, distinct from the pooled medians quoted in the text. The large late four-step errors remain visible. C/R denote transported/retained optimizer state. N, A, and O denote normal, amplitude, and off-diagonal components, respectively. Full uses the complete state response; P uses polar shape with the NS5 pullback, P+A adds amplitude, No O omits off-diagonal forcing, and P rule replaces both local original-rule terms.

Step	State	T	Full	P	P+A	No O	P rule
500	C	1	0.0643 ± 0.014	0.238 ± 0.023	0.152 ± 0.04	0.15 ± 0.04	0.316 ± 0.018
500	C	4	0.0429 ± 0.0096	0.179 ± 0.022	0.141 ± 0.029	0.139 ± 0.028	0.236 ± 0.016
500	R	1	0.207 ± 0.022	0.206 ± 0.027	0.209 ± 0.052	0.21 ± 0.051	0.22 ± 0.027
500	R	4	0.0853 ± 0.02	0.142 ± 0.028	0.13 ± 0.021	0.128 ± 0.021	0.176 ± 0.024
16500	C	1	1.6 ± 2.4	1.64 ± 2.4	1.61 ± 2.4	1.67 ± 2.5	1.52 ± 2.1
16500	C	4	$(7.57 \pm 8.5)10^5$	$(5.2 \pm 4.6)10^5$	$(5.47 \pm 5.8)10^5$	$(5.56 \pm 5.7)10^5$	$(4.11 \pm 3.5)10^5$
16500	R	1	1.27 ± 1.9	1.26 ± 1.8	1.29 ± 1.9	1.28 ± 1.9	1.27 ± 1.7
16500	R	4	$(3.39 \pm 3)10^5$	$(3.67 \pm 2.6)10^5$	$(3.48 \pm 2.8)10^5$	$(3.8 \pm 2.9)10^5$	$(3.3 \pm 2.6)10^5$

Table 24: Mean ± sample SD across three training seeds. At each fixed dose, relative error is first the median over four direction cases within seed, after averaging sixteen windows. Correlation is first computed over those four directions within seed; it is descriptive and based on only four points. C/R denote transported/retained. Directions and windows do not increase the seed count.

Step	T	s	C rel. error	R rel. error	C ρ	R ρ
500	1	-0.125	0.134 ± 0.032	0.4 ± 0.032	-0.0667 ± 0.42	-0.267 ± 0.23
500	1	-0.0625	0.0578 ± 0.0047	0.2 ± 0.047	0.733 ± 0.12	0.267 ± 0.31
500	1	-0.03125	0.0264 ± 0.0033	0.0857 ± 0.045	0.933 ± 0.12	0.867 ± 0.12
500	1	0.03125	0.0329 ± 0.011	0.111 ± 0.019	1 ± 0	0.867 ± 0.12
500	1	0.0625	0.0781 ± 0.026	0.264 ± 0.016	0.933 ± 0.12	0.8 ± 0
500	1	0.125	0.212 ± 0.04	0.528 ± 0.003	0.6 ± 0.35	0.8 ± 0
500	4	-0.125	0.0906 ± 0.047	0.223 ± 0.051	0.733 ± 0.31	0.667 ± 0.23
500	4	-0.0625	0.0491 ± 0.023	0.0603 ± 0.019	0.867 ± 0.12	0.733 ± 0.31
500	4	-0.03125	0.0247 ± 0.0089	0.0331 ± 0.0038	1 ± 0	0.933 ± 0.12
500	4	0.03125	0.0291 ± 0.016	0.0676 ± 0.025	1 ± 0	1 ± 0
500	4	0.0625	0.0573 ± 0.037	0.121 ± 0.053	0.933 ± 0.12	0.933 ± 0.12
500	4	0.125	0.124 ± 0.095	0.264 ± 0.13	0.8 ± 0	0.667 ± 0.42
16500	1	-0.125	2.83 ± 4.1	2.04 ± 3.2	-0.133 ± 0.61	-0.133 ± 0.76
16500	1	-0.0625	1.79 ± 2.6	1.28 ± 2	0 ± 0.8	$-7.4 \times 10^{-17} \pm 0.53$
16500	1	-0.03125	1.21 ± 1.8	0.886 ± 1.3	0.4 ± 0.69	0.733 ± 0.31
16500	1	0.03125	1.02 ± 1.4	0.661 ± 0.96	0.933 ± 0.12	0.733 ± 0.31
16500	1	0.0625	1.98 ± 2.9	1.63 ± 2.5	0.8 ± 0.35	0.667 ± 0.23
16500	1	0.125	2.78 ± 4.5	2.56 ± 3.9	0.733 ± 0.31	0.8 ± 0
16500	4	-0.125	$(4.01 \pm 5.3)10^6$	$(1.32 \pm 1.2)10^6$	0.0667 ± 0.64	$3.7 \times 10^{-17} \pm 0.69$
16500	4	-0.0625	$(1.07 \pm 1.5)10^6$	$(4.44 \pm 3.7)10^5$	-0.0667 ± 0.95	0.2 ± 0.2
16500	4	-0.03125	$(2.95 \pm 3.8)10^5$	$(1.07 \pm 0.86)10^5$	-0.0667 ± 0.76	0.333 ± 0.64
16500	4	0.03125	$(2.6 \pm 2.9)10^5$	$(1.05 \pm 0.67)10^5$	0.0667 ± 0.76	0.4 ± 0.4
16500	4	0.0625	$(1.03 \pm 1.4)10^6$	$(3.34 \pm 3.1)10^5$	0.267 ± 0.23	0.267 ± 0.23
16500	4	0.125	$(3.6 \pm 4.5)10^6$	$(1.37 \pm 1.1)10^6$	-0.0667 ± 0.64	-0.467 ± 0.58

Table 25: Four-step FD resolution. The central logit statistic is mean ± sample SD across three seed medians; each seed median uses four directions and four windows. The last two columns retain global maxima over all twelve checkpoint–direction cases and four windows, so worst-case errors are not averaged away. C/R denote transported/retained.

Step	State	h	Mean seed median	Max logit	Max Fisher
500	C	1.00×10^{-6}	$(6.3 \pm 1.7)10^{-8}$	1.51×10^{-7}	1.76×10^{-7}
500	C	1.00×10^{-5}	$(0.682 \pm 1.2)10^{-6}$	1.42×10^{-5}	1.81×10^{-5}
500	R	1.00×10^{-6}	$(2.06 \pm 2.6)10^{-7}$	3.72×10^{-6}	5.24×10^{-6}
500	R	1.00×10^{-5}	$(1.66 \pm 2.9)10^{-5}$	3.68×10^{-4}	5.18×10^{-4}
16500	C	1.00×10^{-8}	0.002 ± 0.0015	0.0256	0.02887
16500	C	1.00×10^{-7}	$(2.55 \pm 1.5)10^{-4}$	0.003045	0.002269
16500	C	1.00×10^{-6}	0.0146 ± 0.023	0.5774	0.7977
16500	C	1.00×10^{-5}	0.393 ± 0.47	3.392	2.878
16500	R	1.00×10^{-8}	0.00284 ± 0.0017	0.009046	0.008074
16500	R	1.00×10^{-7}	$(1.91 \pm 0.79)10^{-4}$	9.74×10^{-4}	9.93×10^{-4}
16500	R	1.00×10^{-6}	0.00557 ± 0.0072	0.09959	0.1092
16500	R	1.00×10^{-5}	0.332 ± 0.4	9.468	4.094

Table 26: Recovered FP64 diagnostic cost on two H200 hosts. Times are median synchronized seconds, except preparation, which is observed once. Peak is the maximum allocated GiB across repetitions and the two hosts; it includes resident tensors. The trajectory rows include ordinary continuation, and the JVP rows carry one direction. Cost at a horizon does not establish prediction accuracy at that horizon.

Phase	T	n	Host A (s)	Host B (s)	Peak (GiB)
Checkpoint preparation	–	1	4.64	6.964	4.88
Ordinary step	–	5	0.123	0.1468	10.62
Ordinary trajectory	1	5	0.1223	0.1464	10.62
Ordinary trajectory	4	5	0.4884	0.5852	12.64
Ordinary trajectory	8	5	0.9732	1.196	12.65
Ordinary trajectory	16	5	1.949	2.339	12.65
Ordinary trajectory	64	5	7.793	9.499	12.65
One-window readout	–	5	0.04556	0.05813	5.32
One-direction JVP	–	5	1.435	1.979	14.38
Trajectory + JVP	1	5	1.547	2.109	14.37
Trajectory + JVP	4	5	6.183	8.601	18.43
Trajectory + JVP	8	5	12.36	16.93	18.43
Trajectory + JVP	16	5	24.84	33.54	18.43
Trajectory + JVP	64	5	99.18	134.2	18.44

occupancy trace. Other GPUs on each host can remain occupied. Checkpoint preparation is a single observation per host.

The diagnostic has a measurable cost. On one H200, the median time for four control steps increases from 0.488 s to 6.183 s with one-direction sensitivity; peak allocated memory increases from 12.64 to 18.43 GiB. A single-window readout takes 0.0456 s and can serve multiple scalar doses once its tangent is computed. Appendix F.3 gives both hosts, repetitions, and measurement scope.

These timings describe the direct full-state implementation. Sensitivity accumulates per-example gradients over the same eight training sequences to bound activation memory, whereas ordinary continuation evaluates the full batch. A computed logit tangent can be reused across the six scalar doses; each additional direction or probe requires its corresponding propagation or readout work. The component experiment carries several additional tangents and actual dose branches for validation, so its shared-GPU job duration is not substituted for this benchmark.

G Related Work and Discussion

G.1 Related work

Parameter symmetries and balancing. Path-SGD studies the consequences of rescaling-equivalent representations and designs a rescaling-invariant update (Neyshabur et al., 2015). Balancedness identities and their broken conservation laws under discrete updates, momentum, and decay are developed by Du et al. (2018) and Kunin et al. (2021); function-preserving weight balancing is studied by Saul (2023). The SwiGLU reciprocal symmetry used here is explicitly given in DDC (Shirodkar, 2026, Appendix A.6). Soft gauge fixing instead penalizes redundant scale coordinates (Terin, 2026). These results provide the basis for the interventions studied here (see also Dinh et al., 2017; Neyshabur et al., 2016).

Relation to state-aware teleportation. Zhao et al. (2024) compare transforming parameters and momentum, transforming parameters alone, and resetting momentum. Their experiments target optimization after a selected teleportation. Libra uses these alternatives in a fixed-treatment, populated-checkpoint analysis that separates the transported recurrence, the update commutator, its observable projection, and the paired continuation. The precise covector convention here is specified independently of whether a previous implementation used the same map for a particular architecture.

Symmetry-compatible optimizers. DDC explicitly distinguishes Muon's rotation-compatible polar map from its incompatibility with per-channel scaling. Intrinsic Muon studies representative dependence for factor-wise spectral updates on low-rank quotient manifolds (Li et al., 2026). The SwiGLU analysis of Lau & Su (2026) studies intermediate-neuron permutations, a discrete symmetry complementary to the continuous reciprocal action used here. SIB uses function-preserving reparameterization before supervised fine-tuning (Catalan-Tatjer & Geiping, 2026). Libra measures a specified optimizer with unchanged-rule branches and isolates mechanisms with compensated and canonicalized controls.

Which implementation changes enter Libra. The finite NS polynomial and its normalization/precision choices specify the matrix map $\mathcal{A}$ (Jordan et al., 2024); the scalable Muon implementation adds matrix-size update scaling and an auxiliary Adam parameter group (Liu et al., 2025). Fixed size factors enter γ or the block learning rates, while auxiliary history belongs to S and must be propagated. Alternative update spectra (Shumaylov et al., 2026) change the output amplitudes and hence the finite-map derivative; Intrinsic Muon (Li et al., 2026) changes the geometric setting and representative handling. Libra tests a chosen state action and complete update map rather than transferring an exact-polar derivative unchanged to each variant. A changed state definition requires its own ρ_s, and changed normalization or quantized arithmetic requires its own branch or finite-precision checks.

Sensitivity and matrix calculus. Differentiating an optimizer through time is standard in hyperparameter optimization (Maclaurin et al., 2015; Franceschi et al., 2017; 2018). Flow commutation and Lie brackets already appear in the teleportation analysis of Zhao et al. (2024, Section 3.3). The Fréchet derivative of the polar decomposition, including its Sylvester-equation representation, is classical (Kenney & Laub, 1991; Gawlik & Leok, 2016). We specialize these tools to gauge interventions and derive an observable response and a scale-compensated spectral identity. Recent comparisons of alternative update spectra also caution against identifying a particular spectral construction with a performance benefit (Shumaylov et al., 2026). Libra uses the local predictive readout to connect update geometry to the observed response.

G.2 Discussion

Libra connects representation-dependent updates to local predictive change through scale homogeneity, matrix response, full-state propagation, and observable projection. This separation changes how Muon's response should be interpreted. Uniform-scale behavior alone cannot identify a polar-specific mechanism, and finite NS can change vector response as well as spectral amplitudes under heterogeneous scaling. The predictive importance of this discrepancy also depends on state propagation and observable projection: the component experiment suppresses some matrix directions while retaining substantial amplitude and off-diagonal effects. Optimizer history is part of this dynamical description, since memory can reinforce or cancel forcing in prediction space.

These conclusions concern a specified control trajectory, history convention, probe set, and implemented rule. Independent predictive confirmation covers one and four steps on three held-out training seeds in the 106.57M SwiGLU setting; longer-horizon path checks leave derivatives unresolved and finite-dose predictions fail. Exact-polar analysis assumes full rank, finite NS uses a fixed normalization branch, and auxiliary Adam follows the directional convention in Appendix E.7. Within this local scope, Libra connects update geometry to functional response without treating matrix energy as predictive importance.