BEYOND SCALAR EFFECTIVE LEARNING RATES: UPDATE GEOMETRY UNDER PARAMETER SYMMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter symmetries preserve a model's function but can change the speed and direction of its updates. We describe these changes through *effective learning geometry*, using a *quotient update map* that recovers angular effective learning rates (ELRs) for scale-invariant models as a special case. For strictly rectangular value–output (VO) factors of full rank, we prove that Euclidean factor responses admit one common scalar on the entire product tangent space exactly for orthogonal basis changes. Probes with zero optimizer history in a transformer with 123.6M parameters reveal distinct compatibility structures: AdamW responds to both basis orientation and nonuniform scaling, whereas Muon is nearly compatible with orthogonal changes but remains sensitive to nonorthogonal scaling. Directional mismatch dominates the residual, with little contribution from variation in optimal rates across batches. Starting from identical functions with corresponding non-VO updates, fitting a scalar directly to the full output response reduces AdamW's residual from 0.192 to 0.102 and Muon's from 0.082 to 0.064. For both optimizers, the fitted residuals remain larger than those under exact update transport. Optimizer history attenuates the VO mismatch for AdamW and has a small, mixed effect for Muon. Paired pretraining trajectories show how the tested scalar compensations alter loss and prediction dynamics. Symmetry, update direction, and optimizer history thus determine what a change in learning rate can absorb.

1 INTRODUCTION

Scalar effective learning rates (ELRs) describe how parameter norms, weight decay, and angular motion determine the effect of the external learning rate in scale-invariant networks (van Laarhoven, 2017; Hoffer et al., 2018; Li & Arora, 2020; Li et al., 2020; Wan et al., 2021; Kodryan et al., 2022; Heo et al., 2021). Recent experiments on language models show that matching a scalar ELR approximately collapses loss curves across pretraining interventions (Liu et al., 2026), while angular refinements account for the parameter–update angle (Xiao et al., 2026). These observations motivate a broader question: *when do parameter symmetries act only as rescalings of learning speed, and what directional structure remains when scalar compensation is insufficient?*

A symmetry changes the parameters while preserving the represented function. For a positive scaling symmetry, normalization removes the radial coordinate, and plain SGD at scaled weights recovers the same directional motion after a known correction to the learning rate. An invertible change of basis in attention's value–output (VO) factors preserves the function while mixing and stretching parameter directions. Its effect on subsequent updates can exceed a scalar rescaling. We use *effective learning geometry* to describe how these updates depend on direction. Its defining object is the optimizer's *quotient update map*, which maps gradients to descent directions in parameter space modulo symmetry, for a specified representative and optimizer state.

This distinction connects existing ELR analyses to optimizer–symmetry compatibility. Adam's basis dependence is already established (Singh, 2026), and gauge-equivariant conditioners explicitly restore compatibility (Shirodkar, 2026). Our question concerns the weaker remedy of scalar compensation and its limits under nonorthogonal transformations. We characterize this boundary using VO structure and transformation families, then connect parameter responses to outputs and training trajectories.

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 contributions are as follows:

1. We formulate effective learning geometry through a quotient update map that recovers scale and angular ELR and separates rate adjustment from irreducible directional change. For strictly rectangular VO factors of full rank, we characterize exactly which basis changes admit a common scalar for Euclidean responses on the full product tangent space.
2. Orthogonal controls, together with diagonal and dense controls sharing singular values, reveal distinct compatibility structures for AdamW and Muon. A decomposition of squared error shows that mismatch under dense transformations is predominantly directional, even when optimal rates vary little across batches.
3. Full output fitting at identical functions, controlled optimizer histories, and scalar and transport interventions during pretraining connect local responses to history, outputs, and training dynamics. They show how memory modifies mismatch and scalar compensation affects loss and prediction paths.

2 From scale ELR to effective learning geometry

The scaling special case. Let $w \in \mathbb{R}^d \setminus \{0\}$ be a vectorized parameter with $f_{cw} = f_w$ for every positive scale c. Write $r = \|w\|_2$, $q = \pi(w) = w/r$, and $P_q = I - qq^\top$, where I is the identity and P_q projects onto the tangent space of the unit sphere at q. For a fixed raw descent proposal u and external learning rate $\eta > 0$, the update is $w \mapsto w - \eta u$. Differentiating normalization gives

$$D\pi_w = r^{-1} P_q, \qquad \Delta q = -\frac{\eta}{r} P_q u + O(\eta^2), \qquad \frac{\|\Delta q\|_2}{\|u\|_2} \approx \frac{\eta}{r}\sqrt{1 - \cos^2(w,u)}, \tag{1}$$

where $\Delta q = \pi(w - \eta u) - \pi(w)$ and $\cos(w,u) = \langle w, u\rangle/(r\|u\|_2)$ for $u \neq 0$. This angular displacement *per unit norm of the raw update* agrees to first order with the angular ELR of Xiao et al. (2026), whose formula also accounts for finite steps. For matrix blocks, vectorization makes these Euclidean norms Frobenius norms.

If plain SGD supplies $u = \nabla_w L$ for the invariant loss $L(w) = \bar{L}(q)$, the chain rule contributes a second inverse norm:

$$\nabla_w L = r^{-1} P_q \nabla_q \bar{L}, \qquad \Delta q = -\frac{\eta}{r^2} P_q \nabla_q \bar{L} + O(\eta^2). \tag{2}$$

Here $\nabla_q \bar{L}$ is a sphere gradient, or the gradient of any smooth local extension followed by P_q. The factors $1/r$ and $1/r^2$ therefore describe different quantities: the angular response to a given raw proposal and the complete SGD response to a loss gradient. Under $w \mapsto cw$, SGD gradients scale by c^{-1}; multiplying the transformed learning rate by $a = c^2$ preserves the weight orbit for finite SGD steps as well. Figure 1 verifies this prediction with independently computed gradients in a 64-dimensional problem with a scaling symmetry with one parameter (Appendix F.10).

General parameter symmetries. Let a Lie group G act smoothly on the parameter space Θ, with $f_{g\theta} = f_\theta$ for a symmetry $g \in \mathsf{G}$. In a regular region, the quotient $\mathcal{Q} = \Theta/\mathsf{G}$ identifies parameters on the same symmetry orbit. Write $q = \pi(\theta)$ and $J_\theta = D\pi_\theta$, the derivative sending parameter perturbations to quotient perturbations. Fix quotient coordinates and an inner product, and use the ambient parameter inner products; these choices define the transpose $J_\theta^\top$ and all gradients below.

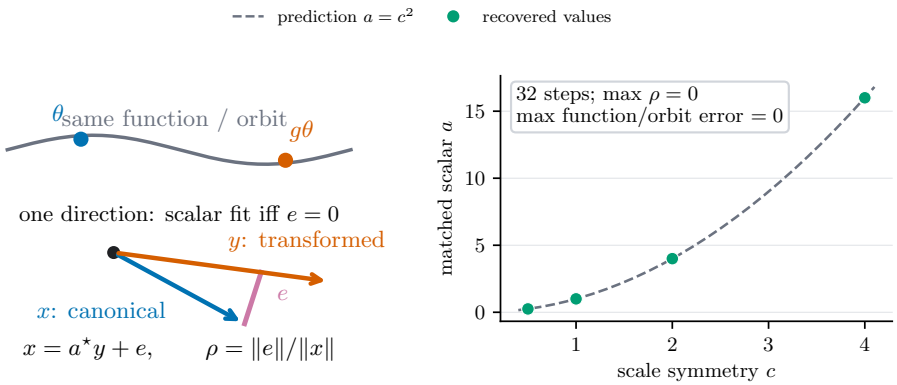
Given optimizer history s, let $\mathcal{U}_{\theta,s}(z)$ be its descent proposal for a parameter gradient z, before applying the external rate: $\Delta\theta = -\eta \mathcal{U}_{\theta,s}(z) + O(\eta^2)$. For a gradient direction $\xi \in T_q\mathcal{Q}$ on the quotient, the chain rule gives $z = J_\theta^\top \xi$. The *quotient update map* gives the projected descent proposal per unit learning rate:

$$\mathcal{R}_{\theta,s}(\xi) = J_\theta \mathcal{U}_{\theta,s}(J_\theta^\top \xi). \tag{3}$$

Thus $\Delta q = -\eta \mathcal{R}_{\theta,s}(\nabla_q \bar{L}) + O(\eta^2)$ for $L = \bar{L} \circ \pi$. This map can be nonlinear and depend on optimizer history. Only when the optimizer response is linear, $\mathcal{U}_{\theta,s}(z) = P_{\theta,s} z$, for a linear map on parameter space $P_{\theta,s}$, can we write the *quotient effective operator*

$$\mathcal{E}_{\theta,s} = J_\theta P_{\theta,s} J_\theta^\top. \tag{4}$$

Equation 2 is the case $J_w = r^{-1} P_q$, $P_{w,s} = I$. Effective learning geometry refers to how quotient updates depend on direction, without assuming curvature or a Riemannian metric.



(a) Responses at the same quotient point.

(b) Independent SGD scaling control.

Figure 1: **Scalar ELR is a special case of effective learning geometry.** (a) For a paired direction, x is the canonical response, y the transformed candidate, and $e = x - a^\star y$ the unmatched component. (b) In a 64-dimensional scale-invariant linear predictor, markers show the median fitted scalar over 32 SGD steps at each scale c; the dashed curve is the prediction $a = c^2$. All three branches compute gradients separately. Matched angular, function, and orbit errors are zero in the stored FP64 records. The scaling symmetry has one parameter.

Proposition 1 (Diagnostic for scalar matching)**.** *For representatives (θ, s), $(g\theta, s_g)$ at the same quotient point, with specified histories, one positive multiplier a matches all responses in a set $\mathcal{V} \subseteq T_q\mathcal{Q}$ exactly to first order if and only if*

$$\mathcal{R}_{\theta,s}(\xi) = a\mathcal{R}_{g\theta,s_g}(\xi) \quad \textit{for every } \xi \in \mathcal{V}. \tag{5}$$

In the linear case this is proportionality of the restricted effective operators. When a nonzero response is present and a matching scalar exists, it is unique.

This criterion defines *scalar-conformal* responses: a change in learning rate matches their magnitudes while preserving direction. Exact proportionality survives a common invertible coordinate change, although a numerical residual depends on the chosen metric (Appendix A).

Direction versus rate. For canonical and transformed responses $x \neq 0$, $y \neq 0$, least squares gives

$$a^\star = \frac{[\langle x, y\rangle]_+}{\|y\|^2}, \qquad \rho(x, y) = \frac{\|x - a^\star y\|}{\|x\|}, \qquad [v]_+ = \max(v, 0). \tag{6}$$

The residual measures the component that remains after the best nonnegative scalar fit in this metric. For microbatch observations (x_b, y_b), $b = 1, \ldots, B$, with every $\langle x_b, y_b\rangle > 0$, let $a_b^\star$ fit batch b and $a^\star = \sum_b \langle x_b, y_b\rangle / \sum_b \|y_b\|^2$ fit them jointly. Then the sum of squared errors (SSE) decomposes as

$$\underbrace{\sum_b \|x_b - a^\star y_b\|^2}_{\mathrm{SSE}_{\mathrm{common}}} = \underbrace{\sum_b \|x_b - a_b^\star y_b\|^2}_{\mathrm{SSE}_{\mathrm{direction}}} + \underbrace{\sum_b (a_b^\star - a^\star)^2 \|y_b\|^2}_{\mathrm{SSE}_{\mathrm{rate}}}. \tag{7}$$

The terms measure directional mismatch and the extra squared error from using one common rate, respectively. Large variation in optimal rates would increase the second term, favoring more frequent scalar adaptation. The first term persists after each observation receives its own optimal scalar. It therefore identifies the mismatch that remains even with this additional fitting freedom.

3 Which VO symmetries admit scalar compensation?

For one attention head, let d_{model} denote model width and d_h head width. Its value matrix $V \in \mathbb{R}^{d_h \times d_{\mathrm{model}}}$ and output matrix $O \in \mathbb{R}^{d_{\mathrm{model}} \times d_h}$ enter the head's contribution through $M = OV$. For any invertible matrix $S \in \mathrm{GL}(d_h)$, the general linear group,

$$(O, V) \mapsto (OS, S^{-1}V) \tag{8}$$

preserves M and the network function while other parameters remain fixed. This includes orthogonal rotations and nonorthogonal changes of basis.

A structural classification for Euclidean response. For general dimensions $O \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{r\times n}$, let $G_M = \nabla_M L$ be the matrix loss gradient. Euclidean gradient flow on the factors induces

$$\dot{M} = -\mathcal{E}_{O,V}(G_M), \qquad \mathcal{E}_{O,V}(G) = OO^\top G + GV^\top V, \tag{9}$$

where $G \in \mathbb{R}^{m\times n}$ is a matrix gradient input, the dot denotes continuous time, and matrix gradients use the Frobenius inner product. At the transformed factors, writing $S^{-T} = (S^{-1})^\top$,

$$\mathcal{E}_S(G) := \mathcal{E}_{OS,S^{-1}V}(G) = OSS^\top O^\top G + GV^\top S^{-T} S^{-1} V. \tag{10}$$

Theorem 1 (Scalar compatibility of rectangular VO factors)**.** *Suppose O has full column rank, V has full row rank, and $1 \le r < \min(m, n)$. Let $\mathcal{M}_r$ be the manifold of rank-r matrices and $\mathcal{T} = T_M\mathcal{M}_r$ its tangent space at $M = OV$. For $S \in \mathrm{GL}(r)$,*

$$\mathcal{E}_S|_{\mathcal{T}} = \gamma \mathcal{E}_{O,V}|_{\mathcal{T}} \text{ for some } \gamma > 0 \quad \Longleftrightarrow \quad S^\top S = I.$$

Here $\gamma = 1$, so the compensating learning rate multiplier $a = \gamma^{-1}$ is also one.

The proof isolates two types of tangent motion. A direction inside the column space of O but outside the row space of V forces $SS^\top = \gamma I$. Reversing these roles forces $(SS^\top)^{-1} = \gamma I$. Both can hold only for $\gamma = 1$. Thus even the reciprocal factor scaling $S = cI$, $c > 0$, $c \neq 1$, scales different response directions by incompatible factors c^2 and c^{-2}. A single change in learning speed cannot absorb both factors. Appendix B gives the full proof, rank conditions, and cases where restricted observations or square factors admit additional scalars. The result requires one scalar to match all tangent directions; individual gradients may still admit a scalar match. It builds on standard fixed-rank quotient geometry (Mishra et al., 2014).

Compatibility depends on the optimizer. The theorem applies to Euclidean factor updates; AdamW and Muon require separate tests. Coordinatewise Adam does not commute with general orthogonal mixing, even with zero history (Singh, 2026). The ideal memoryless polar proposal satisfies $\mathrm{polar}(Q_L Z Q_R) = Q_L \,\mathrm{polar}(Z) Q_R$: here Z is a matrix gradient, Q_L, Q_R are compatible orthogonal matrices, and $\mathrm{polar}(Z)$ replaces its nonzero singular values by one (Bernstein & Newhouse, 2024). This gives an orthogonal compatibility control (Appendix B). Practical Muon also uses finite Newton–Schulz iterations, momentum, and other choices outside this idealized statement.

4 Connecting local responses to pretraining

We use the Llama/Muon implementation of Semenov et al. (2025) from the LLM optimizer benchmark, adding VO actions, response probes, and paired interventions. The study uses one transformer with 12 layers, width 768, 12 heads, and 123,587,328 parameters, trained with sequence length 512 on a fixed FineWeb sample (Penedo et al., 2024). Tokens per parameter (TPP) means cumulative training tokens divided by model parameter count. Paired replicas consume identical minibatches through 20.0001 TPP, with 2,000 warmup steps and then constant learning rates. The primary transform is dense GL with condition number $\kappa = \sigma_{\max}(S)/\sigma_{\min}(S) = 4$. We selected it using numerical criteria in FP32 with TF32 disabled, before inspecting loss outcomes. Architecture, optimizer hyperparameters, seeds, and clipping are in Appendix D.

Local structure and memory. Probes with zero history (cold state) at TPP $0, 5, 10, 15, 20$ compare orthogonal, diagonal, and dense families, using $\kappa \in \{2, 4\}$ for the latter two. Diagonal and dense transforms have the same set of singular values at each κ; only basis orientation differs. Four transformation seeds and 32 paired microbatches give 200 cells across optimizers, transformation families, checkpoints, and seeds. We reset all moments, momentum buffers, and counters for each cold proposal. The *state shadow* compares cold and accumulated histories at fixed weights on the same orbit and fixed minibatches. Each optimizer accumulates history in its native coordinates along a common function path maintained by exact transport.

The main VO residual uses an exact finite-step fit: if $\delta O, \delta V$ are proposed factor increments, the candidate product change is $aA + a^2 B$, with $A = \delta O\, V + O\, \delta V$ and $B = \delta O\, \delta V$. We minimize its

squared Frobenius distance to the canonical product change R over $a \geq 0$, pooling all heads for a global scalar or grouping within layers. Appendix C specifies the quartic solver. The linear diagnostic in Equation 7 uses only first-order product changes; its common residual differs from the finite-step residual by at most 1.68×10^{-4} for AdamW and 1.52×10^{-5} for Muon.

Full output response at identical functions. At the five canonical checkpoints, we reconstruct the transformed weights, reset history, and use exactly corresponding non-VO updates. Let F be the vector of logits centered over the vocabulary on four fixed sequences at eight positions. Write x for its canonical directional derivative, b for the derivative from corresponding non-VO increments at the transformed weights, and y for that from the native transformed VO increment. With $\|x\|, \|y\| > 0$, we directly fit

$$a_{\text{out}} = \frac{[\langle x - b, y \rangle]_+}{\|y\|^2}, \qquad \rho_{\text{out}} = \frac{\|x - b - a_{\text{out}} y\|}{\|x\|}. \tag{11}$$

We compare this output-optimal global scalar with the VO-optimal scalar a_{VO} evaluated in the *same output objective*, and with the exact transported full update. Paired finite-difference scales verify response stability (Appendix F.6). This tests whether a global scalar suffices for a fixed output observation and VO candidate direction.

Training interventions and outcomes. The VO transformation is applied once to a copy of the shared initialization, before constructing both optimizers with zero histories. The canonical replica always takes its native update. The same-LR branch uses native updates and the same nominal learning rate schedule as its paired canonical replica. At each update, the global branch fits one scalar for all transformed VO proposals; the layerwise branch fits one per layer, shared across its VO heads. Both minimize the current finite-step VO-product mismatch and retain native updates outside VO. These three branches evolve without reapplying the transformation or forcing orbit alignment. Transport applies the exact pushforward of the canonical full update, including increments in unchanged coordinates outside VO. Each replica clips its full model gradients separately. Appendix F.4 shows the fitted schedules.

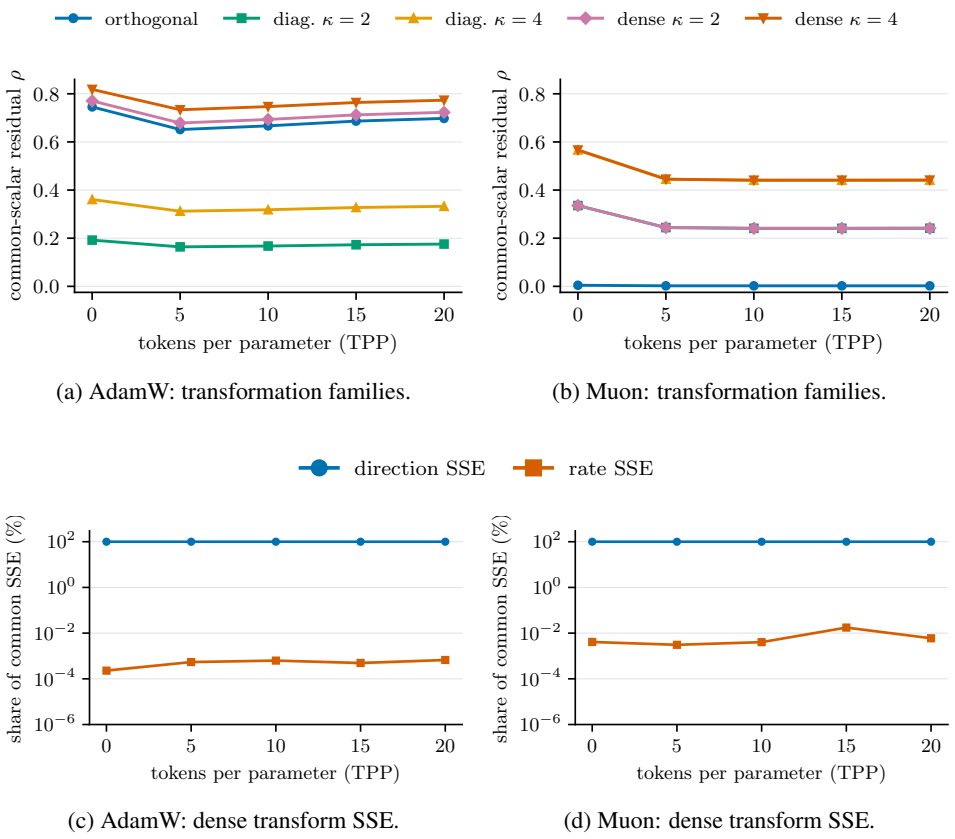
We evaluate loss on 512 held-out sequences every 256 steps, giving a shared grid of 78 evaluations. We measure prediction change on four fixed sequences at eight positions with a fixed 16-dimensional logit projection. We retain the full grid and the fixed $[1, 2, 3, 2, 1]$ smoothing kernel, renormalized at endpoints. Loss gap is the absolute difference between smoothed losses; function gap is the symmetric relative Euclidean distance between projected logits smoothed coordinatewise. Their grid averages are D_{loss} and D_{function}, respectively (Appendix E). D_{loss} measures deviation from the reference loss trajectory: higher or lower validation loss can increase this distance, so it does not rank predictive performance.

5 Results: structure, outputs, and dynamics

5.1 Basis orientation and nonuniform scaling affect the optimizers differently

At TPP 20 and condition number four (Figure 2(a,b)), AdamW's diagonal and dense residuals are 0.332 and 0.773 despite identical singular values; its orthogonal residual is 0.698. Basis orientation therefore matters in addition to nonuniform scaling. Muon gives 0.441 for both diagonal and dense families at condition number four, compared with 0.002 for orthogonal changes. These probes show near compatibility with orthogonal changes and sensitivity to nonorthogonal scaling, without establishing that Muon responses depend only on singular values. Orthogonal residuals measure compatibility, with numerical floors estimated separately.

Across all five checkpoints and four transformation seeds, median global residuals for dense transformations with condition number four are 0.764 for AdamW and 0.441 for Muon. Layerwise fitting changes them only slightly, from 0.7638 to 0.7631 and 0.4413 to 0.4409, respectively. Figure 2(c,d) explains why fitting a separate rate to each microbatch also helps little: directional mismatch accounts for more than 99.99% of common SSE for AdamW and 99.98% for Muon in every cell of the primary family. We verified positive inner products for all 6,400 underlying global microbatch pairs. Table 3 reports the SSE fractions and directional residuals for individual batches; Figure 5 shows variation in $a_b^\star$. A nearly stable optimal rate can therefore coexist with a large directional residual.



(a) AdamW: transformation families.

(b) Muon: transformation families.

(c) AdamW: dense transform SSE.

(d) Muon: dense transform SSE.

Figure 2: **Optimizer–symmetry structure and its directional origin.** (a,b) Exact finite-step common VO residuals; diagonal and dense families share singular values at each condition number κ. (c,d) Direction and rate fractions of linear common SSE for dense transforms with condition number four, plotted as percentages on a log scale; their sum is 100%. Curves show medians and bands the 10th–90th percentiles over four transformation seeds, each pooling 32 microbatches at the given checkpoint. Bands describe variation across transformations within one training run.

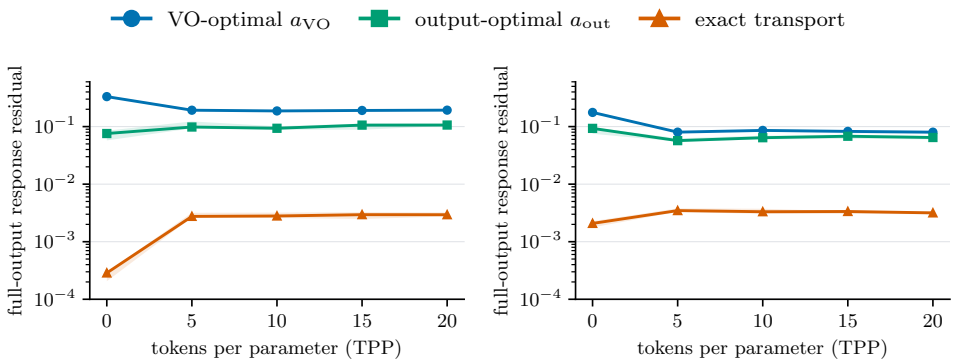
5.2 Direct output fitting reduces the residual but leaves measurable mismatch

Figure 3(a,b) compares the VO-optimal scalar, the directly fitted output scalar, and transport using one output objective and denominator. Post-initialization median residuals decrease from 0.192 to 0.102 for AdamW and from 0.082 to 0.064 for Muon. Both remain above their transport controls, 2.85×10^{-3} and 3.35×10^{-3}. The preferred scalar also depends on the observation: median a_{VO} and a_{out} are 0.492 and 0.718 for AdamW, and 0.747 and 0.831 for Muon. All function equivalence and finite-difference checks satisfy the numerical thresholds in Appendix F.6.

This result concerns a *global* scalar with zero optimizer history. VO-product and output residuals have different targets and denominators, so their magnitudes cannot quantify how much geometry the output Jacobian removes. Both optima describe instantaneous responses without optimizing a learning rate schedule over training.

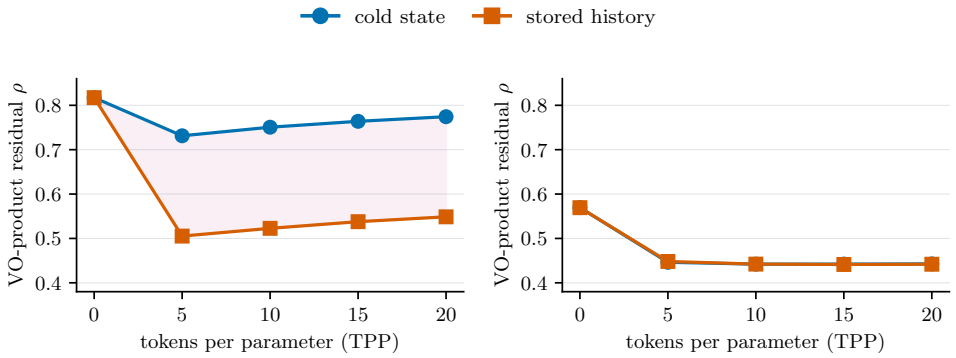
5.3 Memory attenuates AdamW's VO mismatch and only weakly modifies Muon's

At fixed weights on the same symmetry orbit, Figure 3(c,d) compares fresh and stored histories. AdamW history lowers the post-initialization VO-product residual by a median paired change of 0.225–0.228 at every checkpoint. Optimizer history therefore attenuates a mismatch already present in the cold state. Muon's changes have both signs and magnitude below 0.002. History's effect thus differs by optimizer. The experiment does not separate moments, bias correction, or other internal causes. Output responses with accumulated history remain unmeasured.



(a) AdamW: full output response.

(b) Muon: full output response.



(c) AdamW: VO response and history.

(d) Muon: VO response and history.

Figure 3: **Fitting output responses and measuring the effect of optimizer history.** All panels use the primary dense transform with condition number four. (a,b) Residuals of centered full-logit derivatives at identical functions with corresponding non-VO updates, normalized by the complete canonical response norm. Lines are medians and bands the 10th–90th percentiles across four fixed microbatches at each checkpoint, with cold optimizer state. Transport provides a separate numerical control. (c,d) Median VO-product residuals across four paired microbatches at fixed weights on the same orbit; shading joins medians for cold and accumulated histories. The output and VO measurements use distinct objectives and normalizations.

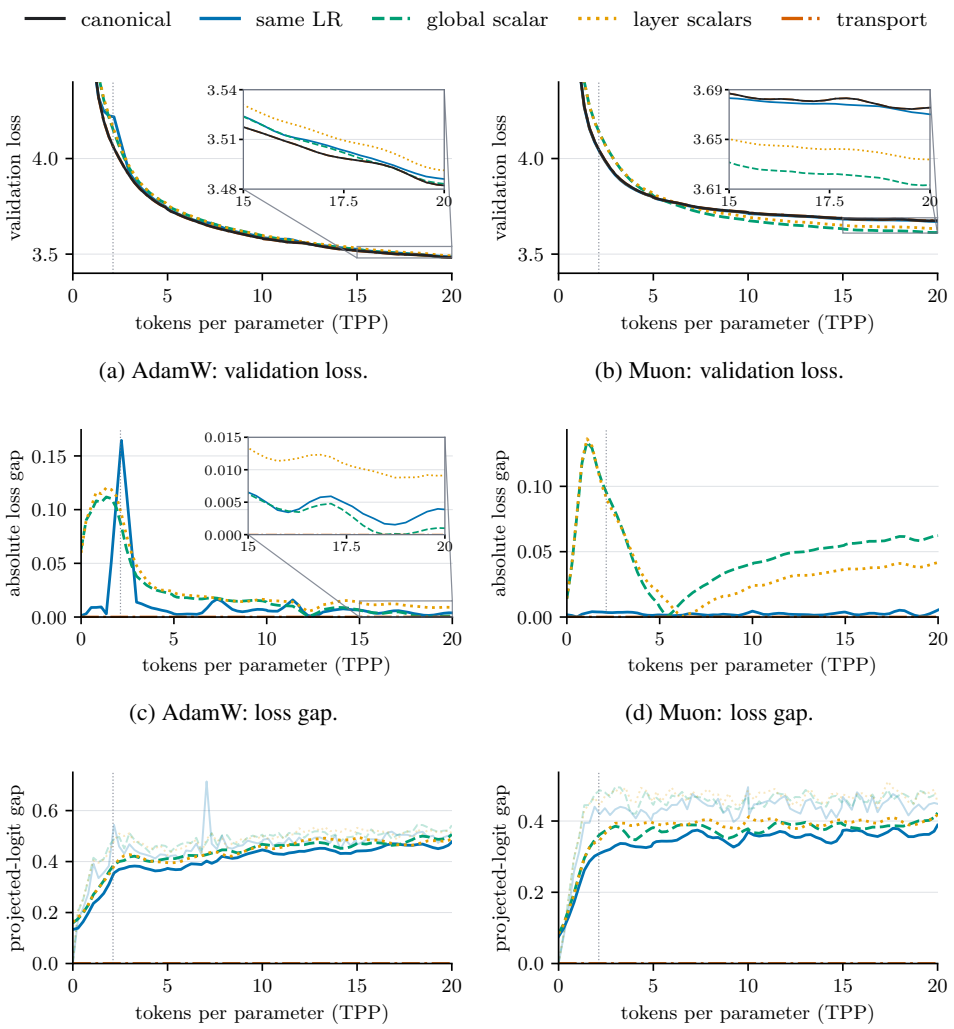
5.4 Scalar compensation changes loss and prediction dynamics

The paired models begin at the same function (Figure 4). Native and scalar paths separate early and remain separated in projected logits through 20 TPP, while exact transport tracks the canonical path. Gap panels reveal changes that are difficult to distinguish in absolute loss curves.

For AdamW, same LR, a global scalar, and layerwise scalars produce similar validation loss curves overall, with mean loss distances of 0.01217, 0.02229, and 0.02725, respectively. Their mean function distances remain substantial at 0.402, 0.431, and 0.431, respectively. Muon's native same-LR loss closely tracks the canonical trajectory, with $D_{\text{loss}} = 0.00247$, yet its mean function gap is 0.337. Global and layerwise VO compensation increase the loss distance to 0.04913 and 0.03717 and the function distance to 0.365 and 0.376. Table 4 gives the complete distance summaries, including the small distances under transport.

From TPP 10 onward, AdamW's same-LR and global-scalar mean loss gaps are close: 0.0053 and 0.0054, respectively. Prediction gaps for the native and both scalar branches persist and generally increase after their early separation. Small validation loss gaps therefore coexist with persistent separation of the observed prediction paths.

Muon's scalar interventions show a different temporal pattern. Their largest smoothed loss gaps occur around TPP 1.09, during warmup, reaching 0.134 for the global fit and 0.136 for layerwise fitting.



(a) AdamW: validation loss.

(b) Muon: validation loss.

(c) AdamW: loss gap.

(d) Muon: loss gap.

(e) AdamW: function gap.

(f) Muon: function gap.

Figure 4: **Pretraining dynamics under VO scalar and transport interventions.** Each column is one optimizer with the same initialization and data order. (a,b) Validation losses (main axes: 3.4–4.4). (c,d) Absolute loss gaps from the canonical replica; (c) spans the full range. Insets magnify boxed TPP 15–20 regions in (a–c), with vertical range 0–0.015 in (c). (e,f) Symmetric relative projected-logit difference. Faint lines use unsmoothed projected logits for the same-LR, global, and layerwise branches; opaque lines use the fixed $[1, 2, 3, 2, 1]$ kernel on projection coordinates with endpoint renormalization. Centered smoothing can carry later differences back to TPP 0; faint curves retain measured initialization values. Raw curves are for display; tabulated distances retain the fixed smoothed definition. The dotted vertical line marks the end of warmup. Canonical replicas and transport overlap; transport gaps are near zero on these linear axes.

The gaps then shrink toward TPP 5 before growing again under the constant-rate phase; final gaps are 0.062 and 0.042, respectively. Its same-LR gap remains small throughout. Later in training, the Muon branches with scalar compensation attain lower validation loss while deviating further from the canonical trajectory across multiple evaluations. The local fit and subsequent training use different objectives, so additional fitting freedom need not improve tracking of the reference. These results motivate studying when scalar summaries of learning speed remain useful despite changes in update direction.

Exact transport's loss distances are 2.72×10^{-8} and 4.22×10^{-8} for AdamW and Muon, respectively; function distances are 8.66×10^{-7} and 1.58×10^{-6}. This verifies that the paired implementation can

preserve the orbit when supplied the corresponding full update. After the native or scalar branches separate, their fits compare different quotient points, with native non-VO updates and separate gradient clipping. These curves evaluate the specified compensation schedules.

The close Muon loss paths are compatible with approximate scalar-ELR loss collapse (Liu et al., 2026); loss proximity alone does not identify a shared function path. Full-vocabulary logits and Jensen–Shannon (JS) divergence on saved checkpoints confirm clear separation from transport, while fine ordering among nearby treatments remains unresolved (Appendix F.8).

Uniform scalar compatibility requires one multiplier to align quotient update maps across all relevant gradient directions. Local fits test observed gradients in a specified metric. Output and memory probes compare equivalent representatives; trajectory experiments follow evolving parameters and optimizer histories.

6 Related work

Effective learning rates. In scale-invariant networks, parameter norms and weight decay affect the effective step size (van Laarhoven, 2017; Hoffer et al., 2018; Li & Arora, 2020; Li et al., 2020). Analyses of spherical dynamics and projected updates connect this effect to angular motion and training regimes (Wan et al., 2021; Kodryan et al., 2022; Heo et al., 2021). Xiao et al. (2026) incorporate the parameter–update angle into angular ELR, while Liu et al. (2026) connect scalar ELR matching to approximate loss collapse in language model pretraining. We recover the scaling case locally and ask when broader parameter symmetries also admit a scalar description.

Transformer parameter symmetries. Continuous attention symmetries extend parameter equivalence beyond permutations. Zhang et al. (2025) use rotations to align models for fusion; Da Silva et al. (2025) define sharpness on a quotient that removes these symmetries. Silverstein et al. (2026) introduce biases to break rotational redundancy, and François & Ravera (2026) formulate representations and optimization through invariant relational quantities. We use exact VO symmetries to compare optimizer updates from the same function.

Quotient geometry and optimization. Mishra et al. (2014) develop invariant metrics and quotient representations for optimization over matrices of fixed rank. Zhao et al. (2022) use symmetries to move along loss level sets and seek faster subsequent descent. Our analysis fixes the Euclidean factor update and characterizes which basis changes admit one scalar on the entire product tangent space.

Optimizer equivariance. Singh (2026) connects gauge equivariance to implicit bias and shows that Adam separates equivalent transformer initializations. Bernstein & Newhouse (2024) interpret matrix updates through steepest descent under matrix norms, providing the setting for our idealized polar control. Shirodkar (2026) constructs gauge-equivariant conditioners built on Adam and Muon. We test the weaker remedy of scalar compensation under nonorthogonal actions, combining directional error decomposition, controlled optimizer history, and direct output fitting.

7 Conclusion

Effective learning geometry extends scalar ELR from angular speed under scaling to changes in update direction under parameter symmetries. The VO classification identifies when one scalar suffices for Euclidean factor responses. The experiments reveal distinct compatibility structures for AdamW and Muon, predominantly directional mismatch, and measurable effects of optimizer history. Direct output fits and paired trajectories connect these local properties to predictions and pretraining dynamics. Scalar ELR remains a useful special case within a response determined by the interaction of symmetry, update direction, and optimizer state.

Limitations. The experiments use one model scale, initialization, training seed, FineWeb sample and order, primary trajectory transform, and configuration per optimizer. Local seeds vary only the transformation. Output diagnostics use finite, fixed observations. Output fitting tests a global scalar with cold state; layerwise output fits and stateful output responses remain unmeasured.

AI USE STATEMENT

Generative AI tools assisted with literature searches, language editing, mathematical derivation and verification, and code for analysis and plotting. The authors verified the cited sources, experimental results, mathematical statements, and generated figures and tables. They did not treat generated model outputs as experimental evidence and take responsibility for the final text, code, and scientific conclusions.

ETHICS STATEMENT

The study uses an existing pretraining sample of web text and involves no interaction with human subjects or deployment. Its principal risks are overclaiming from limited compute and inheriting biases or problematic content from web data. We address the risk of overclaiming through explicit scope limits, numerical checks, and comparisons with positive controls. We do not evaluate inherited biases or problematic content, and the resulting model is not released as a deployed system.

REPRODUCIBILITY STATEMENT

The appendices give complete proofs, the exact finite-step objective, configuration, random seeds, metric definitions, numerical thresholds, and additional tables.

REFERENCES

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024. URL https://opt-ml.org/papers/2024/paper93.pdf.

Marvin F. Da Silva, Felix Dangel, and Sageev Oore. Hide & seek: Transformer symmetries obscure sharpness & Riemannian geometry finds it. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55591–55613. PMLR, 2025. URL https://proceedings.mlr.press/v267/silva25a.html.

J. François and L. Ravera. Toward manifest relationality in transformers via symmetry reduction. *APS Open Science*, 1:000034, 2026. doi: 10.1103/v57m-nhhq. URL https://doi.org/10.1103/v57m-nhhq.

Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Iz3zU3M316D.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: Efficient and accurate normalization schemes in deep networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a0160709701140704575d499c997b6ca-Abstract.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Maxim Kodryan, Ekaterina Lobacheva, Maksim Nakhodnov, and Dmitry P. Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5aea56eefab60e06f35016478e21aae6-Abstract-Conference.html.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJg8TeSFDH.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a7453a5f026fb6831d68bdc9cb0edcae-Abstract.html.

Zihan Liu, Ruiheng Zheng, Shaobo Zhang, Changxin Tian, Kunlong Chen, Zhiqiang Zhang, and Lei Wu. Effective learning rate governs loss dynamics in language model pretraining. *arXiv preprint arXiv:2608.24814*, 2026. URL https://arxiv.org/abs/2608.24814.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29:591–621, 2014. doi: 10.1007/s00180-013-0464-z. URL https://arxiv.org/abs/1209.0430.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025. URL https://arxiv.org/abs/2509.01440.

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

Eva Silverstein, Daniel Kunin, and Vasudev Shyam. Symmetry breaking in transformers for efficient and interpretable training. *arXiv preprint arXiv:2601.22257*, 2026. URL https://arxiv.org/abs/2601.22257.

Devender Singh. The loss does not see the basis, but Adam does. *arXiv preprint arXiv:2608.05136*, 2026. URL https://arxiv.org/abs/2608.05136.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017. URL https://arxiv.org/abs/1706.05350.

Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using SGD and weight decay. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/326a8c055c0d04f5b06544665d8bb3ea-Abstract.html.

Yihao Xiao, Jialong Sun, Zitian Gao, Zeming Wei, Chutian Wang, Ran Tao, Jiaye Teng, and Bryan Dai. Hyperball may not be a free lunch. *arXiv preprint arXiv:2607.22444*, 2026. URL https://arxiv.org/abs/2607.22444.

Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77090–77106. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25dk.html.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/69f7750aa28f75fddf101da038f8b529-Abstract-Conference.html.

A Formal setting and scalarization proofs

A.1 Regular quotient and update map

Let G act freely and properly on an open neighborhood of $\theta \in \Theta$, so $\mathcal{Q} = \Theta/G$ is locally a smooth manifold and $\pi : \Theta \to \mathcal{Q}$ is a submersion. This assumption excludes singular orbit strata from the differential statement. The matrix identities below also apply without a global smooth quotient. Fix quotient coordinates and an inner product at $q = \pi(\theta)$, together with an inner product on the ambient parameter coordinates at each representative. Transposes and gradients below are defined relative to these choices. Because the loss factors through the quotient, $L = \bar{L} \circ \pi$,

$$\nabla_\theta L = J_\theta^\top \nabla_q \bar{L}, \qquad J_\theta = D\pi_\theta.$$

For optimizer state s, define the response per unit external learning rate by

$$\Delta\theta = -\eta \mathcal{U}_{\theta,s}(z) + O(\eta^2).$$

The response $\mathcal{U}$ may be nonlinear. Equation 3 then gives

$$\Delta q = -\eta \mathcal{R}_{\theta,s}(\nabla_q \bar{L}) + O(\eta^2).$$

When $\mathcal{U}_{\theta,s}(z) = P_{\theta,s} z$, Equation 4 follows by substitution. The resulting operator is positive semidefinite if $P_{\theta,s}$ is positive semidefinite. Effective learning geometry also covers responses outside this special case.

A.2 Angular normalization and independent SGD compensation

For $w \neq 0$ and a parameter perturbation h, differentiating $\pi(w) = w/\|w\|_2$ gives

$$D\pi_w[h] = \frac{h}{r} - \frac{w\langle w, h\rangle}{r^3} = r^{-1} P_q h.$$

A Taylor expansion at fixed w, u gives Equation 1. The angle $\Delta\phi = \angle(w, w - \eta u)$ and the spherical chord satisfy $\|\Delta q\|_2 = 2\sin(\Delta\phi/2)$, so they agree to first order. Both our angular coefficient and that of Xiao et al. (2026) divide angular displacement by the proposal norm before multiplication by the external learning rate. A radial decay term of order η has zero first-order projection, although its effect on the parameter norm can affect later steps. For a Hyperball proposal normalized to the parameter radius, the denominator must use the normalized proposal's norm. These local statements retain the first-order scope of the expansion; finite steps require the exact angular expression and its denominator conditions.

For a scale-invariant differentiable loss, differentiate $L(cw) = L(w)$ with respect to w to obtain $\nabla L(cw) = c^{-1}\nabla L(w)$. SGD steps evaluated independently on matched minibatches then satisfy

$$cw - c^2\eta \nabla L(cw) = c\big(w - \eta \nabla L(w)\big).$$

Induction preserves $w_t^c = cw_t$ and equal normalized predictions for as long as the weights remain nonzero. This finite-step identity assumes plain SGD without momentum or weight decay, and matched data. The general expansion of the angular response remains a first-order statement. The control in Appendix F.10 evaluates these gradients separately.

A.3 Proof of the scalarization criterion

Proof of Proposition 1. If $\mathcal{R}_{\theta,s}|_{\mathcal{V}} = a\mathcal{R}_{g\theta,s_g}|_{\mathcal{V}}$, multiplying the transformed external rate by a gives the same first-order quotient displacement for every $\xi \in \mathcal{V}$. Conversely, a single scalar matches every direction precisely when $\mathcal{R}_{\theta,s}(\xi) = a\mathcal{R}_{g\theta,s_g}(\xi)$ for all $\xi \in \mathcal{V}$, which is equality of the restricted maps. In the linear case, substitute $\mathcal{R}(\xi) = \mathcal{E}\xi$. If both a and a' match the restricted maps and some transformed response is nonzero, then $(a - a')\mathcal{R}_{g\theta,s_g}(\xi) = 0$ for a direction with nonzero response, so $a = a'$. □

The criterion is invariant under a common invertible change of quotient coordinates, with the gradient inputs relabeled consistently. For an invertible coordinate derivative C, if both charts use Euclidean coordinate gradients then $\mathcal{R}'(\xi') = C\mathcal{R}(C^\top \xi')$ and $\mathcal{V}' = C^{-T}\mathcal{V}$. Applying this transformation to

both representatives preserves scalar proportionality in both directions. The statement specifies the state pair (s, s_g) without requiring a canonical group action on optimizer states. Fitting a scalar to one realized gradient is weaker than establishing equality on a subspace. For nonlinear quotient update maps, equality on a basis need not imply equality on its span.

If both restricted responses vanish, every positive scalar matches them, so the scalar is unidentifiable. If only one vanishes, no positive scalar matches them. The normalized empirical residuals exclude both cases by checking that both response norms are positive.

A.4 Projection-optimal and common scalars

Proposition 2 (Best nonnegative scalar)**.** *For $x \neq 0$ and $y \neq 0$,*

$$\min_{a \geq 0} \|x - ay\|^2 = \|x\|^2 - \frac{[\langle x, y\rangle]_+^2}{\|y\|^2}, \qquad a^\star = \frac{[\langle x, y\rangle]_+}{\|y\|^2}.$$

If $\langle x, y\rangle > 0$, then $\rho(x, y) = \sqrt{1 - \cos^2(x, y)}$; otherwise $a^\star = 0$ and $\rho = 1$.

Proof. Expand

$$h(a) = \|x\|^2 - 2a\langle x, y\rangle + a^2\|y\|^2.$$

The unconstrained minimizer is $\langle x, y\rangle / \|y\|^2$; projecting it onto $[0, \infty)$ gives $a^\star$. Substitution gives the minimum. Dividing by $\|x\|^2$ yields the cosine expression when the inner product is positive. □

For observations $b = 1, \ldots, B$, suppose all inner products are positive and define

$$a_b^\star = \frac{\langle x_b, y_b\rangle}{\|y_b\|^2}, \qquad a^\star = \frac{\sum_b \langle x_b, y_b\rangle}{\sum_b \|y_b\|^2} = \frac{\sum_b \|y_b\|^2 a_b^\star}{\sum_b \|y_b\|^2}.$$

Proposition 3 (Decomposition for a common scalar)**.** *Under the preceding conditions,*

$$\sum_b \|x_b - a^\star y_b\|^2 = \sum_b \|x_b - a_b^\star y_b\|^2 + \sum_b (a_b^\star - a^\star)^2 \|y_b\|^2.$$

Proof. Let $r_b = x_b - a_b^\star y_b$. First-order optimality gives $\langle r_b, y_b\rangle = 0$, and

$$x_b - a^\star y_b = r_b + (a_b^\star - a^\star) y_b.$$

Apply the Pythagorean identity and sum over observations. Differentiating the pooled objective gives the weighted mean above. □

The first term in Proposition 3 measures the directional mismatch that remains after fitting a scalar to each observation. The second measures the additional error from requiring a common rate. All 6,400 global microbatch observations satisfy strict positivity, as verified from their recorded inner products and norms. Across the 200 cells, the decompositions agree with the common SSE to a maximum error of 4.14×10^{-16} relative to the total reference squared norm. This check verifies the linear diagnostic; the quadratic projection identity does not apply to the finite-step quartic fit.

B VO action, optimizer controls, and transport

B.1 Notation bridge to the implementation

The paper uses $M = OV$, with $O \in \mathbb{R}^{d_{\text{model}} \times d_h}$ and $V \in \mathbb{R}^{d_h \times d_{\text{model}}}$. The implementation stores and forms the transposed composite $M^\top$. Its bank matrix T corresponds to the paper's $S = T^{-T}$. Inverse transpose preserves orthogonality, condition number, and membership in the diagonal or dense family. Transposition preserves Frobenius residuals. The code and paper conventions therefore define the same tests.

B.2 EUCLIDEAN FACTOR RESPONSE

For $G_M = \nabla_M L$, the chain rule gives

$$\nabla_O L = G_M V^\top, \qquad \nabla_V L = O^\top G_M.$$

Under Euclidean gradient flow,

$$\begin{aligned}\dot{M} &= \dot{O}\,V + O\,\dot{V} \\ &= -G_M V^\top V - O O^\top G_M \\ &= -\mathcal{E}_{O,V}(G_M).\end{aligned}$$

At $(O_S, V_S) = (OS, S^{-1}V)$,

$$O_S O_S^\top = OSS^\top O^\top, \qquad V_S^\top V_S = V^\top S^{-T} S^{-1} V,$$

which proves Equation 10.

B.3 CLASSIFICATION ON THE FULL PRODUCT TANGENT SPACE

We prove Theorem 1 with the Frobenius metric on matrix space. Let $C = \mathrm{col}(O) \subset \mathbb{R}^m$ and $R = \mathrm{row}(V) \subset \mathbb{R}^n$, and let P_C, P_R be their orthogonal projectors. Full factor rank implies $\mathrm{rank}(M) = r$ and

$$\mathcal{T} = T_M \mathcal{M}_r = \{Z \in \mathbb{R}^{m\times n} : (I - P_C) Z (I - P_R) = 0\}.$$

This tangent space is the orthogonal sum of blocks whose column and row spaces lie in (C, R), $(C, R^\perp)$, and $(C^\perp, R)$. The factor differential $J(\delta O, \delta V) = \delta O\, V + O\, \delta V$ has image $\mathcal{T}$; its adjoint is $J^*(G) = (GV^\top, O^\top G)$. Consequently, $JJ^* = \mathcal{E}_{O,V}$. This operator maps into $\mathcal{T}$ and annihilates normal inputs. Restricting it to $\mathcal{T}$ therefore tests the entire first-order product response, independently of an ambient extension of the loss.

Proof of Theorem 1. Write $H = SS^\top$, which is symmetric positive definite. Suppose $\mathcal{E}_S(G) = \gamma \mathcal{E}_{O,V}(G)$ for every $G \in \mathcal{T}$, with $\gamma > 0$. Since $n > r$, choose a nonzero $v \in R^\perp$. For any $u \in C$, the rank-one matrix $G = uv^\top$ lies in $\mathcal{T}$, and $Vv = 0$ eliminates the right-multiplication term. Thus

$$O(H - \gamma I) O^\top u\, v^\top = 0 \quad \text{for every } u \in C.$$

Cancel $v^\top$. Because O has full column rank, it is injective and $O^\top : C \to \mathbb{R}^r$ is surjective, so $H = \gamma I$.

Since $m > r$, choose a nonzero $u \in C^\perp$ and let v vary in R. For $G = uv^\top$, the left-multiplication term vanishes and

$$u\, v^\top V^\top (H^{-1} - \gamma I) V = 0 \quad \text{for every } v \in R.$$

Cancel u. The map $V : R \to \mathbb{R}^r$ is surjective, and V has a right inverse because it has full row rank. Hence $H^{-1} = \gamma I$. Combining the identities gives $\gamma = \gamma^{-1}$; positivity forces $\gamma = 1$ and $H = I$. For square invertible S, $SS^\top = I$ is equivalent to $S^\top S = I$.

Conversely, if S is orthogonal, both inserted Gram factors in Equation 10 are identities, so the operators agree on the full ambient matrix space. The response is nonzero on $\mathcal{T}$: for nonzero $G \in \mathcal{T}$, $\langle G, \mathcal{E}_{O,V}(G)\rangle_F = \|O^\top G\|_F^2 + \|GV^\top\|_F^2 > 0$. Thus the proportionality factor is uniquely one. □

An alternative block reconstruction. Choose orthonormal bases U, W for C, R, with complements $U_\perp, W_\perp$, and write $O = UA$, $V = BW^\top$, where $A, B \in \mathbb{R}^{r\times r}$ are invertible. Every tangent matrix is $UKW^\top + ULW_\perp^\top + U_\perp N W^\top$, for arbitrary matrices K, L, N of the indicated compatible sizes. On the second block, operator proportionality reads $AHA^\top L = \gamma AA^\top L$ for every L. On the third it reads $NB^\top H^{-1} B = \gamma N B^\top B$ for every N. Invertibility again yields $H = \gamma I$ and $H^{-1} = \gamma I$. The two complementary blocks thus impose reciprocal constraints on the same scalar.

Boundary cases and quantifiers. The rectangular inequalities provide sufficient conditions for both arguments using complementary spaces above. We make no claim that both are necessary. Without them, additional scalar symmetries can occur: if $m = n = r$, $O = V = I$, and $S = cI$, then $\mathcal{E}(G) = 2G$ and $\mathcal{E}_S(G) = (c^2 + c^{-2})G$ for every G, including positive $c \neq 1$. In the scalar case $m = n = r = 1$, the product tangent space itself has dimension one.

Rank deficiency also permits invisible actions. For $m = n = 3, r = 2$, take $O = [e_1, 0]$ and $V = [e_1, 0]^\top$, where $e_1 = (1, 0, 0)^\top$. The nonorthogonal $S = \mathrm{diag}(1, 2)$ changes neither factor nor its response. This pair lies outside the stratum of full rank. A restricted gradient family can also scalarize with rectangular factors of full rank: if $O = U$, $V = W^\top$ have orthonormal columns/rows and $S = cI$, then every core direction $G = UKW^\top$ has common response ratio $(c^2 + c^{-2})/2$. The complementary blocks rule out scalarization on the *whole* tangent space. Collinearity on particular gradients remains possible, and adaptive optimizer histories lie outside the theorem's scope.

Proposition 4 (Orthogonal invariance and generic failure of scalar proportionality)**.** *If $S^\top S = I$, then $\mathcal{E}_{O_S,V_S} = \mathcal{E}_{O,V}$. If $S = sI$, $s > 0$, $s \neq 1$, and $A = OO^\top G$ and $B = GV^\top V$ are linearly independent, then $\mathcal{E}_{O_S,V_S}(G)$ is not proportional to $\mathcal{E}_{O,V}(G)$.*

Proof. Orthogonality makes both inserted Gram factors in Equation 10 the identity. For $S = sI$,

$$\mathcal{E}_{O_S,V_S}(G) = s^2 A + s^{-2} B, \qquad \mathcal{E}_{O,V}(G) = A + B.$$

If these were proportional, independence of A, B would require $s^2 = s^{-2}$, contradicting positive $s \neq 1$. □

A concrete counterexample is

$$O = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad V = I, \quad G = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad s = 2.$$

Direct substitution gives

$$\mathcal{E}(G) = \begin{bmatrix} 2 & 2 \\ 5 & 5 \end{bmatrix}, \qquad \mathcal{E}_s(G) = \begin{bmatrix} 4.25 & 4.25 \\ 16.25 & 16.25 \end{bmatrix}.$$

The rowwise ratios are 2.125 and 3.25, so the two responses are not collinear. This direction disproves universal scalarization for the specified representative pair. Other choices of factors, gradients, or S may yield collinear responses.

B.4 Optimizer-specific controls

Coordinatewise AdamW. At zero moments and negligible ϵ, the coordinatewise Adam proposal is proportional to $\mathrm{sign}(z)$. General orthogonal mixing does not commute with sign or elementwise squaring. For example, with

$$z = (1, 2), \qquad Q = 2^{-1/2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$

we have $\mathrm{sign}(zQ) = (1, 1)$ but $\mathrm{sign}(z)Q = (\sqrt{2}, 0)$. Moment accumulation remains coordinatewise. Decoupled weight decay is compatible with the linear factor action, but adding an equivariant term does not generically cancel a coordinatewise anisotropy. This permits failure even for the orthogonal subgroup and is consistent with the dependence on basis established by Singh (2026).

Idealized Muon control. For a compact SVD $Z = U\Sigma V^\top$ over its positive singular values, define $\mathrm{polar}(Z) = UV^\top$. Orthogonal Q_L, Q_R obey

$$\mathrm{polar}(Q_L Z Q_R) = Q_L \,\mathrm{polar}(Z) Q_R.$$

For orthogonal S, factor gradients transform as

$$\nabla_{O_S} L = (\nabla_O L)S, \qquad \nabla_{V_S} L = S^\top \nabla_V L.$$

An exact polar proposal without memory consequently transports as $\delta O_S = \delta O\, S$ and $\delta V_S = S^{-1}\delta V$. This statement excludes general invertible S, changing rank, the finite Newton–Schulz approximation, momentum, Nesterov correction, width scaling, weight decay, and AdamW fallback parameters. The empirical orthogonal result therefore serves as a numerical control. The idealized statement does not cover the complete optimizer implementation.

B.5 EXACT UPDATE TRANSPORT

Proposition 5 (Transport under a linear action)**.** *Let $g : \Theta \to \Theta$ be linear. If $\theta^g = g\theta$ and $\Delta\theta^g = Dg\,\Delta\theta = g\Delta\theta$, then*

$$\theta^g + \Delta\theta^g = g(\theta + \Delta\theta).$$

Proof. Linearity gives $g\theta + g\Delta\theta = g(\theta + \Delta\theta)$. □

By induction, the transported branch remains on the same orbit at every step in exact arithmetic. Equation 8 is blockwise linear, so the proposition applies. The experiment applies this pushforward while measuring and discarding the native transformed proposal. Transport thus provides a positive control for the group action, paired data and random numbers, update application, and function metric. This control establishes no equivariance property for the native optimizer. We measure the error from finite precision to establish the numerical floor.

C EXACT FINITE-STEP SCALAR OBJECTIVE

Let R be the canonical finite update to the VO product. For a transformed proposal $(\delta O, \delta V)$,

$$(O + a\delta O)(V + a\delta V) - OV = aA + a^2B, \quad A = \delta O\, V + O\, \delta V, \quad B = \delta O\, \delta V.$$

The exact fit solves

$$\min_{a \geq 0} \|R - (aA + a^2B)\|_F^2. \tag{12}$$

With

$$(r_r, r_a, r_b, a_a, a_b, b_b) = (\|R\|^2, \langle R, A\rangle, \langle R, B\rangle, \|A\|^2, \langle A, B\rangle, \|B\|^2),$$

the objective is

$$r_r - 2r_a a + (a_a - 2r_b)a^2 + 2a_b a^3 + b_b a^4. \tag{13}$$

The solver evaluates the boundary $a = 0$ and every nonnegative real root of the cubic derivative, then chooses the scalar with the lowest objective, breaking exact ties in favor of the smaller scalar. The squared-norm polynomial is coercive whenever it has a nonconstant term. The same enumeration of the boundary and stationary points handles degenerate cases of lower degree. Before storing a fit, the implementation checks that the Gram matrix is positive semidefinite and that the residuals are nonnegative.

For a global fit, we sum the six coefficients over all observed heads and microbatches before solving. Layerwise and headwise fits group these coefficients by layer and head, respectively. Because relaxing a shared scalar cannot increase the optimum, the implementation checks

$$\mathrm{SSE}_{\mathrm{head}} \leq \mathrm{SSE}_{\mathrm{layer}} \leq \mathrm{SSE}_{\mathrm{global}}.$$

The linear diagnostic drops B and becomes Proposition 2. Across the local sweep, the maximum absolute difference between exact and linear residuals is 1.6813×10^{-4} for AdamW and 1.5195×10^{-5} for Muon. Within the primary dense family with condition number four, the maxima are 6.2444×10^{-5} and 1.5110×10^{-5}. The exact finite-step fit refines the measurement while preserving its qualitative relation to the local theorem.

The deterministic algebra check uses seed 20270905. It verifies projection stationarity and the residual identity, the decomposition for a common scalar, the quartic expansion, orthogonal VO invariance, nonconformality under dense GL actions, idealized polar equivariance, and exact transport.

D EXPERIMENTAL DETAILS

D.1 CONFIGURATION

Table 1 gives the training configuration. The model has no biases, so the VO action is exact. It uses neither a mixture of experts nor dropout. Both replicas in each paired run use the same global batch and data order. AdamW uses the adaptive moments of Kingma & Ba (2015) with decoupled weight decay (Loshchilov & Hutter, 2019).

Table 1: Training configuration. Hidden matrices use Muon; embeddings, normalization parameters, and other ineligible tensors use the stated AdamW fallback.

Component	Field	Value
Model	layers / width / heads / KV heads	12 / 768 / 12 / 12
Model	head dim. / context / vocabulary	64 / 512 / 50,304
Model	trainable parameters	123,587,328
Batch	sequences / tokens per update	256 / 131,072
Batch	microbatch seq. / accumulation	64 / 4
Horizon	updates / tokens / final TPP	18,858 / 2,471,755,776 / 20.00007457
Schedule	warmup / post-warmup / clip	2,000 / constant / 0.5
AdamW	rate / β_1, β_2 / ϵ / decay	10^{-3} / (0.8, 0.999) / 10^{-8} / 0.1
Muon	hidden rate / momentum / Nesterov	10^{-2} / 0.95 / yes
Muon	NS steps / decay / fallback rate	5 / 0.1 / 10^{-3}
Execution	precision / TF32 / distributed	full FP32 / off / no

The checkpoint schedule is

label	tpp0	tpp5	tpp10	tpp15	tpp20
step	0	4,715	9,429	14,144	18,858
actual TPP	0	5.0005489	10.0000373	15.0005862	20.0000746

All eight paired training runs completed the stated horizon without early stopping.

D.2 Random seeds and transformations

Initialization and training RNG seeds are 0, and the seed for data ordering is 271828. The primary trajectory uses transformation seed 314159. The local sweep uses $\{314159, 271828, 161803, 141421\}$; the seed for logit projection is 161803, the bootstrap seed is 173205, and the seed for the scaling control is 223607.

For every layer and head, we draw an orthogonal matrix by FP64 Gaussian QR with diagonal sign correction. For a target condition number κ, the 64 singular values are logarithmically spaced from $\kappa^{-1/2}$ to $\kappa^{1/2}$. A diagonal transform randomly permutes these values. A dense transform uses independently drawn orthogonal left and right axes. For each seed, condition numbers two and four share the same axes and ordering, so only the singular values change. Each transform bank covers every layer and head exactly once and stores FP32 matrices and explicit inverses. All banks pass checks for inverse reconstruction, condition number, and function equivalence.

Appendix B relates the implementation's transposed weight convention to the paper's convention. The same routine applies the parameter action and the exact pushforward to updates because the action is linear on each affected tensor.

D.3 Precision selection before outcome inspection

Before inspecting training outcomes, we ordered four numerical precision candidates: symmetry-safe VO arithmetic at dense condition four, then condition two, followed by full FP32 at condition four and condition two. TF32 was disabled throughout. We selected the first candidate that met both the function equivalence criterion and the criterion for error relative to a single update, without using loss outcomes. Full FP32 at condition four was the first candidate to meet both criteria. The maximum allowed function error in projected logits was 10^{-4}, and its maximum allowed ratio to the median one-step change was 0.05. The selected initial values were 1.91×10^{-6} and 0.002372, respectively.

D.4 Paired optimizer step

At update t, both replicas use the same sequence indices and random state. Each replica accumulates its loss and clips its gradients separately to the common threshold. Both optimizers form proposals before either update is applied. The canonical branch always applies its native proposal. The transformed branch applies one of:

1. its unmodified native proposal;

2. the native proposal with one exact finite-step scalar applied to every transformed VO proposal;
3. the native proposal with one exact finite-step scalar per layer applied to its transformed VO proposals; or
4. the exact pushforward of the canonical proposal.

We recompute the global and layer scalars from each step's paired proposals, without caps. Scaling applies only to the VO tensors affected by the reparameterization; other tensors retain their native paired updates. The transported branch forms and records its native proposal, then discards it and applies the transported canonical update. We record symmetry metrics every 10 steps and evaluation metrics every 256 steps.

D.5 Cold probe and state shadow

For each optimizer, the cold grid contains

$$5 \text{ families} \times 5 \text{ checkpoints} \times 4 \text{ transform seeds} = 100$$

summary cells, each containing 32 held-out paired microbatches of 16 sequences. We reset optimizer state to zero for each combination of checkpoint, microbatch, and representation. The validation sequence offset is 1024. Before aggregation, we verify that validation examples and random draws are shared, paired gradients correspond, update norms are positive, the hierarchy has the required order, reconstructed functions are equivalent, and the exact and linear response identities hold.

The state histories come from trajectories with exact transport. At every training step, both native optimizers advance their own state while forming proposals. The canonical branch applies its native proposal; the transformed branch applies the exact pushforward of that canonical update and retains the state it has just advanced in its native coordinates. The weights thus remain on the prescribed orbit and follow a common function history, while each optimizer accumulates its gradient history in native coordinates along that path. The transformed optimizer constructs its state from this gradient history, without applying a group action to the canonical state.

At each of the five checkpoints, we use four paired microbatches of eight sequences starting at validation offset 4096. We load the weights related by the orbit action and their native optimizer states, then compute and separately clip the paired full-model gradients. Optimizers with stored state form one proposal from the loaded histories. Fresh optimizer objects repeat the proposal at identical weights, clipped gradients, schedule multiplier, and minibatch. Fresh construction resets all historical moment or momentum buffers and the step counter; forming the current proposal then creates the ordinary first-step state. We fit the same VO hierarchy to both proposal pairs without applying either. Thus $\rho_{\mathrm{cold}} > 0$ isolates the dependence of VO-product geometry on representation, while $\rho_{\mathrm{state}} - \rho_{\mathrm{cold}}$ measures how history in native coordinates modifies that residual.

E Metrics and statistical summaries

E.1 Local hierarchy

Each head contributes exact finite-step coefficients from Equation 13. The global fit pools all heads and observations; the layer fit pools observations within each layer; the head fit pools them within each head. Reported ρ always uses the canonical reference norm as its denominator. We accept a common fit only if the head, layer, and global objectives obey the required monotonic order. Signed scalars for individual batches and their coefficient of variation supplement the common exact residual.

For each optimizer, the local floor at the precision used in the experiment is the 90th percentile of the transported oracle's quotient error over 1,890 path measurements. A residual is at floor if it is no more than three times that value. This yields floor values 6.5853×10^{-6} for AdamW and 5.4147×10^{-6} for Muon.

A transform seed provides support across checkpoints when at least four of five checkpoints are consistently at floor or above floor. Support across seeds requires agreement among all four transform seeds. For the primary dense condition-four family, both optimizers are above floor at all five checkpoints for every transform seed.

E.2 TRAJECTORY DISTANCES

Let $\tilde{L}_i^c$ and $\tilde{L}_i^g$ be canonical and transformed losses on the shared evaluation grid after smoothing. The primary loss distance is

$$D_{\text{loss}} = \frac{1}{N} \sum_{i=1}^{N} |\tilde{L}_i^c - \tilde{L}_i^g|, \qquad N = 78.$$

The smoothing kernel has weights $[1, 2, 3, 2, 1]$; at endpoints, we renormalize the available weights. This distance measures departure from the canonical loss trajectory. A larger value can arise from either higher or lower losses and therefore does not by itself rank validation performance.

For function distance, we project logits at token positions $\{63, 127, 191, 255, 319, 383, 447, 511\}$ to 16 dimensions with one fixed random projection. We smooth the coordinates before computing distances. For projected vectors p_i, q_i, the pointwise symmetric relative distance is

$$d_i = \frac{\|p_i - q_i\|_2}{\sqrt{(\|p_i\|_2^2 + \|q_i\|_2^2)/2}}, \qquad D_{\text{function}} = \frac{1}{N} \sum_i d_i.$$

Loss evaluation uses 512 held-out sequences at each point. The projected function vector uses four fixed sequences at the eight stated positions, giving shape $4 \times 8 \times 16$ at every point. We also store a separate loss evaluation on 4,096 sequences at the final endpoint.

Four independently executed canonical replicas produce six pairwise distances, whose maximum defines the duplicate floor. The transported branch defines a second, causal floor; the effective floor is the larger of the two. Here the duplicate floors are exactly zero, while the nonzero transport floors reflect arithmetic and measurement error. We report relative closure only when the same-rate distance exceeds three times the effective floor.

E.3 MUTUALLY EXCLUSIVE CLASSIFICATION

We test classes in the fixed order Numerically Inconclusive, Scalar ELR, Layerwise ELR, and Effective Learning Geometry. Numerical interpretation first requires all function checks for exact transport to pass and the same-rate loss and function distances to be resolvable. Conditional on those checks:

- Scalar ELR requires global stateful stability at floor, local cross-seed global support at floor, and global-scalar loss and function paths at floor.
- Layerwise ELR requires failure of global closure but layerwise stateful stability, local cross-seed layer support, and layer-scalar loss and function paths at floor.
- Effective Learning Geometry requires stable layer-resolved stateful mismatch above floor, local cross-seed layer support above floor, and both global and layerwise scalar trajectories above floor.

We conservatively assign any unmatched combination of criteria to Numerically Inconclusive. The labels state-amplified, memory-attenuated, mixed, and none describe how optimizer state modifies the VO-product residual within each primary class. Both optimizers meet the Effective Learning Geometry criteria; the modifier is memory-attenuated for AdamW and mixed for Muon.

E.4 UNCERTAINTY AND REPLICATION UNIT

The local figures show medians and 10th–90th percentiles across four fixed transformation seeds. The analysis also stores nonparametric bootstrap intervals from 10,000 resamples using seed 173205. At TPP 20 for the primary family, the global residual medians and bootstrap intervals are $0.77338\,[0.77310, 0.77347]$ for AdamW and $0.44124\,[0.44104, 0.44143]$ for Muon. These intervals characterize transformation perturbations for the fixed checkpoint and microbatch set. They do not estimate uncertainty across training seeds, architectures, or datasets. Microbatches within a checkpoint share model weights, so we do not treat them as independent training replicates.

Table 2: Probes with cold optimizer state across transformation families at TPP 20. Global residuals report medians [10th, 90th percentiles] across four fixed transformation seeds. The remaining columns report medians over the same seeds. Each seed aggregates 32 paired microbatches.

Opt.	family	κ	global ρ	layer ρ	a	CV(a_b)
AdamW	orthogonal	1.00	0.698 [0.698, 0.698]	0.697	0.705	0.002
AdamW	dense_gl_k2	2.00	0.723 [0.723, 0.723]	0.723	0.630	0.003
AdamW	diagonal_gl_k2	2.00	0.176 [0.176, 0.176]	0.176	0.949	1.72×10^{-4}
AdamW	dense_gl_k4	4.00	0.773 [0.773, 0.773]	0.773	0.473	0.003
AdamW	diagonal_gl_k4	4.00	0.332 [0.332, 0.333]	0.332	0.820	4.59×10^{-4}
Muon	orthogonal	1.00	0.002 [0.002, 0.002]	0.002	1.000	1.42×10^{-6}
Muon	dense_gl_k2	2.00	0.241 [0.241, 0.242]	0.241	0.924	0.001
Muon	diagonal_gl_k2	2.00	0.242 [0.241, 0.242]	0.242	0.924	0.001
Muon	dense_gl_k4	4.00	0.441 [0.441, 0.441]	0.441	0.746	0.004
Muon	diagonal_gl_k4	4.00	0.441 [0.441, 0.442]	0.441	0.746	0.004

Table 3: Decomposition of the linear residual into direction and rate terms at TPP 20. Directional residuals ρ_b and scalars $a_b^\star$ report medians [10th, 90th percentiles] over all 128 combinations of microbatch and seed per row. SSE columns report the median fractions of common squared error across four seeds, in percent. The direction and rate fractions partition the common squared error and sum to 100%.

Optimizer	family	ρ_b	$a_b^\star$	direction (%)	rate (%)
AdamW	orthogonal	0.698 [0.696, 0.700]	0.705 [0.702, 0.706]	99.9995	0.0005
AdamW	diagonal, $\kappa = 2$	0.176 [0.175, 0.176]	0.949 [0.949, 0.950]	99.9999	9.08×10^{-5}
AdamW	diagonal, $\kappa = 4$	0.332 [0.332, 0.333]	0.820 [0.819, 0.820]	99.9998	0.0002
AdamW	dense, $\kappa = 2$	0.723 [0.722, 0.725]	0.630 [0.627, 0.631]	99.9994	0.0006
AdamW	dense, $\kappa = 4$	0.773 [0.772, 0.775]	0.474 [0.471, 0.475]	99.9993	0.0007
Muon	orthogonal	0.002 [0.002, 0.002]	1.000 [1.000, 1.000]	100.0000	4.11×10^{-5}
Muon	diagonal, $\kappa = 2$	0.241 [0.240, 0.243]	0.924 [0.923, 0.924]	99.9980	0.0020
Muon	diagonal, $\kappa = 4$	0.440 [0.439, 0.445]	0.747 [0.745, 0.748]	99.9940	0.0060
Muon	dense, $\kappa = 2$	0.241 [0.240, 0.243]	0.924 [0.923, 0.924]	99.9981	0.0019
Muon	dense, $\kappa = 4$	0.440 [0.439, 0.445]	0.747 [0.745, 0.748]	99.9940	0.0060

F Additional results

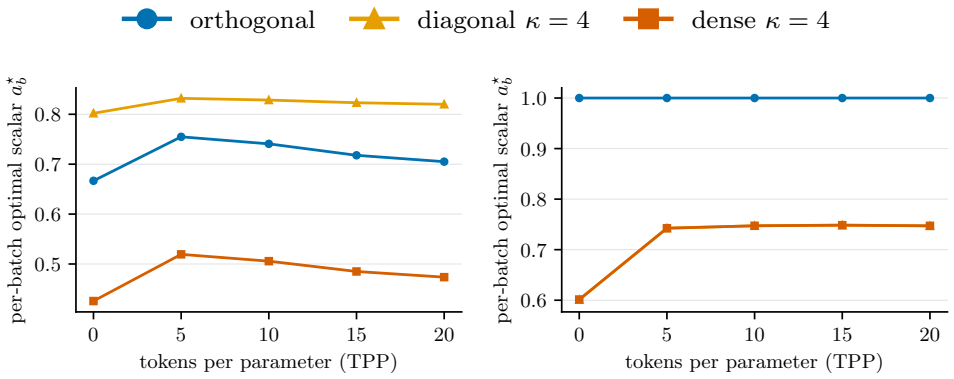
F.1 Sweep across transform families

Table 2 reports all five transform families and their coefficients of variation (CV). We define CV as the sample standard deviation of batchwise signed linear scalars divided by the absolute sample mean. The diagonal condition-two transform produces the smallest non-orthogonal residual at TPP 20 for AdamW. For Muon, diagonal and dense families nearly coincide at a fixed condition number. These differences show that the response also depends on axis orientation and the optimizer's coordinate structure.

F.2 Decomposition into direction and rate, and batchwise variation

Table 3 uses the linear VO response for each microbatch, concatenating product updates from all layers and heads into one vector. For each of the 200 cells, we use the 32 triples $(\|x_b\|^2, \langle x_b, y_b\rangle, \|y_b\|^2)$ to independently compute $a_b^\star$, the pooled $a^\star$, and both SSE terms in Equation 7. All 6,400 global observations have strictly positive norms and inner products, as do all 76,800 layer and 921,600 head observations. Thus the reported decomposition requires no boundary correction. A nonpositive inner product would require separate treatment of the constrained projection; a positive pooled scalar alone would not guarantee positive inner products for every observation.

We compute percentages within each seed and checkpoint cell, then take medians across seeds. The distribution columns pool the 128 observations across microbatches and seeds at TPP 20. Directional shares near 100% should be interpreted alongside residual magnitude: Muon's orthogonal residual is



(a) AdamW: optimal batchwise scalar.

(b) Muon: optimal batchwise scalar.

Figure 5: **Optimal batchwise rates across checkpoints.** The linear global VO diagnostic fits a separate scalar to each microbatch. Lines and bands show medians and 10th–90th percentiles over 128 observations (32 microbatches and four transform seeds) per checkpoint and family. They describe variation at fixed checkpoints; they do not estimate uncertainty across training seeds.

Table 4: Trajectory distances for the fixed dense GL transform ($\kappa = 4$). D averages distance between smoothed trajectories over the shared TPP grid. G and L fit global and layerwise scalars; Tr exactly transports the canonical update.

	Loss D				Projected-logit D			
Optimizer	same	G	L	Tr	same	G	L	Tr
AdamW	0.012	0.022	0.027	2.72×10^{-8}	0.402	0.431	0.431	8.66×10^{-7}
Muon	0.002	0.049	0.037	4.22×10^{-8}	0.337	0.365	0.376	1.58×10^{-6}

small even though most of its remaining error is directional. Figure 5 shows how optimal batchwise scalars evolve across checkpoints. Their narrow bands do not imply scalar sufficiency: the large dense-family ρ_b values in Table 3 persist after fitting each microbatch separately.

F.3 Trajectory summaries over the full evaluation grid

Table 4 averages gaps from the same smoothed trajectories as Figure 4, using the fixed analysis kernel over all 78 evaluation points, including initialization and the final checkpoint. The table also resolves the transport control, which lies near zero on the main figure's linear axes. AdamW's same-LR, global-scalar, and layerwise-scalar loss curves are similar overall, while all three retain substantial function distances. Muon's scalar interventions increase both distances from the canonical reference relative to its same-LR branch, even as their validation losses fall below the canonical reference later in training. The larger loss distances therefore reflect this greater departure from the reference trajectory. These averages complement the changes over time shown in the main figure.

F.4 Fitted scalar schedules

Figure 6 plots the applied global multipliers a_t and layerwise multipliers $a_{\ell,t}$ along their respective training paths. The global coefficients drop early. AdamW's coefficient subsequently rises toward 0.10, while Muon's ends near 0.03. Layerwise coefficients remain distinct: final values range from 0.0078 to 0.1623 for AdamW and from 0.0024 to 0.0659 for Muon.

F.5 Optimizer memory

Table 5 gives the checkpoint medians underlying Figure 3(c,d). AdamW's four post-initialization memory effects are all negative and tightly clustered. Muon's two positive and two negative post-

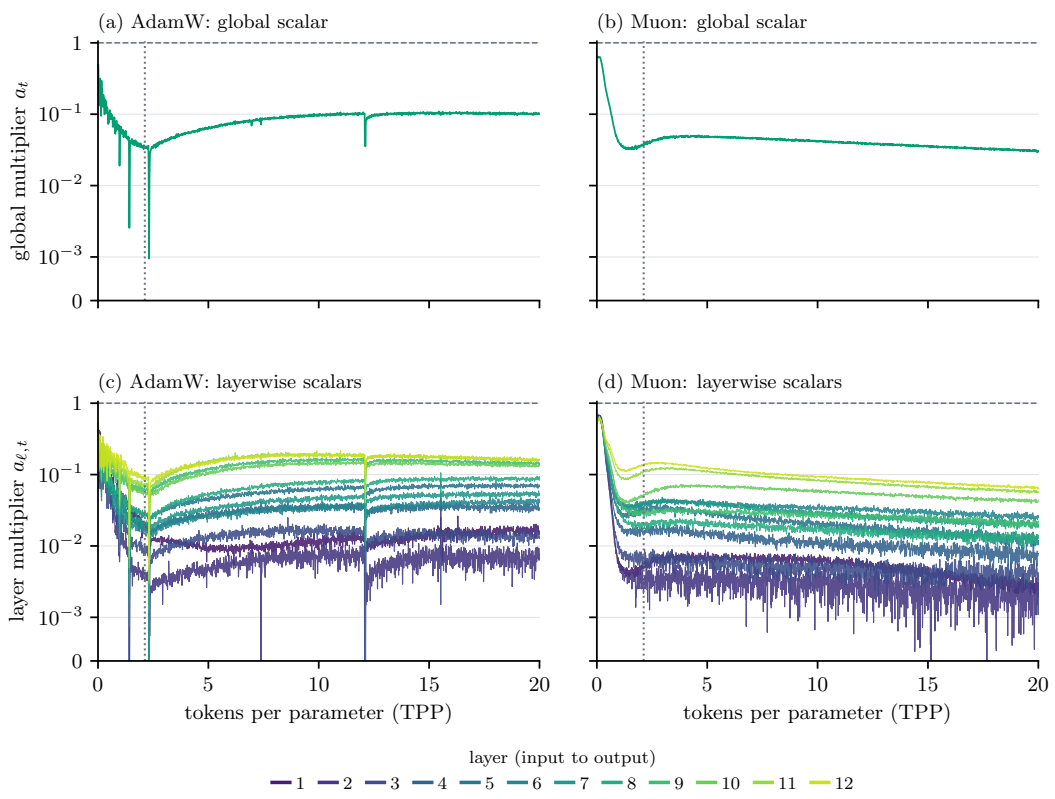


Figure 6: **Fitted scalar schedules in the training interventions.** (a,b) Global multipliers a_t. (c,d) One curve for each of the 12 layers for $a_{\ell,t}$, with colors identifying layers from input to output. Coefficients are recomputed at every update; lines connect all 1,890 logged values per branch, saved every 10 steps and at the first update and checkpoints. These are dimensionless multipliers of the native VO proposals. The vertical dotted line marks the end of warmup, and the horizontal dashed line marks a multiplier of one. Vertical axes are logarithmic above 10^{-3} and linear below it.

Table 5: Same-function state shadow. At each checkpoint, cold and stateful proposals use the same orbit-related weight pair, gradients, and minibatches. Entries are medians across four paired microbatches. The final column is the median paired difference $\rho_{\text{state}} - \rho_{\text{cold}}$, which need not equal the difference of column medians.

Optimizer	TPP	ρ_{cold}	ρ_{state}	state effect
AdamW	0	0.817	0.817	0
AdamW	5	0.731	0.505	-0.226
AdamW	10	0.751	0.523	-0.228
AdamW	15	0.764	0.538	-0.226
AdamW	20	0.774	0.549	-0.225
Muon	0	0.569	0.569	0
Muon	5	0.446	0.448	0.002
Muon	10	0.442	0.442	9.90×10^{-5}
Muon	15	0.442	0.441	-7.48×10^{-4}
Muon	20	0.443	0.442	-6.90×10^{-4}

initialization effects motivate the mixed label. At TPP 0, cold and stateful optimizer states coincide by construction.

Table 6: Scalar fits to the full output response. Values are medians over TPP $\{5, 10, 15, 20\}$ and four fixed gradient microbatches. All residuals use the canonical response norm $\|x\|$ as denominator; ρ_{tr} is the exact transport control.

Optimizer	a_{VO}	a_{out}	$\rho(a_{\mathrm{VO}})$	$\rho(a_{\mathrm{out}})$	ρ_{tr}
AdamW	0.492	0.718	0.192	0.102	2.85×10^{-3}
Muon	0.747	0.831	0.082	0.064	3.35×10^{-3}

Table 7: Scalar fits to the full output response at the same quotient point. Values are medians over four fixed gradient microbatches. The candidate uses corresponding updates outside VO and scales only the native VO proposal at the transformed weights. a_{VO} fits the exact finite-step displacement of the VO product; a_{out} fits the response of centered logits over the full vocabulary. The final column is the exact transport control.

Optimizer	TPP	a_{VO}	a_{out}	$\rho(a_{\mathrm{VO}})$	$\rho(a_{\mathrm{out}})$	ρ_{tr}
AdamW	0	0.426	0.832	0.332	0.076	2.86×10^{-4}
AdamW	5	0.522	0.764	0.193	0.098	2.75×10^{-3}
AdamW	10	0.500	0.727	0.187	0.093	2.80×10^{-3}
AdamW	15	0.485	0.710	0.191	0.106	2.95×10^{-3}
AdamW	20	0.472	0.705	0.193	0.106	2.94×10^{-3}
Muon	0	0.595	0.832	0.176	0.093	2.08×10^{-3}
Muon	5	0.741	0.828	0.080	0.057	3.47×10^{-3}
Muon	10	0.747	0.837	0.086	0.064	3.32×10^{-3}
Muon	15	0.747	0.825	0.083	0.068	3.35×10^{-3}
Muon	20	0.747	0.820	0.080	0.065	3.17×10^{-3}

F.6 Full output response to one update

Without additional training, we evaluate the ten saved checkpoints from trajectories with exact transport at TPP $\{0, 5, 10, 15, 20\}$. We fixed the design and numerical thresholds before computing responses. We reconstruct each transformed representative from the canonical model with the selected dense condition-four VO action; stored transformed weights are unused.

For each optimizer and checkpoint, four fixed microbatches of eight sequences produce proposals with cold optimizer state. The output batch contains a disjoint fixed set of four sequences at the eight trajectory positions. The candidate direction uses the exact corresponding canonical update outside VO and the native transformed proposal on VO slices. We clip canonical and transformed full-model gradients separately, matching the paired training setup.

Let $x = DF_\theta[\delta_c]$, $b = DF_{g\theta}[\delta_{\mathrm{nonVO}}]$, $y = DF_{g\theta}[\delta_{\mathrm{VO}}]$, and $t = DF_{g\theta}[g_*\delta_c]$, where F is the centered full-logit vector, DF denotes its directional derivative, δ_c is the canonical full increment, δ_{nonVO} is its corresponding non-VO increment, δ_{VO} is the native transformed VO increment, and $g_*\delta_c = Dg[\delta_c]$ is the transported full increment. The directly fitted scalar and its normalized residual are

$$a_{\mathrm{out}} = \frac{[\langle x - b, y\rangle]_+}{\|y\|^2}, \qquad \rho_{\mathrm{out}} = \frac{\|x - b - a_{\mathrm{out}} y\|}{\|x\|}.$$

Writing $r = x - b - a_{\mathrm{out}}y$, this factors as $\rho_{\mathrm{out}} = (\|x - b\|/\|x\|)(\|r\|/\|x - b\|)$. It combines residual noncollinearity with the VO response's scale relative to the complete response. The VO-product residual uses a different normalization, which prevents direct comparison. We also evaluate the output residual at the exact finite-step VO-product optimum a_{VO} and compare both fits with $\rho_{\mathrm{tr}} = \|x - t\|/\|x\|$. Directional derivatives use symmetric finite differences at global parameter-relative perturbations 2×10^{-5} and 10^{-5}; each direction has its own multiplier, which is divided out.

Table 6 summarizes the 16 post-initialization cells per optimizer; Table 7 includes every checkpoint, including initialization. Both tables report the output residuals in Figure 3(a,b), normalized by the norm of the complete response.

The maximum reconstructed same-function error is 1.88×10^{-6}, the maximum exact-transport response residual is 3.91×10^{-3}, the maximum cross-scale response disagreement is 0.014, and the largest change in either fitted residual is 9.41×10^{-4}. These are below the respective limits 10^{-4},

Table 8: Numerical controls with FP32 throughout. The path floor is the 90th percentile of quotient error under exact transport. The function and state columns report maximum relative errors in projected logits. The final two columns bound the difference between exact finite-step and linearized residuals for a common scalar.

Opt.	path floor	function	state	exact–lin. all	exact–lin. main
AdamW	6.59×10^{-6}	1.98×10^{-6}	1.91×10^{-6}	1.68×10^{-4}	6.24×10^{-5}
Muon	5.41×10^{-6}	4.17×10^{-6}	3.11×10^{-6}	1.52×10^{-5}	1.51×10^{-5}

Table 9: Checkpoint validation with full vocabulary logits on four sequences at eight positions. Values are medians over TPP $\{5, 10, 15, 20\}$. Full and proj. report relative L_2 distances between centered logits, using the full vocabulary and the primary projection to 16 dimensions, respectively. JS denotes Jensen–Shannon divergence. Spearman correlation for treatment ranking is $\rho = 0.667$, below the criterion of 0.95; the median proj./full ratio is 1.002.

	AdamW			Muon		
Treatment	full	proj.	JS	full	proj.	JS
same LR	0.469	0.467	0.063	0.440	0.454	0.069
global scalar	0.495	0.503	0.068	0.471	0.466	0.070
layerwise scalars	0.494	0.493	0.063	0.465	0.471	0.075
exact transport	1.18×10^{-6}	1.35×10^{-6}	1.69×10^{-8}	2.26×10^{-6}	2.46×10^{-6}	1.85×10^{-8}

0.05, 0.05, and 0.03. The medians of the transport residual, cross-scale response disagreement, and fitted-residual change also meet their common threshold of 0.01. No output fit reaches the nonnegative boundary. After initialization, the ratio of the output-optimal residual to the transport residual within each cell has medians 34.5 and 19.5 for AdamW and Muon, respectively. Refitting to a_{out} lowers the output residual but leaves it above transport. Different targets and denominators prevent interpreting the change across metrics as Jacobian compression. This experiment with cold optimizer state supports a local response claim; it does not test memory effects or subsequent trajectory states.

F.7 Numerical controls and finite-step sensitivity

Table 8 reports numerical controls and sensitivity to the finite-step or linear response. Checks pass for the transported function, state-probe equivalence, and error relative to a single update. Trajectory floors are orders of magnitude below the dense-transform residuals. The difference between exact and linear responses is too small to explain the observed mismatch.

F.8 Checkpoint validation over the full vocabulary

We also evaluate the saved checkpoints over the full vocabulary without further training. For every optimizer, treatment, and TPP checkpoint, we use four fixed validation sequences at eight trajectory token positions. The metrics are centered full-vocabulary relative L_2, symmetric KL, Jensen–Shannon (JS) divergence, and Rademacher projections with dimensions $\{8, 16, 32, 64\}$ and four fixed seeds.

Table 9 reports distances for full and projected centered logits alongside JS at individual checkpoints, without the smoothing used for the full trajectory evaluation grid.

The primary projection's median ratio to full L_2 is 1.002 (median absolute deviation 2.77%). For full logits, every post-initialization native/scalar cell is separated from exact transport by at least 1.47×10^5. Exact transport produces the smallest distance for both full logits and the primary projection in 8/8 optimizer–checkpoint groups. JS gives the same coarse separation.

The results do not meet the deliberately strict criterion for preserving fine rankings. The primary Spearman correlation across the 24 clustered native/scalar cells is 0.667, below the specified 0.95 threshold; the minimum across projection dimensions and seeds is 0.490. Within optimizer–checkpoint groups, the primary projection preserves 41/48 pairwise treatment orders, but JS is not monotone in centered-logit L_2. These metrics support the separation between numerical floors and macroscopic distances. They do not support fine ordering among native and scalar interventions.

Table 10: Sensitivity of trajectory distances to smoothing. The primary kernel is $[1, 2, 3, 2, 1]$, with weights renormalized over available points at the endpoints. Columns labeled 1, 3, and 7 use boxcar windows of those widths. Treatments use the same external learning rate (same), a global scalar (G), layerwise scalars (L), or exact transport (Tr). Across these smoothing choices, only exact transport reaches the effective numerical floor.

		Loss D				Projected-logit D			
Opt.	trt.	primary	1	3	7	primary	1	3	7
AdamW	same	0.012	0.013	0.012	0.012	0.402	0.454	0.413	0.388
AdamW	G	0.022	0.022	0.022	0.022	0.431	0.476	0.441	0.418
AdamW	L	0.027	0.027	0.027	0.027	0.431	0.477	0.441	0.419
AdamW	Tr	2.72×10^{-8}	6.23×10^{-8}	3.17×10^{-8}	1.96×10^{-8}	8.66×10^{-7}	1.26×10^{-6}	9.34×10^{-7}	7.84×10^{-7}
Muon	same	0.002	0.003	0.003	0.002	0.337	0.422	0.352	0.318
Muon	G	0.049	0.049	0.049	0.049	0.365	0.444	0.380	0.346
Muon	L	0.037	0.037	0.037	0.038	0.376	0.452	0.390	0.359
Muon	Tr	4.22×10^{-8}	7.28×10^{-8}	4.76×10^{-8}	3.43×10^{-8}	1.58×10^{-6}	2.12×10^{-6}	1.67×10^{-6}	1.45×10^{-6}

F.9 Trajectory smoothing sensitivity

Table 10 compares the primary kernel with boxcar windows one, three, and seven; the qualitative result is unchanged. For both optimizers, exact transport remains at the numerical scale, while the projected-logit distances for same-rate, global-scalar, and layer-scalar branches remain macroscopic. AdamW's small absolute loss distances remain consistent with similar overall validation-loss curves across these smoothing choices.

F.10 Positive control for scaling symmetry

The learning parameter is $w \in \mathbb{R}^{64} \setminus \{0\}$, and the predictor is $f_w(X) = Xw/\|w\|_2$. With a CPU random generator seeded at 223607, the control draws an initial weight $w_0 \in \mathbb{R}^{64}$, a fixed input matrix $X \in \mathbb{R}^{32\times64}$, and a target direction $v \in \mathbb{R}^{64}$, in that order, with independent standard-normal entries. Targets are $y = Xv/\|v\|_2$. It minimizes

$$L(w) = \frac{1}{2 \cdot 32}\|f_w(X) - y\|_2^2$$

using plain SGD in FP64, learning rate $\eta = 0.05$, no momentum, no weight decay, and the same fixed batch for 32 steps. For each $c \in \{0.5, 1, 2, 4\}$, the canonical branch starts at w_0 and both comparison branches start at cw_0. The same-LR branch uses η and the compensated branch uses $c^2\eta$.

At each step, all three branches independently compute their loss and gradient by automatic differentiation, then form their proposals. Before applying updates, we fit a scalar between the angular derivatives $D\pi_w[-\eta\nabla L(w)]$ and check it against c^2. After applying updates, we measure relative prediction error against the canonical predictor and orbit error $\|w_t^{\text{matched}} - cw_t\|_2/\|cw_t\|_2$.

The fitted multipliers $\{0.25, 1, 4, 16\}$ match the theory at every step. Over all 32 steps at each scale, the maximum absolute error in the scalar fit, angular residual ρ, relative prediction error, and relative orbit error are exactly zero in the stored FP64 records. The scales are powers of two, for which binary scaling can be exact; these zeros are present before rounding for display. Figure 1(b) plots the medians of the measured scalars. This verifies independent SGD compensation for a one-parameter group of positive scalings; the parameter space has 64 dimensions and its normalized direction has 63 degrees of freedom. The control uses scale-invariant SGD; it does not run the Hyperball optimizer.