WHICH GEOMETRY ON WHICH LAYER? MEASURED SIGN–SPECTRAL ASSIGNMENT IN MIXED-OPTIMIZER TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern training of large neural networks now commonly uses mixed recipes, assigning AdamW-like or sign-based updates to embedding and output layers and spectral updates to hidden matrices. These assignments are fixed by layer type for the entire run and are largely inherited heuristics that must be re-tuned across architectures and training regimes, without a clear criterion for selecting the appropriate geometry, that is, the norm under which the steepest descent step is taken. We turn the sign-versus-spectral assignment into a measurable, layerwise decision problem. For ideal memoryless steps, we derive the finite-batch geometry ratio $\mathcal{R}_{\mathrm{LB}}(W;B) := B_{\mathrm{spec}}(W)/B_{\mathrm{sign}}(W)$, where B_{spec} and B_{sign} are conservative descent-lemma lower bounds for the specified spectral and sign steps. Its closed form separates signal efficiency from a geometry-dependent global curvature factor. In the positive-signal regime, the same smoothness assumption gives two-sided bounds that determine which step has the larger conditional expected decrease when the ratio crosses explicit noise-dependent thresholds. We then derive a local directional quadratic score from gradient alignments and Hessian–vector products and use it in Generalized-Scion (G-Scion) to control layerwise routing. In a controlled input-rank experimental setup, including directional curvature matches the measured same-iterate ordering in all seven aggregate settings, whereas the curvature-free proxy selects the opposite geometry. On 31M GPT-2, the output score routes the output layer from spectral warmup to sign, and G-Scion improves validation loss over fixed Scion in memoryless and matched-momentum comparisons. ViT-B/16 results further show that a spectral edge assignment can outperform fixed Scion outside language modelling. These results suggest that optimizer geometry need not be hard-coded by layer type, and they are a first step toward understanding how it should be assigned.

1 INTRODUCTION

The training of large-scale neural networks has entered the era of mixed optimizers (Liu et al., 2025; Zeng et al., 2025). State-of-the-art recipes no longer apply a single update rule to every weight matrix; instead, they partition the model by layer type, typically applying AdamW or sign-based methods such as SignSGD (Kingma & Ba, 2015; Bernstein et al., 2018) to embedding and output layers, while reserving spectral methods like Muon (Jordan, 2024) for hidden weight matrices.

Current practice relies on inherited conventions, e.g., a recipe that works for a 7B language model is ported to a vision transformer or a multimodal architecture under the assumption that "embedding-like" and "hidden-like" roles remain invariant. When these recipes fail, practitioners must resort to exhaustive hyperparameter sweeps to re-validate the mix. This points to a gap in optimization theory: we lack a principled criterion that says which geometry belongs on which layer, how that preference evolves across training stages, and whether architectural choices alter the preference between geometries. Figure 1 shows why the stage question matters: at the output head of a 31M GPT-2, the measured comparison of the two geometries crosses from sign to spectral within the first 400 steps, while a gate that commits at step 150 keeps sign for the rest of training.

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.

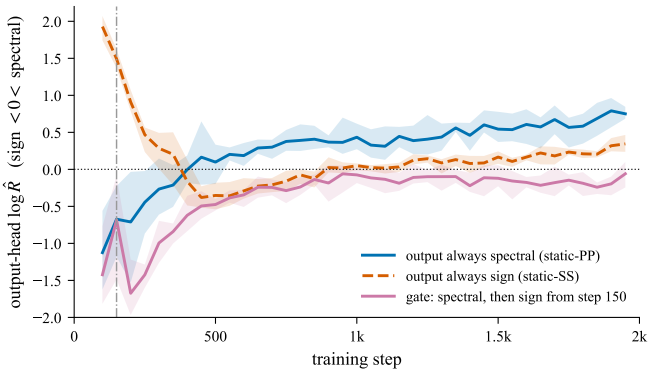

Figure 1: Output-head $\log \widehat{\mathcal{R}}_{\mathrm{dir}}$ logged every 50 steps along the static spectral/spectral trajectory, the static sign/sign trajectory, and the one-way gate of the follow-up (31M, 5 seeds; lines are means and bands are seed ranges). On the spectral trajectory the score crosses from sign to spectral by step 400, whereas the gate locks the output head to sign at step 150 (dash-dotted line) and its own reading afterwards stays between -0.3 and 0.

We argue that the choice between optimizer geometries is a measurable property of the local loss geometry whose usefulness must be tested against tuned static assignments. Every recipe that mixes sign-based and spectral methods makes an implicit layerwise assignment. We analyse this decision through the finite-batch geometry ratio[1]

$$\mathcal{R}_{\mathrm{LB}}(W; B) \;:=\; B_{\mathrm{spec}}(W) \,/\, B_{\mathrm{sign}}(W),$$

where B_{spec} and B_{sign} are conservative lower bounds for the conditional expected decreases of ideal memoryless spectral and sign steps. The ratio factorizes into a *signal-efficiency* term, governed by the gradient's norms and minibatch noise, and a *curvature ratio* κ formed from geometry-specific global smoothness constants. The factorisation yields testable predictions across layer type, training stage, model scale, and architecture. We evaluate the layer and stage predictions with the local directional score of Section 3, which operationalizes the signal–curvature view locally and is not an estimator of $\mathcal{R}_{\mathrm{LB}}$.

Contributions. We summarize our contributions as follows:

- **Analytical geometry comparison**: We derive a finite-batch ratio of geometry-specific lower bounds, factor it into signal efficiency and global curvature, and give explicit thresholds beyond which it orders the conditional expected decreases of the specified conservative steps (Section 2).
- **Measurement-driven assignment**: We derive an HVP-based local directional score and use it in G-Scion to replace Scion's fixed sign–spectral layer convention with online assignment (Section 3).
- **Geometry diagnostics and training outcomes**: Input-rank experiments show that directional curvature is needed to match same-iterate orderings; 31M language-model checkpoint comparisons show a stage-dependent preference at the output head; G-Scion improves over fixed Scion in the 31M and ViT settings; and a five-seed static comparison together with held-out update tests bounds what the score's accuracy implies for the resulting training policy (Section 4).

2 Geometry Comparison: Global Bounds and Directional Scores

Appendix A summarizes the notation.

[1]The subscript LB stands for lower bound: $\mathcal{R}_{\mathrm{LB}}$ is a ratio of two descent lower bounds, and it is distinct from the online score $\widehat{\mathcal{R}}_{\mathrm{dir}}$ of Section 3.

Table 1: Three optimizer geometries as norm-steepest descent.

Method	**Primal norm** $\|\cdot\|$	**Dual norm** $\|\cdot\|_*$	**Direction** $D^*(G)$
(S)GD	Frobenius $\|\cdot\|_{\mathrm{F}}$	Frobenius $\|\cdot\|_{\mathrm{F}}$	$G/\|G\|_{\mathrm{F}}$
SignSGD	Entrywise ℓ_∞	Entrywise ℓ_1	$\mathrm{sign}(G)$
SpecSGD	Spectral $\|\cdot\|_{\mathrm{op}}$	Nuclear $\|\cdot\|_{\mathrm{nuc}}$	$\mathrm{polar}(G) = UV^\top$

Let $W \in \mathbb{R}^{m\times n}$ be a weight matrix and $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ a differentiable loss with gradient $G = \nabla f(W)$, thin SVD $G = U\Sigma V^\top$, and $p = \min(m, n)$. Let $\widehat{G} = G + E$ be a minibatch estimate with additive noise E at batch size B.

2.1 Norm-steepest descent

Given a primal norm $\|\cdot\|$, the steepest-descent direction solves $D^*(G) \in \arg\max_{\|D\|\le 1} \langle G, D\rangle$, with $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. By duality, $\max_{\|D\|\le 1} \langle G, D\rangle = \|G\|_*$, the dual norm of G. Three geometries yield three familiar optimizers (Table 1).

If f is $L_{\|\cdot\|}$-smooth w.r.t. $\|\cdot\|$, the ideal scaled step $W^+ = W - \frac{\|G\|_*}{L_{\|\cdot\|}} D^*(G)$ satisfies

$$f(W^+) \;\le\; f(W) - \frac{\|G\|_*^2}{2\,L_{\|\cdot\|}}. \tag{1}$$

For the least norm-specific smoothness constants, $L_{\mathrm{F}} \le L_{\mathrm{op}} \le pL_{\mathrm{F}}$ and $L_{\mathrm{op}} \le L_\infty \le mnL_{\mathrm{F}}$ (Appendix B.1); hence $1 \le \kappa := L_\infty/L_{\mathrm{op}} \le mn$.

2.2 Stochastic Expected-Decrease Bounds

Let $D_g = D_g^*(\widehat{G})$ satisfy $\|D_g\|_g \le 1$ and define the conditional mean alignment $\mu_g := \mathbb{E}[\langle G, D_g\rangle \mid W]$. For any nonnegative scalar step η, norm-specific smoothness gives

$$\mathbb{E}[f(W) - f(W - \eta D_g) \mid W] \;\ge\; \eta\mu_g - \frac{L_g\eta^2}{2}. \tag{2}$$

The best nonnegative step in this lower model is $\eta = (\mu_g)_+/L_g$, yielding the lower bound $(\mu_g)_+^2/(2L_g)$. For $D_{\mathrm{sign}} = \mathrm{sign}(\widehat{G})$, the sign-bit alignment bound of Balles et al. (2020) gives $\mu_{\mathrm{sign}} \ge \|G\|_1 - 2\nu_1(W)$ with $\nu_1(W) := \mathbb{E}[\|\widehat{G} - G\|_1 \mid W]$, hence the conservative step $\eta_{\mathrm{sign}} = (\|G\|_1 - 2\nu_1(W))_+/L_\infty$ gives

$$B_{\mathrm{sign}}(W) \;:=\; \frac{\big(\|G\|_1 - 2\nu_1(W)\big)_+^2}{2L_\infty}. \tag{3}$$

For $D_{\mathrm{spec}} = \mathrm{polar}(\widehat{G})$, the elementary inequality $\langle G, \mathrm{polar}(G + E)\rangle \ge \|G\|_{\mathrm{nuc}} - 2\|E\|_{\mathrm{nuc}}$ (Appendix B.2) with $\nu_*(W) := \mathbb{E}[\|E\|_{\mathrm{nuc}} \mid W]$ gives

$$B_{\mathrm{spec}}(W) \;:=\; \frac{\big(\|G\|_{\mathrm{nuc}} - 2\nu_*(W)\big)_+^2}{2L_{\mathrm{op}}}, \tag{4}$$

with the analogous conservative spectral step $\eta_{\mathrm{spec}} = (\|G\|_{\mathrm{nuc}} - 2\nu_*(W))_+/L_{\mathrm{op}}$. Following Naganuma et al. (2026), write $c_1 := \sqrt{B}\,\nu_1$ and $c_* := \sqrt{B}\,\nu_*$. Treating these quantities as approximately batch-independent is a modelling assumption, and neither norm alone implies a universal small-batch advantage: the relative penalties $c_1/\|G\|_1$ and $c_*/\|G\|_{\mathrm{nuc}}$ determine the direction of the effect.

2.3 Definition and factorisation

Taking the ratio of the two lower-bound quantities in Eqs. (3)–(4) gives the finite-batch geometry ratio used in the theoretical factorisation.

Definition 1 (Finite-batch geometry ratio)**.** *For a layer with gradient G, batch size B, geometry-specific noise scales c_1, c_*, and smoothness constants $L_\infty, L_{\mathrm{op}}$,*

$$\mathcal{R}_{\mathrm{LB}}(W,B) \;:=\; \frac{B_{\mathrm{spec}}(W)}{B_{\mathrm{sign}}(W)} \;=\; \underbrace{\left(\frac{(\|G\|_{\mathrm{nuc}} - 2c_*/\sqrt{B})_+}{(\|G\|_1 - 2c_1/\sqrt{B})_+}\right)^2}_{\textit{signal efficiency}} \cdot \underbrace{\frac{L_\infty}{L_{\mathrm{op}}}}_{\textit{curvature ratio } \kappa}. \tag{5}$$

This ratio is defined when $B_{\mathrm{sign}} > 0$. If either lower bound is zero, the two nonnegative quantities should be compared directly rather than through a potentially undefined ratio.

$\mathcal{R}_{\mathrm{LB}}(W,B) > 1$ means that the spectral lower-bound quantity is larger than the sign lower-bound quantity at the same iterate; $\mathcal{R}_{\mathrm{LB}} < 1$ means the reverse. The signal-efficiency factor depends on the current gradient, while the global curvature ratio depends on the objective and parameterization.

Proposition 1 (Two-sided expected-decrease bounds)**.** *Assume that f is L_g-smooth with $0 < L_g < \infty$ under each geometry $g \in \{\mathrm{sign}, \mathrm{spec}\}$. Let $M_g = \|G\|_{g,*}$, let $\nu_{\mathrm{sign}} = \nu_1(W)$ and $\nu_{\mathrm{spec}} = \nu_*(W)$, and set*

$$S_g = M_g - 2\nu_g > 0, \qquad \rho_g = \frac{2\nu_g}{M_g}, \qquad \eta_g = \frac{S_g}{L_g}, \qquad B_g = \frac{S_g^2}{2L_g}. \tag{6}$$

For $D_g = D_g^*(\widehat{G})$, define the conditional expected decrease at this conservative step by $\Delta_g := \mathbb{E}[f(W) - f(W - \eta_g D_g) \mid W]$. Then

$$B_g \le \Delta_g \le c_g B_g, \qquad c_g = \frac{3-\rho_g}{1-\rho_g}. \tag{7}$$

Consequently, when both lower-bound quantities are positive,

$$\frac{\mathcal{R}_{\mathrm{LB}}}{c_{\mathrm{sign}}} \le \frac{\Delta_{\mathrm{spec}}}{\Delta_{\mathrm{sign}}} \le c_{\mathrm{spec}} \mathcal{R}_{\mathrm{LB}}. \tag{8}$$

Thus $\mathcal{R}_{\mathrm{LB}} > c_{\mathrm{sign}}$ guarantees that the specified spectral step has the larger conditional expected decrease, whereas $\mathcal{R}_{\mathrm{LB}} < 1/c_{\mathrm{spec}}$ gives the analogous guarantee for sign. At zero noise, these inequalities leave the ordering unresolved for $\mathcal{R}_{\mathrm{LB}} \in [1/3, 3]$.

The proof is given in Appendix J.

2.4 The Finite-Batch Ratio and Its Simplifications

The finite-batch geometry ratio $\mathcal{R}_{\mathrm{LB}}$ is operationally heavy because estimating $c_1, c_*, L_\infty, L_{\mathrm{op}}$ on the fly is expensive. Two simplifications drop terms under stated assumptions.

High-SNR limit $\mathcal{R}_\infty$. When $\|G\|_{\mathrm{nuc}} \gg 2c_*/\sqrt{B}$ and $\|G\|_1 \gg 2c_1/\sqrt{B}$, drop the noise corrections:

$$\mathcal{R}_\infty(W) \;:=\; \frac{\|G\|_{\mathrm{nuc}}^2}{\|G\|_1^2} \cdot \kappa \;=\; \frac{\mathrm{nr}(G)}{mn \cdot \phi_{\mathrm{ent}}(G)} \cdot \kappa, \tag{9}$$

where $\mathrm{nr}(G) = \|G\|_{\mathrm{nuc}}^2 / \|G\|_{\mathrm{F}}^2$ is the nuclear rank and $\phi_{\mathrm{ent}}(G) = \|G\|_1^2/(mn\,\|G\|_{\mathrm{F}}^2)$ is the entry-wise density. This simplification is valid only when both stated relative-noise terms are small. $\mathcal{R}_\infty$ is the high-SNR form of the global-smoothness lower-bound ratio and motivates the local directional score.

Curvature-free proxy $\mathcal{R}_0$. Drop the curvature ratio as well:

$$\mathcal{R}_0(W) \;:=\; \frac{\|G\|_{\mathrm{nuc}}^2}{\|G\|_1^2}. \tag{10}$$

The three forms therefore satisfy

$$\mathcal{R}_{\mathrm{LB}}(W;B) \;=\; \mathcal{R}_\infty(W) \cdot [\text{noise correction factor}], \qquad \mathcal{R}_\infty(W) \;=\; \mathcal{R}_0(W) \cdot \kappa.$$

For nonzero G, $\mathcal{R}_0 \le 1$ because $\|G\|_{\mathrm{nuc}} \le \|G\|_1$; raw $\mathcal{R}_0$ cannot favor spectral at a unit threshold. A dimension-normalized proxy $I_{\mathrm{proxy}} = \sqrt{mn\,\mathcal{R}_0}$ is a different score: its unit-threshold decision is equivalent to substituting $\kappa = mn$, not to estimating curvature. Thus raw and normalized proxy readings must not be interchanged. $\mathcal{R}_0$ approximates $\mathcal{R}_\infty$ only when $\kappa \approx 1$; the input-rank experiment in Appendix C tests measured directional curvature, not the global κ.

Algorithm 1 G-Scion

Input: cadence N, samples K, threshold τ
margin m, EMA factor β, confirmations q, warmup t_{w}, eligible set L_{elig}, initial $\{g_\ell\}$
$L_{\mathrm{lock}} \leftarrow \emptyset$; $h_\ell \leftarrow 0$ for $\ell \in L_{\mathrm{elig}}$
for training step $t = 1, 2, \ldots$ **do**
  Take an update using current g_ℓ
  if $t \geq t_{\mathrm{w}}$ and $t \bmod N = 0$ **then**
    for each $\ell \in L_{\mathrm{elig}} \setminus L_{\mathrm{lock}}$ **do**
      estimate a valid $\widehat{\mathcal{R}}_{\mathrm{dir}\ell}$
      $z_\ell \leftarrow \mathrm{EMA}_\beta(\log \widehat{\mathcal{R}}_{\mathrm{dir}\ell})$
      if $g_\ell = \textsc{Sign}$ **then**
        $b_\ell \leftarrow [z_\ell > \tau + m]$
      else
        $b_\ell \leftarrow [z_\ell < \tau - m]$
      end if
      $h_\ell \leftarrow h_\ell + 1$ if b_ℓ, else $h_\ell \leftarrow 0$
      if $h_\ell \geq q$ **then**
        flip g_ℓ; $L_{\mathrm{lock}} \leftarrow L_{\mathrm{lock}} \cup \{\ell\}$
      end if
    end for
  end if
end for

2.5 Architecture Dependence and Predictions

Davis & Drusvyatskiy (2025, Sec. 7) relate feature-based spectral and Frobenius majorizer coefficients through the stable rank of the incoming activations. Their analysis does not identify L_∞, so it does not determine $\kappa = L_\infty / L_{\mathrm{op}}$; we use it only as motivation for measuring architecture-dependent directional curvature (Appendices D and J). The factorisation predicts that the measured score can depend on layer, training stage, scale, and architecture. Section 4 tests layer and stage dependence; scale and architecture dependence remain predictions for matched experiments that jointly record the score, the realised route, and static geometry controls.

3 G-Scion: Measurement-Driven Geometry Assignment

The ratio $\mathcal{R}_{\mathrm{LB}}$ provides the analytical decomposition, but evaluating its global constants L_∞ and L_{op} online is expensive and conservative. G-Scion operationalizes the same signal–curvature view with a local directional quadratic score and uses that measurable quantity to route each eligible layer between sign and spectral updates.

3.1 Directional Quadratic Score $\widehat{\mathcal{R}}_{\mathrm{dir}}$

For a deterministic quadratic objective, a fixed direction D, alignment $s(D) := \langle G,\, D \rangle$, and positive directional curvature $C(D) := \langle D,\, \nabla^2 f(W)[D] \rangle > 0$,

$$f(W) - f(W - \eta D) = \eta s(D) - \tfrac{1}{2}\eta^2 C(D).$$

The best nonnegative line-search step is $\eta_D^* = (s(D))_+ / C(D)$ and its exact decrease is $(s(D))_+^2 / (2C(D))$. Motivated by this identity, the implementation averages stochastic alignment and HVP probes to obtain $\widehat{s}_g$ and $\widehat{C}_g$ for $g \in \{\mathrm{sign}, \mathrm{spec}\}$ and computes

$$\widehat{\mathcal{R}}_{\mathrm{dir}}(W) \;:=\; \frac{(\widehat{s}_{\mathrm{spec}})_+^2 \,/\, \widehat{C}_{\mathrm{spec}}}{(\widehat{s}_{\mathrm{sign}})_+^2 \,/\, \widehat{C}_{\mathrm{sign}}}, \qquad \widehat{C}_{\mathrm{spec}}, \widehat{C}_{\mathrm{sign}} > 0, \quad (\widehat{s}_{\mathrm{sign}})_+ > 0. \tag{11}$$

For fixed directions on a deterministic positive-curvature quadratic, Eq. (11) equals the ratio of the two exact line-optimal decreases because their common factor $1/2$ cancels. In stochastic training, $\widehat{\mathcal{R}}_{\mathrm{dir}}$ is a Monte Carlo local-quadratic routing score. It need not equal the global-smoothness ratio

$\mathcal{R}_{\mathrm{LB}}$, because directional curvature need not attain the global constants; Appendix E.1 gives the estimator details and safeguards.

3.2 G-SCION: ONE-WAY PER-LAYER SWITCHING

In the 31M experiments, G-Scion uses $N{=}50$, $K{=}12$, nominal threshold $\tau{=}0$, log margin $m{=}0.1$, EMA factor $\beta{=}0.9$, and two consecutive confirmations. Eligible groups are forced to spectral during warmup and may then flip once; hidden layers remain spectral following the matched Scion design. The variants *G-Scion (embedding)*, *G-Scion (output)*, and *G-Scion (embedding, output)* differ only in the eligible edge groups. Full settings are in Appendix E.3.1.

One-way locking is an anti-oscillation safeguard (Algorithm 1). Diagnostic cost grows with the probe count and the number of unlocked groups and is amortized by the cadence; a separate end-to-end timing experiment with $N{=}100$ measured approximately 4% overhead.

3.3 RELATIONSHIP TO EXISTING MIXED-OPTIMIZER RECIPES

Scion is the matched sign-versus-spectral baseline, since replacing only its assignment mechanism isolates the binary decision studied here. Practical Muon recipes place AdamW on embedding and output groups and add momentum and coordinate-wise adaptivity, so we report AdamW+Muon as a practical reference and not as a like-for-like baseline.

4 LM PRE-TRAINING EXPERIMENTS

We evaluate the layer- and stage-dependence predictions on GPT-2 pre-training, where shared hidden architecture makes per-layer diagnostics clean.

Setup. We train a 31M-parameter NanoGPT model on OpenWebText with four variants. **Scion** fixes SignSGD on embedding/output and Muon on hidden matrices. **G-Scion (embedding, output)**, **G-Scion (embedding)**, and **G-Scion (output)** use the gate of Section 3 on the indicated edge group(s); all other assignments match Scion.

Results. The like-for-like baseline is fixed Scion. Table 2(a) shows that G-Scion lowers validation loss relative to Scion both without momentum and with matched Signum momentum, and Table 2(b) shows that the embedding route accounts for most of the memoryless gain, with the joint variant lowest for every activation (Figure 2). These comparisons measure the resulting assignments. Held-out tests of single scheduled updates (Appendix H) and a five-seed comparison of all four static assignments against the gate (Appendix I) bound what those gains imply: the score misranks a share of single updates, and the gate does not reach the best static assignment. AdamW+Muon remains a stronger practical reference outside the memoryless comparison (Appendix E.3.3).

The 31M output supplies a sign-side routing case. The eligible groups are forced to spectral during warmup, so the first embedding readings depend on the gated variant and are not used as selector evidence. All six gated-output runs route the output to sign at $t{=}150$ and remain there (Figure 3). The gate acts on two confirmations of an EMA-smoothed log score, and Appendix E.3.3 separates this quantity from the pooled raw-score summaries; the historical proxy-run comparison is not used as evidence because its normalization and gate configuration remain unverified. The same-iterate experiment in Appendix C.2 isolates the directional-curvature factor within $R_{\mathrm{quad}} = \mathcal{R}_0 \hat{\kappa}_{\mathrm{dir}}$. Table 7 reports a phase-stratified sign change at the highest learning rate, which is an aggregate across runs and not a trace of a second switch.

Static assignments and stage dependence. A separate five-seed follow-up trains all four static edge assignments alongside the gate (Appendix I, Table 10). Spectral updates on both edge groups give the lowest final loss in every seed, and the gate, which locks the output head to sign at step 150, finishes about half a nat higher. Paired held-out decreases on three diagnostic seeds agree with the early sign recommendation and then favour spectral from about step 700, so the locked assignment is stale for most of the run. Figure 1 shows the same reversal in the output-head score logged along the

Table 2: GPT-2/OWT, 31M (2,000 steps, lr 10^{-3}, effective batch 128). (a) Resulting G-Scion versus fixed-Scion trajectories with memoryless SignSGD and with matched Signum momentum on the edge groups (GELU, 3 seeds); Muon is fixed on hidden matrices. (b) Edge-group decomposition in the memoryless setting, averaged over GELU, ReLU^2, and SiLU at seed 1337. Δ is the validation-loss reduction relative to the matched Scion row.

(a) Final val loss, sign-vs-spectral by momentum regime.

Method	Final val loss	Δ vs. matched Scion
w/o Momentum (SignSGD on embed/head, Muon on hidden)		
Scion w/o Momentum	5.529 ± 0.012	—
G-Scion w/o Momentum	$\mathbf{5.174 \pm 0.011}$	$\mathbf{+0.355}$
w/ Momentum (Signum on embed/head, Muon on hidden)		
Scion w/ Momentum	5.084 ± 0.013	—
G-Scion w/ Momentum	$\mathbf{4.955 \pm 0.004}$	$\mathbf{+0.129}$

(b) Final val loss, w/o Momentum (mean over GELU, ReLU^2, SiLU; $\widehat{\mathcal{R}}_{\mathrm{dir}}$).

Method	Final val loss	Δ vs. Scion
Scion w/o Momentum	5.516	—
G-Scion w/o Momentum (embedding)	5.293	+0.22
G-Scion w/o Momentum (output)	5.447	+0.07
G-Scion w/o Momentum (embedding, output)	**5.187**	**+0.33**

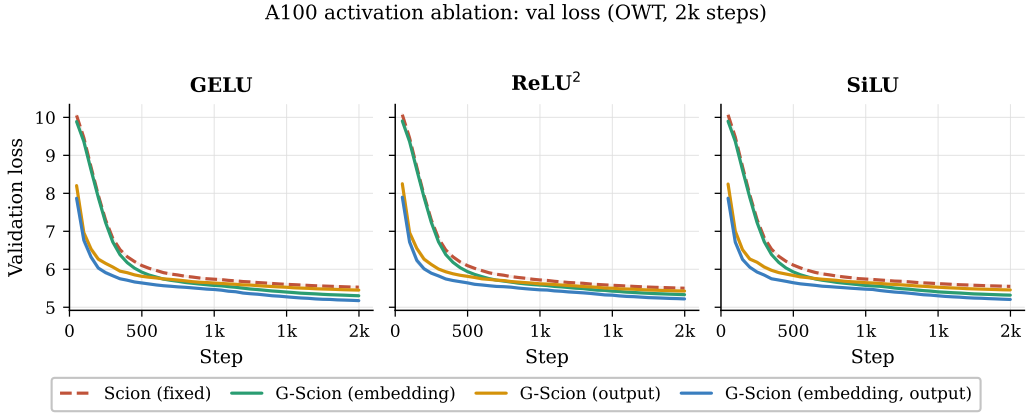

Figure 2: **Validation loss curves** (OWT, 31M, 2 000 steps, $\widehat{\mathcal{R}}_{\mathrm{dir}}$) for all three MLP activations. Colours: G-Scion (embedding, output) (blue), G-Scion (embedding) (green), G-Scion (output) (orange), Scion (vermilion). Ordering is preserved across activations; G-Scion (embedding, output) leads in every panel.

static trajectories. A warm-up variant and a frozen replay of its route reach the same rounded loss, which separates the score's local ranking from the value of ongoing adaptation.

Scale evidence. The available larger-scale runs do not form a matched score-and-assignment test: the 1B panel used the curvature-free proxy, and several 163M/406M method–seed cells lack complete directional records. We therefore treat scale dependence as a prediction and specify the matched directional-score, realised-route, and static-control comparison in Appendix E.3.4.

5 Cross-Architecture Assignment Evidence

ViT assignment outcome. We apply the same four edge-assignment variants to ViT-B/16 trained from scratch on CIFAR-100. At learning rate 3×10^{-3}, the joint spectral-edge trajectory exceeds fixed Scion by 11.9, 11.5, and 7.9 validation-accuracy points at effective batches 128, 256, and 512,

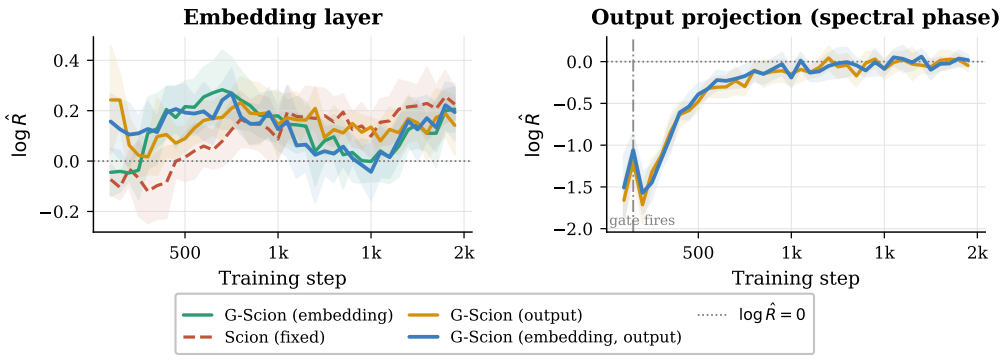

Figure 3: **Per-layer** $\log \widehat{\mathcal{R}}_{\mathrm{dir}}$ **diagnostics** (OWT, 31M, 2,000 steps; mean $\pm$ std across three activations per method). *Left:* embedding diagnostics are descriptive because eligible embeddings were already forced spectral during warmup and their first readings differ by variant. *Right:* output-gated variants start spectral with a clearly negative score, satisfy the two-confirmation rule at t=150 (dashed), and route to sign. Scion is diagnostic only and does not gate.

respectively (Table 8). This shows that Scion's inherited sign assignment on embedding/output is not uniformly best outside language modelling. The ViT score readings are frequently nonfinite, however, so this experiment is assignment-outcome evidence rather than validation of $\widehat{\mathcal{R}}_{\mathrm{dir}}$ or of online gating.

Weight-tied VLM scope. In nanoVLM (HuggingFace, 2025), the language embedding and output head are aliases of one shared parameter. A valid geometry comparison must therefore route the shared tensor exactly once and measure its joint gradient and HVP, rather than treating "embedding" and "head" as independent optimizer groups. The historical configurations did not realise the intended matched sign-versus-spectral head comparison and are excluded from the empirical claims. A corrected VLM evaluation requires explicit alias/identity logging, one realised optimizer route for the shared tensor, static sign and spectral controls, and a jointly measured shared-tensor score.

6 Related Work

Spectral and sign geometries. Most modern optimizers (Duchi et al., 2011; Kingma & Ba, 2015; Martens & Grosse, 2015; Gupta et al., 2018) are preconditioned Euclidean SGD; a complementary line targets the geometry of matrix-valued updates (Carlson et al., 2015), with Muon (Jordan, 2024) as a recent instantiation that scales to LLMs (Liu et al., 2025; Zeng et al., 2025). On the sign side, SignSGD (Bernstein et al., 2018) and Adam's entrywise-steepest interpretation (Balles et al., 2020) are the foundation; we use the sign-alignment lemma of Balles et al. (2020) in Eq. (3), and the non-Euclidean noise scales of Naganuma et al. (2026) cover finite batch size.

Stable rank and activations. Davis & Drusvyatskiy (2025) link a spectral-vs-GD quadratic-model comparison to gradient nuclear rank and post-activation stable rank, which ties activation choice (Hendrycks & Gimpel, 2016; Ramachandran et al., 2018; So et al., 2021) to optimizer suitability. Their result does not determine the sign-vs-spectral ratio L_∞/L_{op}, so our factorisation is derived from the geometry-specific lower bounds together with an empirically evaluated directional score. A causal activation test would vary the activation alone while measuring the gradient, score, and realised route jointly.

7 Discussion

Single-step prediction and training outcomes. We separate two questions. Does the score identify the better single step, and does following it improve the full run? On the first, the follow-up's early sign recommendations at the output head agree with paired held-out decreases, and the better candidate changes later in training (Appendix I); in the scheduled-update development study the

full-group score picks the better held-out step in about three quarters of the pairs, below always-spectral, with its errors concentrated in the sign-side bin (Table 9). This supports stage-dependent preferences and not universal calibration. On the second, the static spectral assignment beats the gate, which locks before the preference changes, and warm-up gating and frozen replay reach the same endpoint. The measurement and the routing rule therefore need to be separated, and the roles of warm-up, smoothing and locking need controlled ablations.

Analytical and operational roles. Proposition 1 orders conditional expected decreases for ideal memoryless steps at a common iterate. Trajectory dominance and joint optimality are separate questions, because the implementation adds stochastic estimation, scheduling, approximate polar factors, and simultaneous block updates. AdamW+Muon is a practical reference and not a theoretical upper bound (Sato et al., 2025).

Practical comparison and theory scope. The AdamW-edge/Muon reference reaches a lower validation loss than G-Muon (Table 6), which remains relevant even though AdamW lies outside the memoryless theorem; equal-budget tuning has not been done. At zero noise the bound certifies a spectral decision only for $\mathcal{R}_{\mathrm{LB}} > 3$ and a sign decision only for $\mathcal{R}_{\mathrm{LB}} < 1/3$, and the interval widens with noise. No $\mathcal{R}_{\mathrm{LB}}$ values were estimated in the LM study, so its score bins are not certification regions. Frobenius pairwise validation remains incomplete (Appendix J).

Limitations. (i) The layerwise score omits cross-block curvature, so single-block predictions need not identify the best joint assignment. (ii) One-way locking can preserve an obsolete assignment, as the follow-up shows, and allowing reversal alone did not recover a switch. (iii) Directional-score evidence is currently concentrated at 31M; scale and VLM generalisation require the matched comparisons described above. (iv) $\widehat{\mathcal{R}}_{\mathrm{dir}}$ requires Hessian–vector products, and the overhead varies with cadence, probes, unlocked groups, model, and implementation (Section 3). (v) ViT currently provides assignment-outcome evidence rather than validation of its directional-score trace. (vi) The follow-up reports five seeds for four static routes and the gate, and three for checkpoint diagnostics; other cells use two or three. Broader robustness and threshold calibration remain open.

8 Conclusion

G-Scion studies measured sign-versus-spectral assignment. The finite-batch lower-bound ratio provides sufficient expected-decrease conditions for specified conservative memoryless steps; it does not certify the implemented gate at a threshold of one. The directional score has a line-optimal quadratic interpretation, but our paired scheduled-update development results show substantial ranking errors. An actual-displacement predictor is promising on the same development data and remains unconfirmed. Reported gains over fixed Scion do not establish superiority over the best static assignment or AdamW/Muon. The present evidence therefore supports a diagnostic framework and motivates independent calibration and controlled assignment studies, rather than a general claim of optimal routing or practical optimizer superiority.

AI use statement

Generative AI was used to polish writing, write code and run experiments, and adapt to the ICLR latex template.

Ethics statement

The work studies optimizer geometry and training dynamics using public datasets, public model architectures, and controlled training experiments. It does not involve human subjects, crowdsourcing, private user data, deception, or deployment of high-risk systems. The paper's claims are technical and are framed within the limitations of the experiments.

Reproducibility statement

The paper specifies the datasets, model families and sizes, optimizer variants, edge groups, cadence N, probe count K, margin, EMA, confirmation rule, learning rates, batch sizes, training steps, seeds, and the alignment/HVP quantities used by the directional score. Section E records the provenance limits of the larger-scale, ViT-score, and historical VLM artifacts. Code-level reproduction is limited by the incomplete submitted repository, as acknowledged in Appendix E.2; the follow-up's per-seed records and full configuration mapping also remain to be archived (Appendix I).

REFERENCES

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. doi: 10.48550/arXiv.2502.02737. URL https://arxiv.org/abs/2502.02737.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *ArXiv*, 2024. doi: 10.48550/arXiv.2409.20325.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning (ICML)*, 2018.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *Artificial intelligence and statistics*, pp. 111–119. PMLR, 2015.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and H. Shi. Clarifying shampoo: Adapting spectral descent to stochasticity and the parameter trajectory. *ArXiv*, 2026. doi: 10.48550/arXiv.2602.09314.

Kevin Frans, Pieter Abbeel, and Sergey Levine. What really matters in matrix-whitening optimizers? *ArXiv*, 2025. doi: 10.48550/arXiv.2510.25000.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

HuggingFace. nanovlm: A minimal vision-language model, 2025. URL https://github.com/huggingface/nanoVLM. GitHub repository.

Keller Jordan. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/. Blog post.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2015. Published as conference paper at ICLR 2015.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *ArXiv*, 2025. doi: 10.48550/arXiv.2503.12645.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *ArXiv*, 2025. doi: 10.48550/arXiv.2510.05491.

Jingyuan Liu, Jianling 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, Meng 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*, 2025. doi: 10.48550/arXiv.2502.16982.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. COSMOS: A hybrid adaptive optimizer for efficient training of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j2QTOOtM8R.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Hiroki Naganuma, Shagun Gupta, Youssef Briki, Ioannis Mitliagkas, Irina Rish, Parameswaran Raman, and Hao-Jun Michael Shi. Adaptive batch sizes using non-euclidean gradient noise scales for stochastic sign and spectral descent. *arXiv preprint arXiv:2602.03001*, 2026.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Shikai Qiu, Zixi Chen, Hoang Phan, Qi Lei, and Andrew Gordon Wilson. Hyperparameter transfer enables consistent gains of matrix-preconditioned optimizers across scales. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Ei6IsmxYrb.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *ICLR Workshop*, 2018.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *ArXiv*, 2025. doi: 10.48550/arXiv.2509.01440.

Essential AI Darsh J Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, A. Tanwer, Darsh J Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, T. Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of muon for pretraining. *ArXiv*, 2025. doi: 10.48550/arXiv.2505.02222.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *ArXiv*, 2025. doi: 10.48550/arXiv.2507.11005.

David R. So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. Primer: Searching for efficient transformers for language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Akiyoshi Tomihari and Issei Sato. Understanding why adam outperforms sgd: Gradient heterogeneity in transformers. *arXiv e-prints*, pp. arXiv–2502, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. doi: 10.48550/arXiv.2502.14786. URL https://arxiv.org/abs/2502.14786.

A. Veprikov, Arman Bolatov, Samuel Horv'ath, A. Beznosikov, Martin Takác, and S. Hanzely. Preconditioned norms: A unified framework for steepest descent, quasi-newton and adaptive methods. *ArXiv*, 2025. doi: 10.48550/arXiv.2510.10777.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and S. Kakade. Soap: Improving and stabilizing shampoo using adam. *ArXiv*, 2024. doi: 10.48550/arXiv.2409.11321.

Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Weinan E, and Lei Wu. The sharpness disparity principle in transformers for accelerating language model pre-training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DZ6iFdVDrx.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms adam in tail-end associative memory learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=twbMFL0DMp.

Kaiyue Wen, David Leo Wright Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2J51qUZ0iG.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PUnD86UEK5.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=X6rqEpbnj3.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham M. Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zfeso8ceqr.

APPENDIX CONTENTS

A Notation 13
B Convergence Theory 16
C $\mathcal{R}_0$ Approximation Bounds 17
D Activation-Dependent Stable Rank Theory 21
E Experimental Details 23
F Comprehensive Related Work 28
G Compute Resources 30

A NOTATION

This appendix collects the notation used throughout the paper, grouped by category. All definitions are consistent with the main-text usage in Section 2.

A.1 PARAMETERS, GRADIENTS, AND BATCH QUANTITIES

Symbol	Meaning
$W \in \mathbb{R}^{m\times n}$	weight matrix of a given layer
m, n	output / input dimension of the layer
$p := \min(m, n)$	smaller side of the matrix
$f : \mathbb{R}^{m\times n} \to \mathbb{R}$	(possibly regularised) training loss at parameter W
$G := \nabla f(W)$	full-batch (population) gradient at W
$G = U\Sigma V^\top$	thin SVD of G, with $\Sigma \in \mathbb{R}^{p\times p}$
$\widehat{G} = G + E$	minibatch gradient estimate; E is mean-zero additive noise
B	minibatch (or effective-batch) size used to form $\widehat{G}$
K	number of independent minibatch gradients used by $\widehat{\mathcal{R}}_{\mathrm{dir}}$ (12 in the reported 31M runs)
N	decision cadence of the G-Scion gate (50 steps in the reported 31M runs)
t	training step index
τ	nominal switching threshold in $\log \widehat{\mathcal{R}}_{\mathrm{dir}}$ (0)
m	log-score margin (0.1 in the reported 31M runs)
β	EMA factor (0.9 in the reported 31M runs)
q	required consecutive confirmations (2 in the reported 31M runs)

A.2 NORMS, INNER PRODUCTS, AND MATRIX OPERATORS

Symbol	Meaning
$\langle A,\, B\rangle := \mathrm{tr}(A^\top B)$	Frobenius inner product
$\|\cdot\|_{\mathrm{F}}$	Frobenius norm, $\|A\|_{\mathrm{F}} = \sqrt{\langle A,\, A\rangle}$
$\|\cdot\|_{\mathrm{op}} = \|\cdot\|_{\mathrm{op}}$	spectral (operator) norm, largest singular value
$\|\cdot\|_{\mathrm{nuc}} = \|\cdot\|_*$	nuclear (trace) norm, sum of singular values
$\|\cdot\|_1 = \|\cdot\|_1$	entrywise ℓ_1 norm, $\sum_{ij} \lvert A_{ij}\rvert$
$\|\cdot\|_\infty$	entrywise ℓ_∞ norm, $\max_{ij} \lvert A_{ij}\rvert$
$\|\cdot\|_*$	dual norm of $\|\cdot\|$
$\mathrm{sign}(G)$	entrywise sign of G (with $\mathrm{sign}(0) := 0$)
$\mathrm{polar}(G) = UV^\top$	polar factor of G (partial isometry from its SVD)
$\mathrm{tr}(\cdot)$	matrix trace
$\mathrm{rank}(\cdot)$	matrix rank

The three geometries of interest correspond to the pairs $(\|\cdot\|, \|\cdot\|_*) = (\|\cdot\|_{\mathrm{F}}, \|\cdot\|_{\mathrm{F}})$ for (S)GD, $(\|\cdot\|_\infty, \|\cdot\|_1)$ for SignSGD, and $(\|\cdot\|_{\mathrm{op}}, \|\cdot\|_{\mathrm{nuc}})$ for SpecSGD; see Table 1.

A.3 SMOOTHNESS CONSTANTS AND THE CURVATURE RATIO

Symbol	Meaning
$L_{\|\cdot\|}$	smoothness constant of f w.r.t. a primal norm $\|\cdot\|$
L_{F}	Frobenius smoothness constant
L_{op}	operator (spectral) smoothness constant
L_∞	entrywise ℓ_∞ smoothness constant
$\kappa := L_\infty / L_{\mathrm{op}}$	*curvature ratio* (central configuration-side quantity)
$C(D) := \langle D,\, \nabla^2 f(W)[D] \rangle$	directional curvature along unit direction D

The smoothness constants satisfy $L_{\mathrm{op}} \le p \cdot L_{\mathrm{F}}$ and $L_{\mathrm{F}} \le L_\infty$ (Appendix B.1, Proposition 2).

A.4 UPDATE DIRECTIONS AND ONE-STEP IMPROVEMENTS

Symbol	Meaning
$D^*(G) \in \arg\max_{\|D\| \le 1} \langle G,\, D \rangle$	norm-steepest-descent direction for primal norm $\|\cdot\|$
$D_{\mathrm{sign}} := \mathrm{sign}(\widehat{G})$	SignSGD update direction
$D_{\mathrm{spec}} := \mathrm{polar}(\widehat{G})$	SpecSGD / Muon-like update direction
W^+	next iterate after an ideal scaled step from W
$\Delta_g(W;\eta)$	conditional expected one-step decrease under geometry g at step η
$B_{\mathrm{sign}}(W)$	SignSGD conservative lower bound (Eq. (3))
$B_{\mathrm{spec}}(W)$	SpecSGD conservative lower bound (Eq. (4))

A.5 NOISE SCALES

Symbol	Meaning
$E := \widehat{G} - G$	stochastic gradient noise matrix
$\nu_1(W) := \mathbb{E}[\|\widehat{G} - G\|_1 \mid W]$	ℓ_1 (sign-geometry) noise scale
$\nu_*(W) := \mathbb{E}[\|E\|_{\mathrm{nuc}} \mid W]$	nuclear-norm (spectral-geometry) noise scale
$c_1 := \sqrt{B} \cdot \nu_1$	batch-size-normalised ℓ_1 noise scale
$c_* := \sqrt{B} \cdot \nu_*$	batch-size-normalised nuclear noise scale
$\mathrm{nr}(G) := \|G\|_{\mathrm{nuc}}^2 / \|G\|_{\mathrm{F}}^2$	nuclear rank (effective rank in the $*$-vs-F sense)
$\phi_{\mathrm{ent}}(G) := \|G\|_1^2 / (mn\, \|G\|_{\mathrm{F}}^2)$	entrywise density of G

A.6 GLOBAL LOWER-BOUND RATIOS AND DIRECTIONAL SCORES

The paper distinguishes the ratios derived from global-smoothness lower bounds from the practical directional score (main text, Section 2; Definition 1):

Symbol	Meaning
$\Delta_g(W;\eta)$	actual expected decrease for geometry g at step size η
$B_g(W, B)$	lower-bound quantity for the stated conservative step
$\mathcal{R}_{\mathrm{LB}}(W, B)$	finite-batch geometry ratio $B_{\mathrm{spec}}/B_{\mathrm{sign}}$ (Eq. (5))
$\mathcal{R}_\infty(W)$	high-SNR form of the global-smoothness ratio (Eq. (9))
$\mathcal{R}_0(W) := \|G\|_{\mathrm{nuc}}^2 / \|G\|_1^2$	curvature-free proxy (Eq. (10))
$\widehat{\mathcal{R}}_{\mathrm{dir}}(W)$	empirical directional quadratic score (Eq. (11))

In the positive-signal regime, $\mathcal{R}_{\mathrm{LB}} = \mathcal{R}_\infty \cdot [\text{noise-correction factor}]$ and $\mathcal{R}_\infty = \mathcal{R}_0 \cdot \kappa$. The directional score $\widehat{\mathcal{R}}_{\mathrm{dir}}$ is not generally an estimator of either global-smoothness quantity.

A.7 ACTIVATION-DEPENDENT STABLE RANK

Symbol	Meaning
$\sigma(\cdot)$	pointwise activation function
X_ℓ	pre-activation at layer ℓ: $X_\ell = W_\ell A_{\ell-1}$
$A_\ell := \sigma(X_\ell)$	post-activation at layer ℓ
$\mathrm{st}(A) := \|A\|_{\mathrm{F}}^2 / \|A\|_{\mathrm{op}}^2$	stable rank of a matrix A
$\mathrm{er}(A)$	effective rank (alternative rank surrogate; see Appendix D)
$s_\sigma := \mathbb{E}[\sigma(\gamma)^2]/(\mathbb{E}[\sigma(\gamma)])^2$	activation noise-to-signal ratio, $\gamma \sim \mathcal{N}(0,1)$
NSR	noise-to-signal ratio (synonym for s_σ when context is clear)

Proposition 7 bounds $\mathrm{st}(\sigma(WX))$ by an activation moment ratio, subject to the stated random-feature assumptions. This controls the spectral–Frobenius feature-majorizer ratio only; it supplies no bound on the sign–spectral global ratio $\kappa = L_\infty / L_{\mathrm{op}}$.

A.8 OPTIMIZERS AND METHODS

The paper compares three primal geometries for matrix-valued parameters, two named single-geometry baselines, and the mixed-optimizer recipes built on top of them. The table below collects the abbreviations used in the main text and figures.

Name	Meaning
(S)GD	(stochastic) gradient descent; Frobenius-norm steepest descent
SignSGD, SIGNSGD	sign-based steepest descent under entrywise ℓ_∞ geometry
SpecSGD, SPECSGD	spectral steepest descent under operator-norm geometry; direction $\mathrm{polar}(\widehat{G})$
Muon	practical spectral-geometry optimizer used as our SpecSGD instantiation
AdamW	Adam with decoupled weight decay (standard baseline)
Scion	fixed mixed recipe: SignSGD on embed / head, Muon on hidden layers
G-Scion	adaptive gate: after forced warmup, an eligible layer may switch once after margin, EMA, and confirmation checks driven by $\widehat{\mathcal{R}}_{\mathrm{dir}}$
G-Muon	variant of G-Scion with AdamW (not SignSGD) as the "sign side"

A.9 STANDARD MATHEMATICAL NOTATION

We follow standard probability and analysis conventions; symbols below are listed for completeness.

Symbol	Meaning
$\mathbb{R}, \mathbb{N}, \mathbb{Z}$	real numbers, natural numbers, integers
$\mathbb{E}[\cdot]$	expectation (over minibatch sampling unless specified otherwise)
$\mathrm{Var}[\cdot], \mathrm{Cov}[\cdot,\cdot]$	variance, covariance
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$\mathbb{1}[\cdot]$	indicator function
$(x)_+ := \max(x, 0)$	non-negative part
$O(\cdot), \Theta(\cdot), o(\cdot)$	standard asymptotic notation
$\log(\cdot)$	natural logarithm unless otherwise stated

B Convergence Theory

B.1 Smoothness Relations

We establish the relationships between the least admissible norm-specific global smoothness constants used throughout the paper. Recall that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is $L_{\|\cdot\|}$-smooth w.r.t. a norm $\|\cdot\|$ if $\|\nabla f(X) - \nabla f(Y)\|_* \le L_{\|\cdot\|} \|X - Y\|$ for all X, Y.

Proposition 2 (Norm-specific smoothness relations). *Let $p = \min(m, n)$ and $d = mn$. Then*

$$L_{\mathrm{F}} \le L_{\mathrm{op}} \le pL_{\mathrm{F}}, \tag{12}$$

$$L_{\mathrm{op}} \le L_\infty \le dL_{\mathrm{F}}. \tag{13}$$

Consequently,

$$1 \le \kappa := \frac{L_\infty}{L_{\mathrm{op}}} \le mn. \tag{14}$$

Proof. Use the norm relations

$$\|A\|_{\mathrm{F}} \le \|A\|_* \le \sqrt{p}\, \|A\|_{\mathrm{F}}\,, \qquad \|A\|_* \le \|A\|_1\,,$$

and

$$\|M\|_\infty \le \|M\|_{\mathrm{op}} \le \|M\|_{\mathrm{F}} \le \sqrt{mn}\, \|M\|_\infty\,.$$

For example, $L_{\mathrm{op}} \le L_\infty$ follows from

$$\|\Delta\nabla f\|_* \le \|\Delta\nabla f\|_1 \le L_\infty \|\Delta W\|_\infty \le L_\infty \|\Delta W\|_{\mathrm{op}}\,.$$

Applying the remaining norm comparisons on the input and output sides gives $L_{\mathrm{F}} \le L_{\mathrm{op}} \le pL_{\mathrm{F}}$ and $L_\infty \le mnL_{\mathrm{F}}$. □

The upper bound is dimensionally sharp: for $f(W) = \frac{1}{2}\langle J, W\rangle^2$ with J the all-ones matrix, $L_\infty/L_{\mathrm{op}} = mn$. Global smoothness depends on the Hessian of the objective, not on the rank of the current gradient.

B.2 Proof of the Polar Misalignment Lemma

The following lemma bounds the alignment loss when computing the polar factor from a noisy gradient.

Lemma 3 (Polar misalignment). *Let $G = U\Sigma V^\top$ be the SVD of the true gradient and $\widehat{G} = G + E$ the noisy observation. Then:*

$$\left\langle G,\, \mathrm{polar}(\widehat{G})\right\rangle \ge \|G\|_{\mathrm{nuc}} - 2\|E\|_{\mathrm{nuc}}. \tag{15}$$

Proof. Write $\mathrm{polar}(\widehat{G}) = \widehat{U}\widehat{V}^\top$ where $\widehat{G} = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$ is the SVD of $\widehat{G}$. By the definition of the polar factor as the maximizer of the inner product over the set of partial isometries:

$$\left\langle \widehat{G},\, \mathrm{polar}(\widehat{G})\right\rangle = \|\widehat{G}\|_{\mathrm{nuc}} \ge \left\langle \widehat{G},\, UV^\top\right\rangle = \langle G,\, UV^\top\rangle + \langle E,\, UV^\top\rangle\,. \tag{16}$$

Thus:

$$\left\langle G,\, \mathrm{polar}(\widehat{G})\right\rangle = \left\langle \widehat{G},\, \mathrm{polar}(\widehat{G})\right\rangle - \left\langle E,\, \mathrm{polar}(\widehat{G})\right\rangle \tag{17}$$

$$= \|\widehat{G}\|_{\mathrm{nuc}} - \left\langle E,\, \widehat{U}\widehat{V}^\top\right\rangle. \tag{18}$$

Now, by the triangle inequality for nuclear norms:

$$\|\widehat{G}\|_{\mathrm{nuc}} = \|G + E\|_{\mathrm{nuc}} \ge \|G\|_{\mathrm{nuc}} - \|E\|_{\mathrm{nuc}}.$$

Also, $|\left\langle E,\, \widehat{U}\widehat{V}^\top\right\rangle| \le \|E\|_{\mathrm{nuc}} \cdot \|\widehat{U}\widehat{V}^\top\|_{\mathrm{op}} = \|E\|_{\mathrm{nuc}}$, since $\|\widehat{U}\widehat{V}^\top\|_{\mathrm{op}} = 1$ (it is a partial isometry). Combining:

$$\left\langle G,\, \mathrm{polar}(\widehat{G})\right\rangle \ge \|G\|_{\mathrm{nuc}} - \|E\|_{\mathrm{nuc}} - \|E\|_{\mathrm{nuc}} = \|G\|_{\mathrm{nuc}} - 2\|E\|_{\mathrm{nuc}}. \tag{19}$$

□

This bound is used in deriving the SpecSGD expected improvement (Eq. 4 in the main text). The factor of two arises because noise affects both the choice of direction (through the polar factor) and the inner product evaluation.

B.3 Ideal Arbitrary-Switching Norm-Steepest Descent

We record a convergence guarantee for an idealized deterministic method that may choose either the sign or spectral geometry at every iteration. The choice is arbitrary; the result does not require a gate or a one-step score.

Theorem 4 (Ideal arbitrary-switching descent and stationarity). *Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ be differentiable and bounded below by $f_{\inf} > -\infty$. Assume that f is L_∞-smooth with respect to the entrywise ℓ_∞ norm and L_{op}-smooth with respect to the spectral norm. At iteration t, choose arbitrarily*

$$(\|\cdot\|_t, \|\cdot\|_{t,*}, L_t) \in \{(\|\cdot\|_\infty, \|\cdot\|_1, L_\infty), (\|\cdot\|_{\mathrm{op}}, \|\cdot\|_{\mathrm{nuc}}, L_{\mathrm{op}})\}.$$

Let $G_t = \nabla f(W_t)$ and define

$$D_t \in \arg\max_{\|D\|_t \le 1} \langle G_t,\, D\rangle, \qquad \eta_t := \frac{\|G_t\|_{t,*}}{L_t}, \qquad W_{t+1} := W_t - \eta_t D_t.$$

Then, for every t,

$$f(W_{t+1}) \le f(W_t) - \frac{\|G_t\|_{t,*}^2}{2L_t}. \tag{20}$$

Consequently, with $L_{\max} := \max\{L_\infty, L_{\mathrm{op}}\}$, for every $T \ge 1$,

$$\sum_{t=0}^{T-1} \|\nabla f(W_t)\|_{\mathrm{F}}^2 \le 2L_{\max}\big(f(W_0) - f_{\inf}\big), \tag{21}$$

$$\min_{0\le t<T} \|\nabla f(W_t)\|_{\mathrm{F}}^2 \le \frac{2L_{\max}\big(f(W_0) - f_{\inf}\big)}{T}. \tag{22}$$

In particular, $\|\nabla f(W_t)\|_{\mathrm{F}} \to 0$ as $t \to \infty$.

Proof. Norm-specific smoothness gives

$$f(W_t - \eta D_t) \le f(W_t) - \eta\,\langle G_t,\, D_t\rangle + \frac{L_t\eta^2}{2}\,\|D_t\|_t^2.$$

By construction, $\|D_t\|_t \le 1$ and $\langle G_t,\, D_t\rangle = \|G_t\|_{t,*}$. Substituting $\eta = \eta_t$ proves Eq. (20). For both candidate geometries,

$$\|G_t\|_1 \ge \|G_t\|_{\mathrm{F}}, \qquad \|G_t\|_{\mathrm{nuc}} \ge \|G_t\|_{\mathrm{F}}.$$

Since $L_t \le L_{\max}$, summing Eq. (20) and using $f(W_T) \ge f_{\inf}$ proves Eqs. (21)–(22). Finiteness of the infinite sum of squared gradient norms gives the final claim. □

Remark 1 (Scope of the result). *Theorem 4 concerns an ideal deterministic single-block update with exact norm-steepest directions and geometry-specific step sizes $\eta_t = \|G_t\|_{t,*}/L_t$. It does not analyze the implemented G-Scion gate, which uses stochastic gradient and curvature estimates, finite decision cadence, one-way locking, shared learning-rate schedules, approximate polar directions, and simultaneous updates of coupled parameter blocks. The theorem therefore does not imply that the gate selects the better realised update, that G-Scion dominates either fixed optimizer, or that the per-layer decisions are jointly optimal.*

C $\mathcal{R}_0$ Approximation Bounds

We derive upper and lower bounds on the finite-batch geometry ratio $\mathcal{R}_{\mathrm{LB}}(W, B)$ (Eq. 5) in terms of the curvature-free proxy $\mathcal{R}_0$ (Eq. 10).

C.1 Exact Decomposition of $\mathcal{R}_{\mathrm{LB}}$

C.1.1 Exact decomposition

Define the relative noise ratios:

$$\alpha := \frac{2c_*}{\sqrt{B}\,\|G\|_{\mathrm{nuc}}}, \qquad \beta := \frac{2c_1}{\sqrt{B}\,\|G\|_1}, \tag{23}$$

and the curvature ratio $\kappa := L_\infty / L_{\mathrm{op}}$.

Proposition 5 (Exact decomposition)**.** *Assuming $0 \le \alpha, \beta < 1$ (signal dominates noise):*

$$\mathcal{R}_{\mathrm{LB}}(W, B) \;=\; \mathcal{R}_0(G) \;\cdot\; \underbrace{\left(\frac{1-\alpha}{1-\beta}\right)^2}_{\textit{noise correction}} \cdot \underbrace{\kappa}_{\textit{curvature ratio}} . \tag{24}$$

Proof. Factor the noise-corrected norms as $\|G\|_{\mathrm{nuc}} - 2c_*/\sqrt{B} = \|G\|_{\mathrm{nuc}}(1-\alpha)$ and $\|G\|_1 - 2c_1/\sqrt{B} = \|G\|_1(1-\beta)$. Substituting into Eq. (5):

$$\mathcal{R}_{\mathrm{LB}}(W, B) = \left(\frac{\|G\|_{\mathrm{nuc}}}{\|G\|_1}\right)^2 \left(\frac{1-\alpha}{1-\beta}\right)^2 \kappa.$$

From Eq. (10), $\|G\|_{\mathrm{nuc}}/\|G\|_1 = \sqrt{\mathcal{R}_0}$, giving (24). □

This decomposition reveals a three-factor structure: $\mathcal{R}_0$ captures gradient geometry, α and β encode finite-batch noise, and κ is the ratio of global norm-smoothness constants.

C.1.2 High-SNR approximation

When the signal dominates noise ($\alpha \ll 1$ and $\beta \ll 1$), the noise correction factor satisfies $\left(\frac{1-\alpha}{1-\beta}\right)^2 \approx 1$, so:

$$\mathcal{R}_{\mathrm{LB}}(W, B) \;\approx\; \mathcal{R}_0(G)\,\kappa. \tag{25}$$

If the curvature ratio is also approximately unity ($\kappa \approx 1$):

$$\mathcal{R}_{\mathrm{LB}}(W, B) \;\approx\; \mathcal{R}_0(G). \tag{26}$$

This is the regime in which $\mathcal{R}_0$ numerically approximates the global-smoothness geometry ratio. If κ is merely common rather than close to one, $\mathcal{R}_0$ preserves rankings but not magnitudes.

C.1.3 Upper and lower bounds

Since $0 \le \alpha, \beta < 1$, we have $1 - \beta \le 1$ and hence $1/(1-\beta)^2 \ge 1$, giving:

$$(1-\alpha)^2 \;\le\; \left(\frac{1-\alpha}{1-\beta}\right)^2 \;\le\; \frac{1}{(1-\beta)^2}.$$

Combining with the exact decomposition:

Corollary 6 (Upper and lower bounds)**.**

$$\kappa\,(1-\alpha)^2\,\mathcal{R}_0 \;\;\le\;\; \mathcal{R}_{\mathrm{LB}}(W, B) \;\;\le\;\; \frac{\kappa\,\mathcal{R}_0}{(1-\beta)^2}. \tag{27}$$

Reading the bounds. Relative to $\kappa\mathcal{R}_0$, the noise factor $((1-\alpha)/(1-\beta))^2$ is below one exactly when $\alpha > \beta$. A large spectral-side relative noise α shrinks the numerator, but does not alone imply $\mathcal{R}_{\mathrm{LB}} < \mathcal{R}_0$; that comparison also depends on β and κ. As the sign-side relative noise β approaches 1, the denominator can nearly vanish and the factor can diverge.

Curvature-free bounds. The norm equivalences from Proposition 2 yield:

$$1 \;\le\; \kappa = \frac{L_\infty}{L_{\mathrm{op}}} \;\le\; mn. \tag{28}$$

Substituting into (27) eliminates the curvature ratio entirely:

$$(1-\alpha)^2\,\mathcal{R}_0 \;\;\le\;\; \mathcal{R}_{\mathrm{LB}}(W,B) \;\;\le\;\; \frac{mn\,\mathcal{R}_0}{(1-\beta)^2}. \tag{29}$$

In the noiseless limit ($\alpha, \beta \to 0$):

$$\mathcal{R}_0 \;\le\; \mathcal{R}_{\mathrm{LB}}(W,B) \;\le\; mn\,\mathcal{R}_0. \tag{30}$$

The upper bound shows that $\mathcal{R}_0 > 1/(mn)$ is necessary for $\mathcal{R}_{\mathrm{LB}} > 1$ in the noiseless regime. Because $\mathcal{R}_0 \le 1$, norm equivalence alone supplies no nontrivial curvature-free sufficient condition that is uniformly strict. These inequalities compare the lower-bound quantities, not realised optimizer decreases.

C.1.4 First-order noise correction

For $\alpha, \beta \ll 1$, a Taylor expansion gives:

$$\frac{1-\alpha}{1-\beta} = (1-\alpha)(1+\beta+O(\beta^2)) = 1-(\alpha-\beta)+O(\alpha\beta+\beta^2),$$

so:

$$\mathcal{R}_{\mathrm{LB}}(W,B) \;\approx\; \kappa\,\mathcal{R}_0\,\big(1-2(\alpha-\beta)\big). \tag{31}$$

Reading the first-order correction. When $\alpha > \beta$, the spectral-side relative noise exceeds the sign-side and $\mathcal{R}_{\mathrm{LB}}$ falls below the noiseless proxy $\kappa\mathcal{R}_0$; when $\beta > \alpha$ the situation reverses. No universal monotonic dependence on batch size follows without an assumption on the ordering and batch dependence of these two relative noise terms.

C.1.5 When is $\mathcal{R}_0$ alone sufficient?

The proxy $\mathcal{R}_0$ numerically approximates $\mathcal{R}_{\mathrm{LB}}$ when the following conditions hold:

1. **High SNR in both geometries:** $\sqrt{B}\,\|G\|_{\mathrm{nuc}} \gg 2c_*$ and $\sqrt{B}\,\|G\|_1 \gg 2c_1$ (i.e., $\alpha \ll 1$ and $\beta \ll 1$).
2. **Approximately isotropic curvature:** $L_\infty \approx L_{\mathrm{op}}$ (i.e., $\kappa \approx 1$).

Under these conditions, $\mathcal{R}_{\mathrm{LB}}(W,B) \approx \mathcal{R}_0(G)$ as a lower-bound ratio.

Conversely, $\mathcal{R}_0$ alone can be misleading when:

- the batch size is small (large α or β), so that noise corrections dominate the signal;
- the gradient signal is weak (small $\|G\|_{\mathrm{nuc}}$ or $\|G\|_1$);
- the curvature asymmetry is large ($\kappa \gg 1$).

Three-factor decomposition. Eq. (24) decomposes $\mathcal{R}_{\mathrm{LB}}$ into three factors:

$$\mathcal{R}_{\mathrm{LB}}(W,B) = \underbrace{\mathcal{R}_0(G)}_{\text{gradient structure}} \cdot \underbrace{\left(\frac{1-\alpha}{1-\beta}\right)^2}_{\text{finite-batch noise}} \cdot \underbrace{\kappa}_{\text{global smoothness ratio}}.$$

Within the global lower-bound model, discrepancy between $\mathcal{R}_0$ and $\mathcal{R}_{\mathrm{LB}}$ is attributable to the relative noise terms, the global curvature ratio, or both. Discrepancy with realised optimizer outcomes can additionally arise from bound slack and mechanisms outside the model.

C.2 CONTROLLED SMALL-SCALE VALIDATION: INPUT-RANK SWEEP

We use a controlled toy NanoGPT ($n_{\text{layer}} = 4$, $n_{\text{embd}} = 128$, vocab 256, GELU activation) to test whether retaining measured directional curvature improves on the curvature-free proxy. The input-rank sweep changes both gradient structure and directional curvature, so this experiment is a controlled curvature sweep but not a proof that κ varies in isolation.

Directional-curvature sweep. We attach a forward pre-hook to a single layer (`block0.mlp.c_fc`) that left-multiplies its input by a fixed orthogonal projection $P_r = Q_r Q_r^\top$ of rank r, where $Q_r \in \mathbb{R}^{n_{\text{embd}} \times r}$ has orthonormal columns drawn at initialisation. The hook leaves the layer's weight tensor and parameterization unchanged while constraining the effective input subspace to dimension r. Sweeping $r \in \{2, 4, 8, 16, 32, 64, 128\}$ (with $r = 128$ the unconstrained control) moves the measured directional-curvature ratio $\hat{\kappa}_{\text{dir}}$ over a $\sim 5\times$ range without modifying the optimizer, the loss, or the activation function.

Per-trial measurement. At each measurement point (r, t, seed) and on a fresh probe batch B we compute, all at the same W: (i) $\mathcal{R}_0 = (\|G\|_{\text{nuc}}/\|G\|_1)^2$ from the layer-block gradient G; (ii) two unit-norm descent directions $D_{\text{spec}} = -\text{polar}(G)$ ($\|D_{\text{spec}}\|_{\text{op}} = 1$) and $D_{\text{sign}} = -\text{sign}(G)$ ($\|D_{\text{sign}}\|_\infty = 1$); (iii) directional curvatures $C_{\text{spec}} = \langle D_{\text{spec}}, H D_{\text{spec}} \rangle$ and $C_{\text{sign}} = \langle D_{\text{sign}}, H D_{\text{sign}} \rangle$ via single Hessian-vector products through the layer block; (iv) the directional curvature ratio $\hat{\kappa}_{\text{dir}} := C_{\text{sign}}/C_{\text{spec}}$ and deterministic quadratic score $R_{\text{quad}} := \mathcal{R}_0 \cdot \hat{\kappa}_{\text{dir}}$; (v) the directional quadratic-model step sizes $\eta_{\text{spec}} = \|G\|_{\text{nuc}}/C_{\text{spec}}$ and $\eta_{\text{sign}} = \|G\|_1/C_{\text{sign}}$ when $C_{\text{spec}}, C_{\text{sign}} > 0$; (vi) the empirical one-step decreases $\Delta_{\text{spec}} = f(W; B) - f(W + \eta_{\text{spec}} D_{\text{spec}}; B)$ and Δ_{sign} by the analogous step on the same batch with all other parameters held fixed. Here R_{quad} is the deterministic same-batch analogue of Eq. (11), whereas $\widehat{\mathcal{R}}_{\text{dir}}$ uses stochastic averaged probes during training. When $\Delta_{\text{spec}}, \Delta_{\text{sign}} > 0$, we define $R_{\text{emp}} := \Delta_{\text{spec}}/\Delta_{\text{sign}}$ and report log-ratios. All winner statements below are restricted to these valid, same-batch, geometry-specific probe steps; outside this regime the ratio is left undefined.

Empirical directional-curvature sweep. Table 3 reports geometric means over 24 trials per cell (3 seeds $\times$ 8 probe batches) at training step 200. The curvature-free proxy $\mathcal{R}_0$ predicts a sign-side win at every r, while the spectral probe step has the larger positive same-batch decrease at every reported point. The deterministic score $R_{\text{quad}} = \mathcal{R}_0 \hat{\kappa}_{\text{dir}}$ tracks the empirical one-step ratio much more closely; the same qualitative pattern holds at the other measurement steps and across all three seeds (Fig. 4). This is empirical calibration in this controlled setting, not an error bound that transfers to the LM/VLM gates.

Table 3: Toy directional-curvature sweep at `block0.mlp.c_fc`, step 200, GELU, 24 trials per row (geometric means). $R_{\text{quad}}/R_{\text{emp}} \in [1.01, 1.31]$ across the aggregate rows; $\mathcal{R}_0$ disagrees in sign at every r.

r	$\hat{\kappa}_{\text{dir}}$	$\mathcal{R}_0$	R_{quad}	R_{emp}
2	20,100	$5.7{\cdot}10^{-5}$	1.14	1.13
4	14,100	$9.6{\cdot}10^{-5}$	1.36	1.33
8	13,700	$1.5{\cdot}10^{-4}$	2.04	2.03
16	8,100	$2.9{\cdot}10^{-4}$	2.31	2.24
32	7,000	$4.7{\cdot}10^{-4}$	3.27	3.09
64	5,200	$7.9{\cdot}10^{-4}$	4.09	3.54
128	4,200	$1.2{\cdot}10^{-3}$	4.92	3.77

Across 504 controlled trials, retaining the measured directional-curvature factor yields the reported qualitative agreement, whereas the curvature-free proxy selects the wrong side at every reported point. The range $[1.01, 1.31]$ refers to the aggregate rows in Table 3, not a universal or per-trial error guarantee.

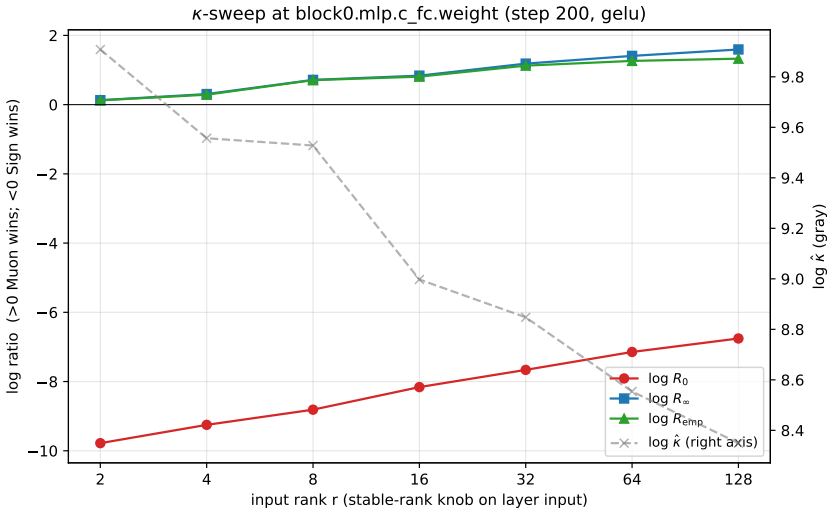

Figure 4: **Toy directional-curvature sweep at `block0.mlp.c_fc` (step 200, GELU, 24 trials per point).** As input rank r varies $2 \to 128$, the directional curvature ratio $\hat{\kappa}_{\mathrm{dir}}$ (gray, right axis) moves over $\sim 5\times$. The deterministic score R_{quad} (blue; labelled $\mathcal{R}_\infty$ in the plotted artifact) and empirical one-step ratio R_{emp} (green) move together. The curvature-free proxy $\mathcal{R}_0$ (red) predicts the wrong side throughout. The rank projection also changes $\mathcal{R}_0$, so the figure demonstrates the empirical value of retaining directional curvature rather than varying curvature alone.

D Activation-Dependent Stable Rank Theory

This section states precisely what the stable-rank framework of Davis & Drusvyatskiy (2025) does and does not imply for our sign–spectral comparison. Their result directly concerns feature-based spectral–Frobenius quadratic majorization; we use it as architectural motivation, not as a theorem determining $L_\infty / L_{\mathrm{op}}$.

Table 4: Unit-scale activation moment ratios. For non-homogeneous SiLU and GELU these values are descriptive and are not the scale-uniform constants required by the full theorem.

Activation $\sigma(t)$	Unit-scale s_σ	Scope
ReLU $\max(0,t)$	$\pi \approx 3.14$	homogeneous corollary
SiLU $t(1+e^{-t})^{-1}$	≈ 8.33	descriptive; non-homogeneous
GELU $t\,\Phi(t)$	≈ 5.34	descriptive; non-homogeneous
ReLU^2 $(\max(0,t))^2$	6	homogeneous corollary

D.1 Stable Rank in a Shared Spectral–Frobenius Majorizer

Consider a layered model where layer ℓ computes $X_\ell = W_\ell A_{\ell-1}$ with post-activation matrix $A_{\ell-1} = \sigma(X_{\ell-1})$. The shared quadratic upper model of Davis & Drusvyatskiy (2025, Sec. 7) uses two admissible feature-based coefficients:

$$\widetilde{L}_\ell^{\mathrm{F}} := C_F(W)\,\|A_{\ell-1}\|_{\mathrm{op}}^2\,, \qquad \widetilde{L}_\ell^{\mathrm{op}} := C_F(W)\,\|A_{\ell-1}\|_{\mathrm{F}}^2\,, \tag{32}$$

where $C_F(W)$ is a feature-based curvature coefficient. Their ratio is

$$\frac{\widetilde{L}_\ell^{\mathrm{op}}}{\widetilde{L}_\ell^{\mathrm{F}}} = \frac{\|A_{\ell-1}\|_{\mathrm{F}}^2}{\|A_{\ell-1}\|_{\mathrm{op}}^2} = \mathrm{st}(A_{\ell-1}). \tag{33}$$

These are feature-based majorization coefficients, not generally the least global norm-smoothness constants of the full network; the full bound also contains a parameter-only correction and a Taylor remainder.

D&D's nuclear-rank/stable-rank condition therefore directly motivates a spectral–Frobenius comparison. It does not determine L_∞, and hence does not by itself determine our sign–spectral ratio $\kappa = L_\infty / L_{\mathrm{op}}$. In particular, the coefficient identity above cannot be substituted into a ratio of the least global norm-smoothness constants without an additional theorem.

D.2 Activation Noise-to-Signal Ratio

Davis & Drusvyatskiy (2025, Cor. 2.7) give a homogeneous specialization assuming a fixed feature matrix X with nonzero columns, iid Gaussian W with the stated scaling, $\sigma(0) = 0$, homogeneity, polynomial growth, and a nonzero Gaussian mean. Under its finite-width probability conditions, the stable rank of $A = \sigma(WX)$ is bounded by the corresponding $(1 \pm \varepsilon)$ factor times

$$s_\sigma := \frac{\mathbb{E}[\sigma(\gamma)^2]}{(\mathbb{E}[\sigma(\gamma)])^2}. \tag{34}$$

The full Theorem 2.6 handles non-homogeneous activations using a supremum of this moment ratio over an input-scale interval, together with scale-wise nonvanishing-mean and growth assumptions. Unit-scale moments for SiLU and GELU are therefore not substitutes for that scale-uniform quantity.

D.3 Explicit s_σ Computations for VLM Activations

ReLU $\sigma(t) = \max(0, t)$**.** $(1,1)$-homogeneous. $\mathbb{E}[\sigma(\gamma)] = 1/\sqrt{2\pi}$, $\mathbb{E}[\sigma(\gamma)^2] = 1/2$.

$$s_\sigma = \frac{1/2}{1/(2\pi)} = \pi \approx 3.14.$$

Squared ReLU $\sigma(t) = (\max(0,t))^2$**.** $(2,2)$-homogeneous. $\mathbb{E}[\sigma(\gamma)] = \mathbb{E}[\gamma^2 \cdot \mathbb{1}[\gamma > 0]] = 1/2$, $\mathbb{E}[\sigma(\gamma)^2] = \mathbb{E}[\gamma^4 \cdot \mathbb{1}[\gamma > 0]] = 3/2$ (half of the fourth Gaussian moment).

$$s_\sigma = \frac{3/2}{(1/2)^2} = 6.$$

This is approximately twice the ReLU unit-scale moment ratio; it does not by itself predict the sign–spectral curvature ratio.

SiLU $\sigma(t) = t \cdot (1 + e^{-t})^{-1}$**.** Non-homogeneous, so the ratio depends on input scale. Numerical computation gives $s_\sigma \approx 8.33$ at unit scale. The Hermite relation is $s^{-1}\widehat{\sigma}_1(s) = \mathbb{E}[\sigma'(s\gamma)] = 1/2$, not $\widehat{\sigma}_1(s) = 1/2$ for all s.

GELU $\sigma(t) = t \cdot \Phi(t)$**.** Non-homogeneous. Numerical computation gives $s_\sigma \approx 5.34$ at unit scale. As with SiLU, this is descriptive and does not verify the scale-uniform assumptions of the full theorem.

D.4 The Pretrained-Weight Regime

The cited random-feature bounds do not apply directly to pretrained weights or to arbitrary changes of activation. Consequently, observed failures or recoveries after changing an activation cannot be attributed to stable rank from this theorem alone; weight structure, gradient structure, and local curvature all change together.

D.5 Empirical Stable Rank Across Activations

Across our LM MLP measurements, the post-activation stable rank remains numerically small for SiLU, ReLU, and ReLU^2 under both AdamW and Muon (typical values 2–4). These measurements are descriptive: they do not identify the global sign–spectral curvature ratio or isolate a causal activation effect.

E EXPERIMENTAL DETAILS

This section collects reproducibility details for GPT-2/OpenWebText pre-training, the controlled input-rank sweep, and ViT-B/16 trained from scratch on CIFAR-100. It also records the protocol correction required before a weight-tied VLM comparison can enter the empirical claims.

E.1 COMMON DIAGNOSTIC IMPLEMENTATION

The gate computes the directional score in Eq. (11), not the ratio $B_{\mathrm{spec}}/B_{\mathrm{sign}}$ based on global smoothness constants. A signal estimate $\widehat{G}_K$ is formed from K minibatch gradients. Fresh gradient probes produce natural-norm directions $D_{\mathrm{sign}} = \mathrm{sign}(g_b)$, with $\|D_{\mathrm{sign}}\|_\infty = 1$, and $D_{\mathrm{spec}} = \mathrm{polar}(g_b)$, with $\|D_{\mathrm{spec}}\|_{\mathrm{op}} = 1$. Separate probe batches provide alignments with $\widehat{G}_K$ and autodiff Hessian–vector products. The implementation averages alignments and curvatures separately before forming $\widehat{s}_g^2/\widehat{C}_g$; finite-sample ratios are therefore Monte Carlo scores rather than unbiased estimators. An HVP is exact only for the selected finite batch and numerical graph, not for the population Hessian. Nonpositive or nonfinite curvature/score values trigger implementation safeguards rather than a theorem-backed decision. Frobenius rescaling of a direction leaves $s(D)^2/C(D)$ invariant only when the same scaling is applied to both terms; raw directional-curvature ratios are reported in the natural sign/spectral normalization.

E.2 CODE AVAILABILITY

The anonymous URL https://anonymous.4open.science/r/g-scion-4EB6/README.md contained a README placeholder at submission. It did not contain the full training, evaluation, and diagnostic implementation needed for code-level reproduction; we acknowledge this artifact limitation explicitly.

E.3 LANGUAGE MODEL PRE-TRAINING

E.3.1 G-SCION CONFIGURATION DETAILS

Generalized-Scion (G-Scion) extends the Scion baseline by making the optimizer assignment of the embedding and output head layers *adaptive* during training. The optimizer assignment is denoted as (Embedding, Hidden, Output):

- **G-Scion (embedding):** Embedding is forced spectral through warmup and is then eligible for one spectral-to-sign decision. Hidden remains Muon and output remains SignSGD.
- **G-Scion (output):** Embedding remains SignSGD, hidden remains Muon, and output is forced spectral through warmup before becoming eligible for one spectral-to-sign decision.
- **G-Scion (embedding, output):** Both edge groups are forced spectral through warmup and independently eligible for one spectral-to-sign decision; hidden remains Muon.

Evaluation protocol. $\widehat{\mathcal{R}}_{\mathrm{dir}}$ is evaluated every $N{=}50$ steps after warmup. Each valid log score is smoothed with EMA factor 0.9; a layer switches only after two consecutive observations beyond the 0.1 margin around the nominal zero threshold. Fresh signal, direction, and HVP probes are aggregated as described above. Once switched, a layer remains locked.

Baselines.

- **Scion:** SignSGD (embedding), Muon (hidden), SignSGD (output).
- **Muon:** AdamW (embedding), Muon (hidden), AdamW (output).

Cross-layer interaction effects. Although $\widehat{\mathcal{R}}_{\mathrm{dir}}$ is informative in the reported per-layer comparisons, it does not recover the oracle joint optimizer assignment in approximately 50% of cases (see Section 4). The oracle sometimes prefers a mixed assignment (SpecSGD on the output, SignSGD

on the embedding) even though SpecSGD wins on each layer in isolation. This occurs because the spectral step on the embedding can be relatively large, and the joint update suffers from a negative cross-layer interaction (synergy penalty). This observation shows that the layerwise score omits cross-block curvature and is not a joint-assignment guarantee; we make no claim about how the error scales in larger multi-component architectures.

E.3.2 Batch-Size Sweep on Tiny Shakespeare

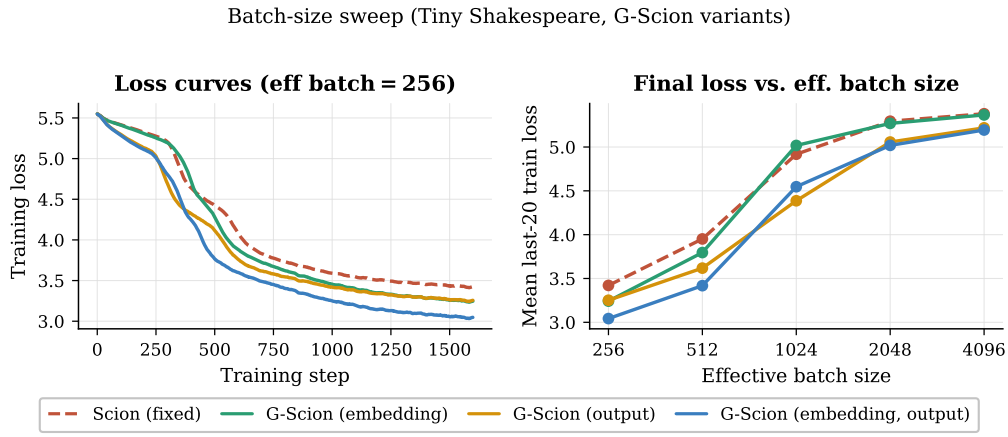

Figure 5: **Batch-size sweep (Tiny Shakespeare, G-Scion variants).** *Left:* Training loss curves at effective batch 256 (1,600 steps). G-Scion (embedding, output) leads throughout; G-Scion (embedding) tracks Scion closely, while G-Scion (output) and G-Scion (embedding, output) pull away via output-layer adaptation. *Right:* Mean last-20 training loss vs. effective batch size ($\log_2$ scale). All G-Scion variants beat Scion at every reported batch size; the observed margin is largest at small batch and shrinks as the effective batch grows.

The batch-size sweeps observe that G-Scion's advantage is largest at small B. This pattern is compatible with a smaller spectral relative-noise penalty, but the theory does not predict it universally without an ordering between α and β. The high-SNR global-smoothness ratio contains no batch-size dependence. These results are therefore empirical observations rather than confirmation of a universal small-batch theorem.

E.3.3 OWT Activation Ablation with $\widehat{\mathcal{R}}_{\text{dir}}$ (12 Runs)

Twelve NanoGPT (31M, 6 layers) experiments on OpenWebText use the directional Monte Carlo score $\widehat{\mathcal{R}}_{\text{dir}}$ (Eq. 11) for gating decisions, organised in a 3×4 grid of MLP activations (GELU, ReLU^2, SiLU) $\times$ optimizer variants (Scion, G-Scion (embedding), G-Scion (output), G-Scion (embedding, output)). All runs: 2,000 steps, effective batch 128 (physical 8×16 microbatches), peak learning rate 10^{-3} (cosine schedule, min 10^{-4}, 100-step warmup).

Validation losses. Table 5 reports final validation losses for all 12 runs. The ranking G-Scion (embedding, output) $>$ G-Scion (embedding) $>$ G-Scion (output) $>$ Scion holds for all three activations. The best-to-worst spread within any method column is < 0.05 in validation loss, while the cross-method spread reaches 0.37 (GELU column); activation choice is therefore a second-order effect at this scale.

Diagnostic readings and gate decisions. The raw diagnostic at a valid measurement j is $s_{\ell,j} = \log \widehat{\mathcal{R}}_{\text{dir}\ell,j}$. The gate instead acts on its exponential moving average, initialized by $z_{\ell,1} = s_{\ell,1}$ and subsequently updated as $z_{\ell,j} = \beta z_{\ell,j-1} + (1-\beta)s_{\ell,j}$. For the stated spectral-to-sign rule, the decision requires two consecutive smoothed readings below $\tau - m$, followed by locking (Algorithm 1). A fraction of raw readings below zero is therefore not a switch frequency or a confirmation rate. Shadow measurements on a fixed route likewise do not constitute routing decisions.

Table 5: Final validation loss, OWT 31M, $\widehat{\mathcal{R}}_{\text{dir}}$. **Top:** single-seed runs (12 runs, seed 1337). **Bottom:** two-seed replication (24 runs, seeds 42 and 2024; mean ± population std). The ranking G-Scion (embedding, output) > G-Scion (embedding) > G-Scion (output) > Scion is uniform across all activations and both seed sets.

Activation	Scion	G-Scion (embed.)	G-Scion (out.)	G-Scion (embed., out.)	Best–Worst
Single seed (seed 1337)					
GELU	5.526	5.278	5.459	**5.161**	0.365
ReLU^2	5.490	5.293	5.421	**5.196**	0.294
SiLU	5.531	5.309	5.462	**5.205**	0.326
Mean	5.516	5.293	5.447	**5.187**	
Δ vs. Scion	—	+0.22	+0.07	**+0.33**	
Two-seed replication (seeds 42 & 2024; mean ± std, $n = 2$)					
GELU	5.531 ± 0.011	5.315 ± 0.011	5.448 ± 0.012	**5.181 ± 0.000**	0.350
ReLU^2	5.507 ± 0.001	5.351 ± 0.001	5.431 ± 0.001	**5.232 ± 0.000**	0.275
SiLU	5.559 ± 0.015	5.324 ± 0.020	5.454 ± 0.008	**5.205 ± 0.001**	0.354
Mean	5.532	5.330	5.444	**5.206**	
Δ vs. Scion	—	+0.20	+0.09	**+0.33**	

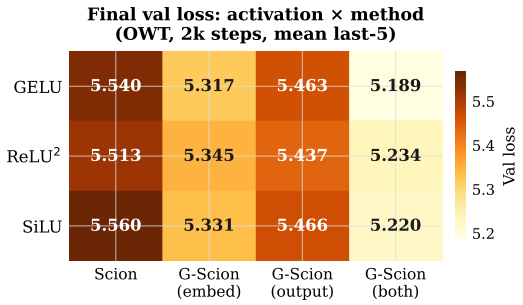

Figure 6: **Activation × method validation loss** (OWT, 31M, 2 000 steps, $\widehat{\mathcal{R}}_{\text{dir}}$; mean of last-5 checkpoints). Green = lower (better). The ranking G-Scion (embedding, output) > G-Scion (embedding) > G-Scion (output) > Scion holds uniformly across all three activations. Activation function is a second-order effect (within-column spread < 0.05 in validation loss).

Figure 3 is retained as a descriptive diagnostic trace. We omit the historical pooled-score table and the exact first-decision range because their raw-versus-smoothed definitions, sampled steps and aggregation weights cannot be reconstructed from the available summaries. In particular, a summary over the training horizon must not be labelled as exclusively pre-switch. Reproducible numerical summaries must identify the runs and activations, raw score or EMA, included steps and phase boundaries, valid-observation count, and weighting across runs. The reported route change at step 150 is stated separately below; no pooled sign-side percentage is used to certify it.

Table 6: AdamW-side reference at 31M (same setting as Table 2). *Muon*: AdamW on embed/head + Muon on hidden, the standard practical recipe. *G-Muon*: one-way AdamW→spectral gate on embed/head + Muon on hidden. AdamW adds momentum and per-parameter adaptivity outside the one-step bound, so this is a reference rather than a like-for-like sign-vs-spectral comparison; the gate does not improve over Muon in this regime, consistent with the theory not covering AdamW.

Method	Final val loss	Δ vs. Muon
Muon (AdamW embed/head + Muon hidden)	4.755 ± 0.035	—
G-Muon (AdamW→Spectral embed/head + Muon hidden)	4.851 ± 0.016	−0.096

Output gate timing. For G-Scion (output) and G-Scion (embedding, output), the output gate fires at step 150 universally across all 6 runs (two methods $\times$ three activations). The stated schedule begins decisions at $t = 100$, with cadence $N = 50$; a switch at $t = 150$ is the second scheduled opportunity. This timing is distinct from the numerical value of either the raw score or its EMA. The exact first-decision score range is not used as quantitative evidence until the original logging and aggregation records are reconciled (see the diagnostic-reporting paragraph above). After locking to sign at step 150, the output stays in sign for the remaining $\approx$1,850 steps.

Comparison to $\mathcal{R}_0$-proxy runs. Replacing the directional score with the curvature-free proxy $\mathcal{R}_0$ on the same 3×4 grid keeps the output proxy on the spectral side, whereas the HVP-based directional score is on the sign side and routes the output to sign. The proxy-routed output variants have slightly higher final validation loss than their matched directional-score counterparts. Thus the two scoring rules produce different routes, but this run-level comparison does not isolate curvature alone because $\widehat{\mathcal{R}}_{\mathrm{dir}}$ also replaces gradient norms with stochastic directional alignments. The controlled same-iterate experiment in Appendix C.2 isolates the added directional-curvature factor within $R_{\mathrm{quad}} = \mathcal{R}_0 \hat{\kappa}_{\mathrm{dir}}$; neither comparison identifies the global ratio κ or equates $\widehat{\mathcal{R}}_{\mathrm{dir}}$ with $\mathcal{R}_\infty$.

Replication across seeds. The identical 3×4 grid was replicated with two additional seeds (42 and 2024), for 24 runs in total; per-cell means are in the lower block of Table 5. The ordering G-Scion (embedding, output) $<$ G-Scion (embedding) $<$ G-Scion (output) $<$ Scion holds for all three activations in both seed sets, and the cross-seed spread is well below the smallest between-method gap. The mean Δ vs. Scion across the two seeds matches the single-seed values within 0.02 per cell.

Learning-rate stratification at the output head. A sweep over four learning rates and five effective batch sizes (80 runs) shows that the phase-aggregated output score is learning-rate dependent (Table 7). At conservative learning rates both phase means are negative; at 3×10^{-3} the spectral-phase mean is negative and the subsequent sign-phase mean is positive. This is a phase-stratified aggregate, not a single-run bidirectional transition trace.

Table 7: Output $\log \widehat{\mathcal{R}}_{\mathrm{dir}}$ by learning rate and phase (OWT 31M, 80-run sweep, $\widehat{\mathcal{R}}_{\mathrm{dir}}$). "Spectral phase": steps 0–100 (before gate). "Sign phase": steps 150–1950 (after gate). All gated-output runs switch at step 150 regardless of LR.

LR	**Out** $\log \widehat{\mathcal{R}}_{\mathrm{dir}}$**, spectral phase**	**Out** $\log \widehat{\mathcal{R}}_{\mathrm{dir}}$**, sign phase**
10^{-4}	-1.83 (100% sign-side)	-1.76 (100% sign-side)
3×10^{-4}	-2.46 (100% sign-side)	-1.94 (100% sign-side)
10^{-3}	-1.12 (100% sign-side)	-0.29 (72% sign-side)
3×10^{-3}	-0.27 (100% sign-side)	$+\mathbf{0.70}$ (19% sign-side)

E.3.4 Scale-Sweep Provenance and Scope

The original scale analysis combined two different diagnostics. Its 1B plot displayed the curvature-free proxy $\mathcal{R}_0$, whereas the available HVP-based output score is negative and the realised route becomes sign after warmup. Several 163M/406M method–seed cells lack complete directional records. We therefore exclude those tables and figures from score validation. A matched scale test records, for every seed and method, the directional score, parameter route, and static sign/spectral edge controls.

E.3.5 Split Invariance Results

The spectral and nuclear norms are unitarily invariant: replacing W by $UWV^\top$ for unitary U, V leaves both norms, and hence the SpecSGD geometry, unchanged. The ℓ_∞-steepest descent underlying SignSGD is basis-dependent, so rotating the coordinate system can increase L_∞ and degrade convergence (Balles et al., 2020).

We confirmed this empirically in the GPT-2 setting: for each layer, the spectral-side quantity $\|G\|_{\mathrm{nuc}}^2$ computed in the original basis and after a random orthogonal transformation of the weight matrix

agrees to relative error $< 10^{-12}$, so the choice of coordinate system (e.g., attention-head ordering, neuron permutation) does not affect the spectral-geometry diagnostic, whereas SignSGD diagnostics are sensitive to such choices.

E.4 VISION: VIT-B/16 ON CIFAR-100

This appendix reports image-classification experiments that apply the four edge-assignment trajectories of Section 4 to ViT-B/16 (Dosovitskiy et al., 2020) trained from scratch on CIFAR-100 (Krizhevsky et al., 2009). The evidence concerns final assignment outcomes rather than validation of the directional score.

E.4.1 SETUP

Architecture and data. ViT-B/16 (12 encoder blocks, width 768, 86M parameters) with `num_classes`=100, image size 224, trained from random initialisation on CIFAR-100 (50k train, 10k test; standard random-crop+flip augmentation). Random initialization alone does not verify all assumptions of Proposition 7; the experiment is an empirical cross-modal check rather than an application of that theorem.

Optimizer layer partition. Hidden attention/MLP layers always use Muon. The edge groups are $\{\text{conv_proj} + \text{class_token} + \text{pos_embed},\ \text{heads.head}\}$, matching the LM "embedding" and "output" roles. The runs use a 100-step forced warmup and one-way locking. The HVP score is nonfinite at many ViT patch-embedding measurements, so neither the implementation flag nor the plotted proxy is used to infer a geometry preference.

Sweep grid. We sweep two learning rates $\{3 \times 10^{-4},\ 3 \times 10^{-3}\}$ and three effective batch sizes $\{128, 256, 512\}$, with two seeds (1337 and 3407) per cell, for 48 runs at 1,000 steps each. A 3-seed replicate at effective batch 256 and learning rate 3×10^{-3} (12 further runs) corroborates the corresponding sweep cell.

E.4.2 VALIDATION ACCURACY AND BATCH-SIZE SENSITIVITY

Table 8: **Final validation accuracy on ViT-B/16 CIFAR-100** (1,000 steps, 2 seeds per cell, lr=3×10^{-3}). The joint spectral-edge assignment trajectory exceeds fixed Scion at every reported effective batch size. Δ is its accuracy minus Scion accuracy in percentage points.

Effective batch	Scion	G-Scion (embedding)	G-Scion (output)	G-Scion (embedding, output)	Δ
128	30.9 ± 1.2	32.5 ± 2.1	43.8 ± 2.8	$\mathbf{42.8 \pm 0.4}$	+11.9
256	39.5 ± 0.7	41.0 ± 0.4	45.3 ± 0.4	$\mathbf{51.0 \pm 1.1}$	+11.5
512	40.0 ± 0.7	43.9 ± 2.7	44.1 ± 1.6	$\mathbf{47.9 \pm 1.2}$	+ 7.9

The joint edge-assignment trajectory exceeds fixed Scion at every reported effective batch size (Table 8). At batch 128, the output-only and joint cells overlap within seed spread; at batches 256 and 512, the joint cell is highest. The observed gap shrinks with batch size, but this sweep does not isolate the relative noise terms required for a theoretical direction-of-effect claim.

E.4.3 SCORE PROVENANCE

The historical ViT plot was labelled $\widehat{\mathcal{R}}_{\mathrm{dir}}$, but the patch-embedding directional score is valid at only three of eight recorded measurements and becomes nonfinite. The plot is therefore removed, and the ViT result is used only as evidence that the alternative assignment trajectories improve final accuracy. A matched static spectral-edge control is required to attribute this gain to online identification.

E.4.4 PER-LAYER DECOMPOSITION: SUPERADDITIVE INTERACTION

At effective batch 256, the joint gain over Scion is larger than the sum of the two single-group gains, whereas the 31M decomposition in Table 2(b) is close to additive. This superadditive outcome is

direct evidence that independent layerwise scores omit cross-block interaction; the experiment does not identify the mechanism or establish a generally preferable joint assignment.

E.5 Weight-Tied VLM Protocol

E.5.1 Architecture and estimand

nanoVLM (HuggingFace, 2025) combines a SigLIP2 vision encoder (Tschannen et al., 2025), an MLP projector, and a SmolLM2 decoder (Ben Allal et al., 2025). The decoder's token embedding and output head are aliases of the same $\mathrm{Parameter}$. The relevant estimand is therefore the geometry assigned to one shared embedding–head tensor, including gradient contributions from both roles, not separate "embedding" and "head" decisions.

E.5.2 Audit of the historical configurations

The historical runs set the full directional diagnostic off. During optimizer construction, the shared tensor was first encountered through its embedding alias and assigned once to the AdamW group. Run labels referring to a head route therefore do not identify the realised optimizer on that tensor. The old proxy values and downstream metrics cannot support a sign-versus-spectral head comparison, a VLM head reversal, or activation causality, and are not included as empirical evidence.

E.5.3 Required matched design

A matched evaluation must (i) enumerate all aliases and verify a single parameter identity, (ii) assign exactly one realised optimizer route to the shared tensor, (iii) compute its joint alignment and HVP, (iv) compare static sign and static spectral routes at matched hyperparameters, and (v) evaluate the adaptive route against those static corners. The projector should be crossed independently because it is the interface receiving visual features. Untying the embedding and head changes the architecture and is therefore a secondary control rather than a substitute for the tied comparison.

E.5.4 Sweep grid and seeds

The corrected tied comparison and its static controls are required before reporting VLM performance or a cross-architecture reversal. No result from that planned comparison is claimed in this version.

F Comprehensive Related Work

Spectral and matrix geometry in large-model training. Many widely used optimizers for deep networks can be viewed as preconditioned first-order methods, including coordinatewise adaptive methods such as AdaGrad and Adam and matrix-preconditioned methods such as K-FAC and Shampoo (Duchi et al., 2011; Kingma & Ba, 2015; Martens & Grosse, 2015; Gupta et al., 2018). A complementary perspective emphasises the geometry of the update for matrix-valued parameters (Carlson et al., 2015; Bernstein & Newhouse, 2024; Veprikov et al., 2025), recently exemplified by Muon (Jordan, 2024), which performs a spectral-norm steepest-descent-style update and has emerged as a competitive alternative to AdamW in language-model pretraining (Liu et al., 2025; Zeng et al., 2025; Shah et al., 2025); convergence bounds and a critical-batch-size analysis covering Muon with Nesterov momentum and weight decay are given by Sato et al. (2025). Muon has been shown to scale to billion-parameter LLMs and MoE training (Liu et al., 2025), and related matrix-preconditioned methods such as SOAP also improve over AdamW in large-batch pretraining regimes (Vyas et al., 2024; Qiu et al., 2025). The remaining open question is which geometry should be used at which layer of a given architecture, and under what conditions its advantage persists.

Mixed recipes and the assignment problem. The practical answer so far has been to mix geometries. Successful Muon-style recipes typically apply matrix-aware updates to hidden weight matrices while retaining AdamW-like updates on embeddings, normalization parameters, biases, or output layers (Liu et al., 2025; Eschenhagen et al., 2026). Hybrid methods make this selectivity even more explicit. SOAP can be read as Adam in a rotated basis induced by a matrix preconditioner (Vyas

et al., 2024), COSMOS splits optimization across eigensubspaces using SOAP on leading directions and Muon on the remainder (Liu et al., 2026), and adaptive Muon variants reintroduce variance adaptation rather than treating orthogonalization as sufficient (Frans et al., 2025; Si et al., 2025; Li et al., 2025). This line of work strongly suggests that the field is converging on mixed geometry rather than on a single universally best optimizer. Yet the assignment itself remains largely heuristic: the usual partition into "matrix" and "non-matrix" parameters is inherited from practice and often must be re-validated across scale, architecture, and tuning budget (Wen et al., 2026; Qiu et al., 2025; Semenov et al., 2025).

Why AdamW-like and sign-based geometry persists. Selectivity in mixed-optimizer recipes is consistent with transformer heterogeneity. Hessian-based analyses show that transformer blocks have markedly different local curvature, so a single Euclidean learning-rate geometry is a poor fit across the whole network (Zhang et al., 2024). Empirical work finds that Adam-like adaptivity is disproportionately important on a small set of sensitive parameter groups, especially the last layer and LayerNorm parameters, while other blocks tolerate simpler updates (Zhao et al., 2025). Sharpness and gradient diagnostics reach a similar conclusion: blockwise sharpness disparity persists throughout training (Wang et al., 2025), and gradient heterogeneity helps explain why coordinatewise or sign-like normalization remains effective where SGD fails (Tomihari & Sato, 2025). SignSGD (Bernstein et al., 2018) and the geometric interpretation of sign descent as entrywise steepest descent (Balles et al., 2020) provide the sign-side foundation, and recent theory argues that Adam benefits from favourable ℓ_∞-geometry of the loss landscape (Xie et al., 2025). Together these results indicate why AdamW-like behaviour remains difficult to displace entirely on the embedding and output side, even when spectral methods improve training on the internal weight matrices.

Criteria for spectral advantage and the role of stable rank. Among existing attempts to move beyond heuristic assignment, the closest precursor is the spectral-versus-GD analysis of Davis & Drusvyatskiy (2025). That work derives a layerwise condition under which spectral updates should outperform Euclidean gradient descent, governed by the gradient's nuclear-to-Frobenius ratio relative to the stable rank of the incoming activations. It is especially relevant because it ties optimizer preference to measurable properties of the current layer rather than to an architecture label alone, and because its empirical results support the view that internal transformer matrices are especially favorable to spectral geometry. The same work also links activation choice to optimizer suitability through stable rank, providing a first concrete bridge between architectural nonlinearities and geometry selection. A complementary Muon-specific account is given in Wang et al. (2026), which traces Muon's gains mainly to value-output attention and feedforward matrices and links that advantage to more isotropic singular spectra under heavy-tailed data. These results bring the literature close to a principled layerwise criterion, but they do not yet resolve the comparison that modern mixed recipes actually instantiate: spectral geometry versus AdamW-like coordinatewise or sign-based geometry.

Activation dependence and non-Euclidean noise. The architecture dependence of optimizer preference remains underdeveloped in the literature. Modern transformers rely on a range of activations, including GELU, Swish, and related variants (Hendrycks & Gimpel, 2016; Ramachandran et al., 2018; So et al., 2021), but optimizer choice is usually treated as largely separate from activation design. The stable-rank perspective of Davis & Drusvyatskiy (2025) begins to close that gap by tying activation-induced representation structure to the case for spectral updates. On the stochastic side, non-Euclidean gradient noise scales provide a way to compare geometries at finite batch size rather than only in deterministic limits (Naganuma et al., 2026). Together, these ideas suggest that geometry preference should vary not only across layers, but also across batch regimes, activations, and architectures.

Unified geometry views and positioning of the present work. A broader theoretical literature clarifies that these comparisons should be posed in geometric terms. Norm-based and trust-region views interpret sign-based and spectral updates through different metrics or constraint sets (Bernstein & Newhouse, 2024; Pethick et al., 2025; Kovalev, 2025). Relative to Davis & Drusvyatskiy (2025), the key shift here is from a spectral-versus-Euclidean quadratic-model comparison to the spectral-versus-sign assignment faced by modern mixed recipes. Relative to fixed prescriptions such as Scion (Pethick et al., 2025), our contribution is a measurable layerwise instrument: a signal–curvature

factorisation derived from global lower bounds together with an empirically evaluated directional score. The input-rank sweep tests the score at a common iterate, the 31M output supplies an operational sign-side case, and the 31M/ViT trajectories test alternative assignments against fixed Scion.

G Compute Resources

Hardware. GPT-2 and ViT/CIFAR-100 experiments ran on NVIDIA A100 80 GB GPUs, split between RunPod cloud instances (for the GPT-2 scale sweep, Appendix E.3.4, and the ViT-B/16 sweep, Appendix E.4) and an institutional cluster (for the OWT 31M activation ablation, Appendix E.3.3, and the small-scale directional-curvature diagnostic, Appendix C.2). Peak parallelism was four A100s, dominated by a 3-GPU RunPod pod for the largest scales. Historical nanoVLM pilots ran on a single NVIDIA H100 80 GB per run; they are excluded from the empirical claims for the protocol reasons in Appendix E.5.

Per-experiment wall-clock. The 1B GPT-2 runs are the dominant cost at ≈ 140 single-GPU hours per seed; the 163M and 406M runs scale roughly with parameter count. The OWT 31M activation ablation completes at ≈ 1 hour per run, the 12-run ViT-B/16 paper sweep at ≈ 6.4 hours per run, the 48-run ViT learning-rate sweep at ≈ 2.8 hours per run, and the small-scale directional-curvature probes within ten minutes per run. The historical 1h-to-2h nanoVLM comparison included exhaustive per-component diagnostics and does not isolate gate cost. In a separate end-to-end timing experiment with decision cadence $N{=}100$, G-Scion measured approximately 4% overhead over its matched baseline. This measurement is configuration-specific; cost depends on cadence, probe count, unlocked groups, model, and implementation.

Total compute budget. The experimental program used approximately $2{,}400$ A100-GPU-hours plus 100 H100-GPU-hours, excluding unlogged pilot runs. The A100 total includes about $1{,}700$ hours for the 1B scale runs even though those runs are not used as directional-score evidence; the H100 total includes the historical nanoVLM studies excluded from the empirical claims.

H HELD-OUT SCHEDULED-UPDATE CALIBRATION: DEVELOPMENT STUDY

Design and estimands. We trained four fresh 31M language-model sources using static SS and PP assignments and seeds 1337 and 2024, for 2,000 updates. Here S denotes sign and P spectral, with the embedding assignment first and output assignment second. Matrix updates were memoryless; this is not a momentum-Muon study. The diagnostic grid uses checkpoints after 100, 700 and 1,999 updates, four units (joint token/position embeddings, output head, attention and MLP), gradient batches $B \in \{32, 128, 512\}$ and two repetitions, giving 288 paired probes. We use $K = 12$, $M = 8$, Hessian batch $H = 128$, FP32 diagnostics and five Newton–Schulz iterations. Each sign/spectral candidate applies its actual scheduled displacement to the selected unit from the identical checkpoint; other units remain fixed. Model, optimizer and RNG restoration checks passed. Signal, direction, Hessian and update samples occupy separate training-token regions. Both candidates are evaluated on the same 32 disjoint held-out spans, aggregated into eight paired blocks.

We distinguish the implemented self-curvature score from a full-group diagnostic that also includes cross-member Hessian terms. Neither estimates $\mathcal{R}_{\mathrm{LB}}$. The alternative predictor $Q_{\mathrm{actual}} = -g^\top u - \frac{1}{2}u^\top H u$ uses the actual rounded scheduled displacement u and independently estimated signal and curvature. It is a separate development predictor, not a retrospectively validated replacement for the gate.

Table 9: Development scheduled-update ranking. Regret is selected held-out loss minus the better sign/spectral loss, averaged over the 288 tuples. The four undefined directional ratios retain the source-route fallback.

Selection rule	Correct / 288	Mean regret (nats/token)
Full-group directional diagnostic + fallback	219	0.005037
Always sign	24	0.024432
Always spectral	264	0.003471
Q_{actual} (development only)	288	0.000000

Failures below and near a score of one. With prespecified log-score bins below -0.2, between -0.2 and 0.2, and above 0.2, the full-group threshold makes 7/30, 129/169 and 83/85 correct choices, respectively. The policy itself switches its preference at score one. Four ratios are undefined because of negative sign alignment in late SS output-head probes at $B = 32$; the source-route fallback chooses sign and loses all four. No such cases are excluded from the primary table. Among the 284 defined pairs, self-curvature and full-group scores rank 205 and 219 correctly. All 28 changed choices occur in joint embeddings, where accuracy increases from 39/72 to 53/72. This diagnoses within-group cross curvature, not full embedding/head interactions.

Noise, route dependence and uncertainty. The full-group score makes 64/96, 81/96 and 74/96 correct choices at $B = 32, 128, 512$, while the recorded uncorrected variance-to-squared-mean-gradient proxy decreases from 1.838 to 0.730 to 0.231. Lower measured noise therefore does not monotonically repair agreement. Always spectral has lower pooled regret, but the directional diagnostic has lower regret on each SS source; all PP-source winners are spectral. The balanced diagnostic grid is not an estimate of deployment frequencies. Its 288 tuples are correlated measurements from two training seeds, not 288 independent replications. Of the paired probes, 235 have the same winner in every held-out block and 53 show block heterogeneity.

Limits of the evidence. There are 24 sign wins, 264 spectral wins and no measured ties at the descriptive 10^{-5}-nat band. This band has not been qualified against full-grid numerical repeatability, and actual LM loss near-ties remain untested. The historically used finite holdout lacks document identifiers; an untouched, document-independent final test is not established. The perfect development ranking of Q_{actual} requires frozen-predictor confirmation on fresh seeds and independent evaluation data. Separate quadratic software controls span both winners and ties but do not supply missing LM evidence. These results establish limitations of scheduled-update ranking in this setting; they do not establish an end-to-end optimizer ranking or invalidate the theorem's conservative-step guarantee.

Outstanding validation. The remaining experiments include independent calibration and numerical tie qualification; known-constant bound-certification checks; qualified Frobenius pairwise ordering; genuine stateful optimizer and equal-budget AdamW/Muon comparisons; joint embedding/head updates; and sensitivity, precision and cost measurements. The separately reported four-static-route comparison and warm-up/replay results are summarized in Appendix I; they should not be described as unrun. The nine-arm implementation prepared for this development campaign has passed software checks but has not supplied those training results. Full per-seed/configuration reconciliation, one-shot and prespecified schedule controls, and independent testing of a frozen warm-up rule remain necessary. Score, realised geometry, switch time and lock state should be reported together. Broader activation, scale and correctly tied VLM claims need matched confirmation; reversible switching remains an extension.

I STATIC ASSIGNMENTS, EARLY LOCKING AND STAGE-DEPENDENT PREFERENCES

Study and evidence scope. A follow-up trains a 31M-parameter GPT-2 on OpenWebText for 2,000 updates under all four fixed sign/spectral assignments to embedding and output layers and the adaptive gate, with five seeds per setting. Paired checkpoint diagnostics separately compare the two candidate updates by their held-out loss decrease on three seeds. Table 10 transcribes the rounded final-loss labels from the follow-up summary figure. These are a separate experimental series from Table 2 and Appendix H; their values must not be pooled across configurations.

Table 10: Reported 31M GPT-2/OWT follow-up. S denotes sign and P spectral; the embedding assignment is listed first. Lower validation loss is better. Values are rounded figure labels; per-seed values are in the run logs from which Figure 1 is drawn, and no significance tests are inferred.

Assignment or routing variant	Final validation loss
PP: spectral on both edge groups	4.92
Frozen-route replay	5.23
G-Scion with warm-up	5.23
SP: spectral on output only	5.35
PS: spectral on embedding only	5.40
G-Scion	5.45
G-Scion with reversal allowed	5.45
SS: sign on both edge groups (fixed Scion)	5.68

Local agreement and later mismatch. On the output head, every reported early recommendation for sign agrees with the larger held-out decrease across the three diagnostic seeds. This is evidence for those sampled decisions, not an error-rate estimate over all stages, layers or noise levels: the summary does not give the number of checkpoint comparisons. The gate chooses sign at step 150 and remains locked through step 2,000, whereas the paired checkpoint measurements report a spectral advantage around step 700. Thus the assignment is already fixed for roughly 550 updates when the preferred candidate changes. The summary figure distinguishes the post-switch score and its running average from a no-switch counterfactual trace. Their checkpoint correspondence is not specified in the summary; their crossings should not be equated with the held-out-loss crossover.

Assignment quality versus online adaptation. The rounded gap between the original gate and PP is $5.45 - 4.92 = 0.53$; PP is reported to outperform the gate on all five seeds. The warm-up variant improves the displayed gate loss to 5.23, and a frozen-route replay also displays 5.23. This equality concerns the warm-up variant, not the original 5.45 gate, and rounded endpoints alone do not prove identical trajectories. Allowing reversal leaves the displayed final loss at 5.45, and no return switch is reported. Inertia in the smoothed decision statistic is a plausible explanation, but the separate causal effects of the averaging horizon, lock and warm-up have not been isolated. These results support stage-dependent candidate preferences and diagnose limitations of the routing rule; they do not establish that ongoing measurement beats a static assignment, that the score is universally calibrated, or that the conservative-step bounds are empirically certified.

Separating the two evaluation questions. The checkpoint comparison tests whether the score predicts the better candidate update at the current state; the assignment matrix tests whether acting on the signal improves a complete training run. Agreement in the former does not imply improvement in the latter. The Q_{actual} result in Appendix H comes from a separate development protocol and a different predictor, so it neither explains the follow-up's misses nor establishes that step size causes them. Likewise, the replay's rounded endpoint does not establish that the routing rule alone causes the gate's gap to the static spectral baseline. A diagnosis requires linking checkpoint calibration, raw and smoothed scores, realised routes and matched controller ablations.

Warm-up without hindsight. Choosing step 700 because the same runs reveal a crossover would use evaluation outcomes to select the policy. A follow-up should instead choose a warm-up duration

or stability rule using designated selection runs, freeze it, and evaluate on fresh seeds with all four static assignments, early one-shot selection and a fixed schedule. Equal tuning budgets and the same training protocol are needed. Raw score, smoothed score, realised geometry, switch times and lock state should be logged together. Reversal is a separate extension, with its averaging rule and thresholds frozen before evaluation.

Reproducibility boundary. The summary establishes the model/data family, training horizon, main five-seed comparison and three-seed diagnostic scope. It does not establish exact seed identities, all optimizer and schedule settings, candidate step scaling, checkpoint/holdout sampling, the warm-up rule, the replayed policy, or seed counts for every additional arm. Per-seed results, configuration hashes, route traces and the error-bar definition remain to be linked to the reported figure before treating this series as a fully reproducible benchmark. The checkpoint evidence is therefore retained at its reported scope, and is not used to override the broader scheduled-update development errors in Appendix H.

J Additional details for the theoretical comparison

Proof of Proposition 1.

Proof. The two-sided Taylor remainder implied by L_g-smoothness gives

$$\eta\mu_g - \frac{L_g\eta^2}{2} \le \Delta_g(\eta) \le \eta\mu_g + \frac{L_g\eta^2}{2}, \qquad \mu_g := \mathbb{E}[\langle G,\, D_g\rangle \mid W]. \tag{35}$$

The alignment bound and Hölder's inequality give $S_g \le \mu_g \le M_g$. Substituting $\eta_g = S_g/L_g$ yields $B_g \le \Delta_g \le S_gM_g/L_g + S_g^2/(2L_g) = c_gB_g$; taking ratios proves Eq. (8). □

Activation moment assumptions.

Definition 2 (Activation noise-to-signal ratio)**.** *For a pointwise activation σ and $\gamma \sim \mathcal{N}(0,1)$ with $\mathbb{E}[\sigma(\gamma)] \neq 0$,*

$$s_\sigma \;:=\; \frac{\mathbb{E}[\sigma(\gamma)^2]}{(\mathbb{E}[\sigma(\gamma)])^2}. \tag{36}$$

Proposition 7 (Homogeneous activation specialization of Davis & Drusvyatskiy, 2025, Cor. 2.7)**.** *Let X have nonzero columns, let W have iid Gaussian entries with the scaling of the cited result, and let σ satisfy its homogeneity, growth, and nonzero-mean assumptions. Then, with the probability and finite-width factors stated in the cited corollary, $\mathrm{st}(\sigma(WX))$ is bounded by the corresponding $(1 \pm \varepsilon)$ factor times s_σ.*

For homogeneous ReLU and ReLU ^{2}, the unit-scale ratios are π and 6, respectively. SiLU and GELU are non-homogeneous and require the scale-uniform form of the full theorem; their unit-scale values (≈ 8.33 and ≈ 5.34) are descriptive statistics, not constants supplied by the homogeneous corollary. Appendix D states the scope in detail.

Frobenius comparison. For $E = \widehat{G} - G$, take $D_{\mathrm{F}} = \widehat{G}/\left\|\widehat{G}\right\|_{\mathrm{F}}$ when $\widehat{G} \neq 0$, and $D_{\mathrm{F}} = 0$ otherwise. Cauchy–Schwarz and the triangle inequality give

$$\langle G,\, D_{\mathrm{F}}\rangle \ge \|G\|_{\mathrm{F}} - 2\,\|E\|_{\mathrm{F}}\,.$$

With $\nu_{\mathrm{F}} = \mathbb{E}[\|E\|_{\mathrm{F}} \mid W]$ and Frobenius smoothness $L_{\mathrm{F}} > 0$, the conservative step yields $B_{\mathrm{F}} = (\|G\|_{\mathrm{F}} - 2\nu_{\mathrm{F}})_+^2/(2L_{\mathrm{F}})$. In the positive-signal regime, the same proof gives pairwise bounds involving B_{F}. This extends the analytical comparison without requiring a three-way selector; it does not show that Frobenius updates are empirically dominated.