BEYOND NON-ZERO INITIALIZATION: INVESTIGATING UPDATE BALANCE IN STRUCTURED LORA INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation (LoRA) fine-tunes pretrained models through two small factor matrices. Structured initializers choose a nonzero adapter using spectral or contextual information, but even the same initialized adapter can be represented by factor pairs that learn differently. We identify their relative scale as an overlooked degree of freedom: reciprocal rescaling leaves the initial adapter and frozen residual unchanged while redistributing the factors' contributions to an update. Our analysis shows how the selected structure determines this allocation and distinguishes correctable imbalance from limits that scalar rescaling cannot overcome. For orthogonal spectral factors, we derive an exact population scaling rule under isotropic Gaussian inputs and single-sample SignSGD. This leads to spectral and activation-based scale choices; InitCal complements them with reversible AdamW probes that measure the optimizer's response. Across Qwen3 and Llama models, calibration improves the four-task average in all nine structured model–initializer settings, with gains in 35 of 36 task-level comparisons. Paired controls show that update magnitude alone does not explain these improvements. Factor-scale allocation thus provides a practical way to improve structured initialization without changing the adapter it selects. The code is available here.

1 INTRODUCTION

Low-Rank Adaptation (LoRA) adapts a pretrained model by freezing its weights and training a low-rank update sBA, substantially reducing the number of trainable parameters (Hu et al., 2022). Its initialization determines not only the starting model but also how the two factors begin learning. Standard LoRA initializes one factor randomly and the other to zero, leaving the pretrained function unchanged. Nonzero initialization opens a broader design space: both factors can participate from the first update, and their initial product can encode information useful for adaptation (Li et al., 2025).

Structured initialization methods exploit this freedom in different ways. PiSSA initializes the trainable factors from principal singular components of pretrained weights (Meng et al., 2024), whereas MiLoRA uses minor components (Wang et al., 2025a); CorDA incorporates context information into the selected directions (Yang et al., 2024). For each method, $P_0 = sB_0A_0$ is the selected initial structure in the trainable branch, and $R_0 = W_0 - P_0$ remains frozen. Holding this decomposition fixed retains the initializer's structural choice, not just the starting predictions.

Selecting this structure, however, does not uniquely determine how it is used during learning. The factors cB_0 and $c^{-1}A_0$ represent the same adapter with unchanged factor directions for any $c > 0$, yet the optimizer acts on the factors separately. An update to A is multiplied by B, and an update to B is multiplied by A. Their relative scales can therefore change the two contributions to the layer output, even when the adapter matrix, frozen residual, and starting predictions are identical (Yen et al., 2025).

Prior analyses show why this distinction matters. LoRA initialization studies and LoRA+ connect factor initialization and factor-specific learning rates to effective feature learning (Hayou et al., 2024a;b). Beyond Zero studies how nonzero initialization improves robustness to suboptimal learning rates (Li et al., 2025), while μA characterizes stable update scales as width and rank grow together (Chen et al., 2026). Together, these works motivate examining initialization through the updates it induces.

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.

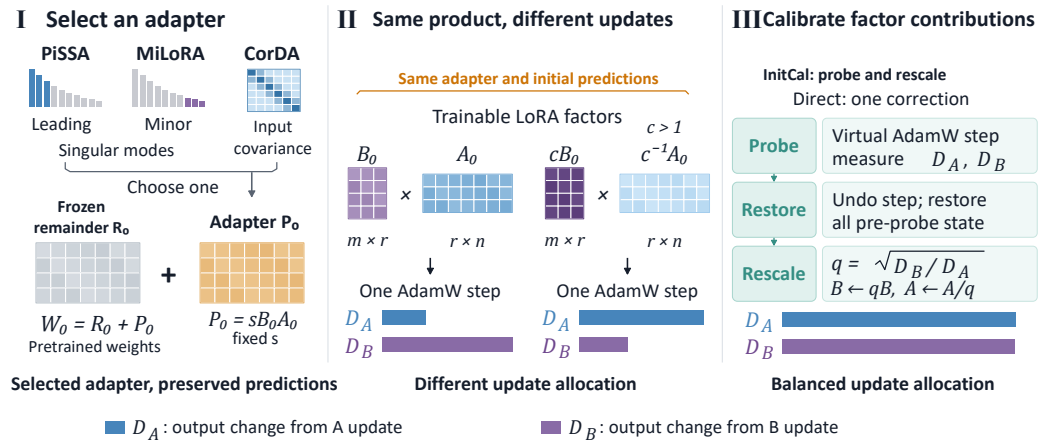

Figure 1: **The same initialized adapter can begin learning differently.** A structured initializer fixes P_0 and $R_0 = W_0 - P_0$. Reciprocal factor scaling preserves P_0 and f_0 but changes the contributions $B_0\Delta A$ and $\Delta B A_0$. The allocation can be selected analytically or measured by an AdamW probe. Probe calibration restores the state before product-preserving rescaling.

For a selected structured adapter, the remaining problem is *how to allocate scale between its factors so both contribute effectively without changing the initial adapter or residual.*

Our analysis distinguishes correctable allocation imbalance from scalar-family feasibility: reciprocal rescaling changes the contribution ratio, but cannot repair incompatible linear and interaction scales. Even at fixed shape and rank, the full selected spectrum can change the required ratio. For rectangular or square layers with orthogonal square-root spectral factors, Spectrum-full follows an exact Gaussian single-sample population formula. Forward-moment instead uses actual activation moments, retaining the single-sample update and column-orthogonality conditions without requiring Gaussian inputs.

Building on initialization-time scale calibration (Krähenbühl et al., 2016; Zhu et al., 2021), InitCal complements these predictions with actual AdamW responses (Figure 1). Each probe is undone before product-preserving rescaling; training uses a fresh optimizer. The resulting comparison separates information available from the initializer, calibration activations, and optimizer responses.

Both routes select ratios, rather than introduce a distinct optimization mechanism: same-probe dual learning rates provide an equivalent implementation under ideal updates. A unit target is a symmetric default, not a loss optimum. Paired comparisons examine which information helps: shared and shuffled scales test layer correspondence, while an additional seed group uses previously selected hyperparameters unchanged. Allocation, update magnitude, and downstream utility are assessed separately.

Our contributions are:

- A structured extension of nonzero LoRA scale analysis: selected spectra and signal statistics determine allocation beyond shape and rank, while feasibility imposes a separate magnitude condition.
- Two scale-selection routes: theory-derived Spectrum-full and Forward-moment, and optimizer-response InitCal, assessed through numerical checks and paired controls of allocation and downstream utility.

2 Structure-Preserving Two-Sided Updates

Consider a pretrained linear map $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and a rank-r adapter with $A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$, and multiplier $s > 0$. A dual-nonzero initializer constructs A_0, B_0 and compensates the frozen weight, as in prior nonzero constructions (Meng et al., 2024; Li et al., 2025):

$$P_0 := sB_0A_0, \qquad W_0 = R + P_0, \qquad h_0(x) = Rx + sB_0A_0x = W_0x. \tag{1}$$

Our core initializers are PiSSA, MiLoRA, and CorDA (Instruction-Previewed Mode, CorDA-IPM) (Meng et al., 2024; Wang et al., 2025a; Yang et al., 2024). Random Init[AB] uses independent,

per-factor fan-in Kaiming-uniform draws and the same residual compensation; the exact LoRA convention is in Appendices D.3 and G.

With R fixed and $c > 0$, reciprocal rescaling $(B_0, A_0) \mapsto (cB_0, c^{-1}A_0)$ preserves P_0 and the initial function exactly (Appendix B.1). It also preserves factor row and column spaces, but not norms or all factor-level objectives (Appendix E.5); finite-precision errors are monitored separately.

Hold the pre-step input x fixed and write $\Delta A_t = A_t - A_{t-1}$ and $\Delta B_t = B_t - B_{t-1}$. Following the feature-update decomposition used in earlier LoRA analyses (Hayou et al., 2024b; Li et al., 2025), the local output change is

$$\Delta Z_B^t = \underbrace{sB_{t-1}\Delta A_t x}_{\delta_A^t} + \underbrace{s\Delta B_t A_{t-1} x}_{\delta_B^t} + \underbrace{s\Delta B_t \Delta A_t x}_{\delta_{AB}^t}, \tag{2}$$

$$D_j^t := \frac{\|\delta_j^t\|_F}{\sqrt{N_{\text{tok}} d_{\text{out}}}}, \qquad j \in \{A, B, AB\}. \tag{3}$$

The first two terms are the linear update contributions; the third is their interaction. Each D_j^t is an output-coordinate RMS over N_{tok} tokens. We study balance of these output contributions, rather than factor-norm or Gram-matrix balance (Du et al., 2018; Castin et al., 2026).

Using the update decomposition and scaling tools of prior feature-learning analyses (Hayou et al., 2024b; Li et al., 2025; Chen et al., 2026), we isolate the following componentwise criterion:

Definition 1 (Two-sided non-degenerate update)**.** Along a width–rank scaling family, an analyzed step has two non-degenerate linear contributions and a bounded interaction when

$$D_A^t = \Theta(1), \qquad D_B^t = \Theta(1), \qquad D_{AB}^t = O(1). \tag{4}$$

The bounds hold uniformly under Appendix A. This criterion permits a vanishing interaction, unlike the three-active-term efficiency definition in μA. A non-vanishing combined update additionally requires no leading-order cancellation between the three vectors; loss improvement is a separate property. Standard $B_0 = 0$ LoRA has $D_A^1 = 0$ at its first normalized step, but later steps can satisfy the criterion.

3 Structure Determines Update Allocation

We extend the normalized-update analysis of Non-Zero LoRA (Li et al., 2025) and μA (Chen et al., 2026) to structured factorizations. Width and rank may grow jointly while the analyzed horizon stays fixed. The key distinction is between an exact reciprocal law, structural predictions under signal assumptions, and the additional requirement that both contributions remain non-degenerate.

3.1 Reciprocal law and structural information

Reciprocal gradient scaling (Yen et al., 2025) leaves SignSGD directions unchanged (Bernstein et al., 2018); the old factors still multiply the two linear paths.

Theorem 2 (Initial scaling of the three update terms)**.** *For the first coordinate-normalized update, assume bounded non-vanishing signals and the applicable aggregation conditions in Appendix A. With positive coefficients K_A, K_B, K_{AB} independent of c, η_A, η_B,*

$$D_A = \Theta(s\eta_A c K_A), \qquad D_B = \Theta(s\eta_B c^{-1} K_B), \qquad D_{AB} = \Theta(s\eta_A \eta_B K_{AB}). \tag{5}$$

For the same batch and learning rates, the first SignSGD step satisfies exactly

$$D_A(c) = cD_A(1), \qquad D_B(c) = c^{-1}D_B(1), \qquad D_{AB}(c) = D_{AB}(1). \tag{6}$$

The exact first-step identities hold for a fixed batch without rank-aggregation assumptions. Coefficient orders require those assumptions; their derivation is in Appendix E. This separates a representation identity from predictions about a particular initializer.

3.2 THE SELECTED SPECTRUM PREDICTS THE SCALE

PiSSA and MiLoRA choose different singular components (Meng et al., 2024; Wang et al., 2025a). For orthogonal square-root factors with positive selected values σ_i, isotropic Gaussian inputs, and fresh single-sample SignSGD, assume the coordinates of the output gradient and its projections onto nonzero B_0 columns are nonzero almost surely. With population output-coordinate RMS and $\rho = \eta_B/\eta_A$, Proposition 12 gives

$$c_{\mathrm{pop}}^4 = \rho^2 \frac{d_{\mathrm{out}}}{d_{\mathrm{in}}} \frac{(1-a)T_r + a(\sum_{i\in I}\sqrt{\sigma_i})^2}{T_r[1+a(d_{\mathrm{in}}-1)]}, \qquad T_r := \sum_{i\in I}\sigma_i, \quad a := 2/\pi. \tag{7}$$

The spectral concentration $q_\sigma = (\sum_i \sqrt{\sigma_i})^2/T_r \in [1, r]$ gives order $c_{\mathrm{pop}}^4 = \Theta(\rho^2 d_{\mathrm{out}} q_\sigma / d_{\mathrm{in}}^2)$. Thus identical layer shape and rank do not determine the same scale. The familiar $(r/n)^{1/4}$ order requires a sufficiently diffuse spectrum in a square layer; concentrated spectra can remove its rank factor. This is a single-sample population result, not an exact batch-AdamW formula. CorDA requires its own input statistics (Appendix M); the empirical forward rule below removes the Gaussian assumption.

3.3 BALANCE AND FEASIBLE UPDATE MAGNITUDE

Reciprocal scaling changes D_A/D_B but leaves $\sqrt{D_A D_B}$ and the ideal interaction unchanged. It therefore cannot rescue a vanishing common magnitude or an excessive interaction merely by making the two paths equal. For a fixed factorization's reciprocal scalar family, Corollary 10 gives the precise feasibility condition and its proof; it does not cover every product-preserving matrix transformation.

In particular, under the uniform orders of Theorem 2 and a rate ratio ρ bounded away from zero and infinity, the coefficient-balanced scale and non-degenerate common rate are

$$c_\star = \sqrt{\rho K_B/K_A}, \qquad \bar{\eta} := \sqrt{\eta_A \eta_B} = \Theta\Big((s\sqrt{K_A K_B})^{-1}\Big). \tag{8}$$

At this rate, $D_{AB} = \Theta(K_{AB}/(sK_AK_B))$; bounded interaction requires $K_{AB} = O(sK_AK_B)$. The two controls have distinct roles: factor scale allocates the linear update, whereas the common rate sets its strength. This is why our experiments compare both contribution ratios and combined update magnitude rather than treating either as a sufficient diagnostic.

Equal contributions minimize the larger linear term at fixed geometric mean, not downstream loss (Lemma 11). Under ideal normalized updates, the chosen scale also admits a factor-specific learning-rate implementation (Hayou et al., 2024b); Proposition 15 states its conditions. These results motivate selecting a useful allocation, rather than claiming a new optimizer or a loss-optimal ratio. Fixed-horizon extensions require the conditions at each analyzed state and provide no full-training AdamW guarantee.

4 PRODUCT-PRESERVING SCALE CALIBRATION

Three scale-selection paths use different information while preserving the adapter and residual. **Spectrum-full** uses the selected spectrum (Equation (7)); **Forward-moment** uses calibration activations; **InitCal-Direct** (Direct) measures one optimizer response. InitCal-Warm (Warm) extends the response-based path with a coarse analytic start and an additional correction, as detailed below.

4.1 FORWARD-MOMENT: EMPIRICAL ACTIVATION STATISTICS

For factors with nonzero orthogonal columns of B_0, the same-sample SignSGD paths yield a distribution-free rule (Appendix N):

$$c_{\mathrm{fwd}}^4 = \rho^2 \frac{d_{\mathrm{out}} \widehat{\mathbb{E}}\|A_0 x\|_1^2}{\|B_0\|_F^2 \widehat{\mathbb{E}}\|x\|_1^2}. \tag{9}$$

Forward-moment accumulates these squared norms over valid calibration tokens in one forward pass, without centering activations or computing gradients. Unlike Spectrum-full, it does not assume Gaussian inputs. Nonzero signal coordinates and the single-sample update model remain necessary; applying it to batch AdamW is a scale-selection rule, not an exact balance identity.

4.2 DIRECT: ONE MEASURED CORRECTION

InitCal follows optimizer-aware initialization (Zhu et al., 2021) by measuring a virtual update with the training AdamW configuration (Loshchilov & Hutter, 2019). After restoring all virtual state, positive, numerically resolved contributions give

$$q_\ell := \sqrt{\frac{D_{B,\ell}}{D_{A,\ell}}}, \qquad B_\ell \leftarrow q_\ell B_\ell, \qquad A_\ell \leftarrow q_\ell^{-1} A_\ell. \tag{10}$$

Under the ideal law, one correction removes any positive starting scale (Corollary 13). Warm adds the coarse start in Equation (94) and a second correction; it is an empirical enhancement, not a theoretically necessary step. Random Init[AB] uses two direct corrections because its factors need not satisfy the forward rule's orthogonality condition.

The measured contributions D_A, D_B live in output space, rather than factor-gradient space. The virtual step uses the training loss, masking, clipping, and AdamW configuration. Temporary parameter increments and optimizer state are restored before rescaling, and formal training starts with a fresh optimizer. The procedure flags unresolved channels and adds no parameters or persistent training operations (Algorithm 1). Each layer receives its own scalar from the shared calibration observations, without adding a formal training step or changing the initial adapter.

5 EXPERIMENTS

5.1 PROTOCOL AND STRUCTURAL PREDICTION

We first test whether selected spectra explain response differences at fixed shape and rank. All raw–calibrated comparisons preserve the initial product and residual. Models are Qwen3-0.6B/1.7B/4B (Team, 2025) and Llama-3.2-3B. Fixed Tulu-3 (Lambert et al., 2024) splits provide 80k training and 32k selection conversations. Final-step validation NLL selects rates; held-out NLL and GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), MMLU-Pro (Wang et al., 2024b), and BBH (Suzgun et al., 2023) never enter selection. Focused configurations use five paired seeds; the broader matrix uses three. Appendix G gives the complete data, training, and statistical protocol.

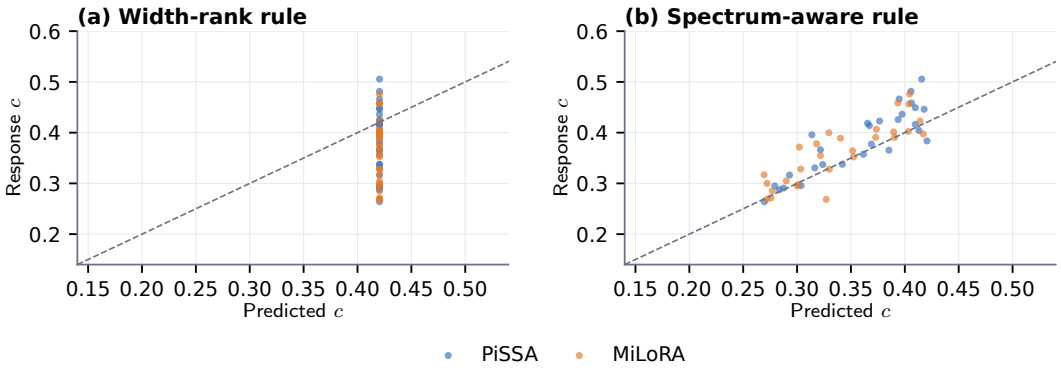

Figure 2: **Structural predictions under batch AdamW.** Qwen3-0.6B, $r = 32$, $n = 1024$; 28 square `k_proj` layers per initializer. Both panels use the same five-seed median held-out response scales. Table 19 compares update rules and post-probe imbalance.

Structure beyond width and rank. Figure 2 shows layer-to-layer variation that the width–rank rule cannot express. The selected spectrum follows the measured response more closely. Errors are larger under batch updates than single-example normalization (Table 19), making the spectral rule a predictor rather than an exact AdamW identity.

Separately executed Gaussian matrix checks give median/maximum relative scale errors of 0.11%/0.51% across 60 rectangular and square configurations (Appendix I).

5.2 ALLOCATION CHOICES AND INFORMATION ACCESS

We next separate scale-selection methods from controls of search, implementation, and update magnitude. Spectrum-full applies the complete rule to all 112 eligible modules; Direct and Warm use optimizer responses. Table 1 groups these methods alongside independent ratio search and mechanism controls. Five candidate settings are matched, with a 25-setting dual-rate search as a strong reference.

Table 1: **Scale-selection methods and matched controls.** Four-task gains over paired Raw, in percentage points; five-seed mean ± SD. Candidates are complete-training settings per seed; "–" inherits Warm's selected rate. PiSSA: Qwen3-0.6B, $r = 32$; MiLoRA: Qwen3-1.7B, $r = 64$.

Method	Candidate settings	PiSSA paired gain	MiLoRA paired gain
Scale-selection methods			
Spectrum-full	5	+0.80±0.16	+0.54±0.14
Direct	5	+0.70±0.14	+0.55±0.06
InitCal-Warm	5	+1.00±0.21	+0.70±0.06
Independent ratio search			
Tuned dual LR (5)	5	+0.66±0.04	+0.41±0.18
Tuned dual LR (25)	25	+0.70±0.04	+0.47±0.19
Mechanism controls			
Layer-norm match	5	+0.26±0.17	+0.23±0.11
Warm-mapped LR	–	+1.00±0.21	+0.71±0.11

Spectrum-full gains 0.80/0.54 points over Raw on PiSSA/MiLoRA; the structural prediction itself thus gives a competitive probe-free choice. Warm adds 0.21/0.16 points on average. Paired intervals resolve this additional gain differently across configurations (Table 22); the analytic rule and response correction are complementary choices rather than a uniformly ranked pair. Layer-norm match retains Raw factors while matching Warm's initial combined RMS in every module; its smaller gains show that amplitude alone does not recover these outcomes. Same-probe and cumulative-scale learning-rate controls test the alternative implementation (Hayou et al., 2024b); their close results support the selected allocation rather than a unique benefit of rescaling. Full-network and paired controls appear in Appendices K and L (Milsom et al., 2025; Zhang et al., 2025c).

With selected settings frozen, an additional five-seed group does not establish a stable ranking of Warm and the strong alternatives. All four paired 95% reference intervals for Warm versus Spectrum-full or Tuned dual LR (25) include zero (Table 34). These comparisons motivate treating Warm as an empirical extension, rather than a uniformly preferred route.

Activation statistics versus optimizer response. Forward-moment and Direct use the same four calibration conversations. The former reads raw activation moments without backward computation; the latter probes the actual optimizer. Their three paired difference intervals contain zero (Table 32), so the results do not establish a consistent advantage for backward information. On CorDA, Forward-moment exceeds the Gaussian covariance reference by $+0.536$ points, with a paired 95% reference interval of $[+0.067, +1.005]$. This contrast is consistent with a benefit from using the observed activation distribution, even without a backward pass.

Layer-specific allocation. Type-shared replaces scales by their geometric mean within module-type/shape groups; shuffling permutes the same scales inside each group. Both retain Direct's information and independently select a rate. Direct minus sharing has a positive interval in both Qwen configurations; Direct minus shuffling does so only for MiLoRA. The benefit of retaining layer correspondence therefore depends on the configuration (Figure 3). Sharing removes within-group scale variation, whereas shuffling preserves its distribution but breaks correspondence with individual layers. The controls therefore distinguish the availability of heterogeneous scales from their assignment, without changing which calibration observations the methods can access.

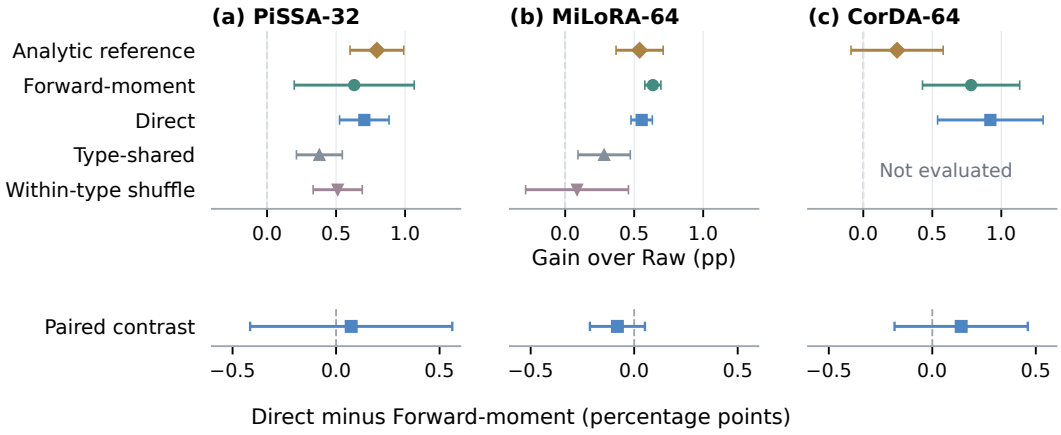

Figure 3: **Information source and layer assignment.** Upper panels: four-task gains over paired Raw. Lower panels: seed-paired Direct-minus-Forward contrasts. Points are five-seed means; bars are 95% t reference intervals conditional on selected settings, without multiplicity adjustment. The analytic reference is Spectrum-full for PiSSA/MiLoRA and Covariance-aware for CorDA. All methods use five candidates; shared/shuffled controls are evaluated only on the two Qwen configurations. Models and ranks match Tables 1 and 31.

5.3 Update dynamics and selection budget

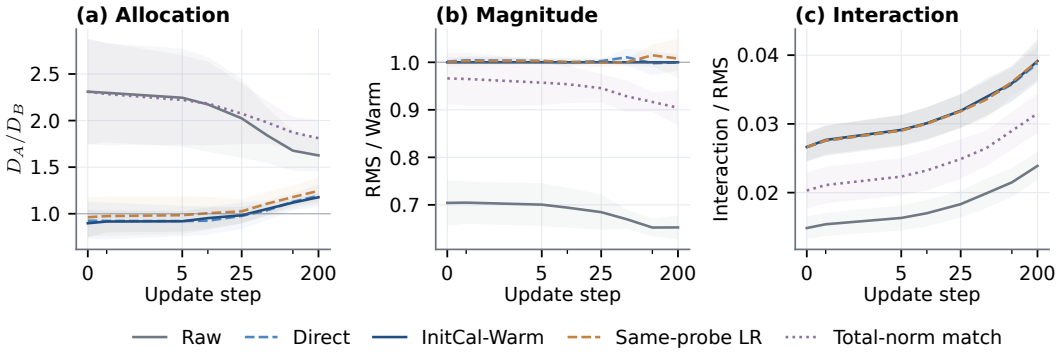

Figure 4: **Early-update allocation diagnostics.** Five-seed mean and SD. A common held-out batch probes each method's own evolving state at its fixed selected peak rate. Left to right: D_A/D_B; total RMS relative to paired Warm; interaction RMS relative to total RMS. Norms and signed cross terms use the output-space Gram statistics in Equation (96).

Figure 4 tracks allocation, magnitude, and interaction after the initial correction. Raw and Total-norm match both begin at $D_A/D_B = 2.31$ but reach 1.63 and 1.81 at step 200. Direct and Warm drift from 0.93/0.90 to 1.19/1.18. Calibration changes the initial allocation without fixing the subsequent trajectory. Matching the initial total response does not reproduce the calibrated ratio trajectory, separating allocation from a simple increase in update size. The average interaction grows with training but remains small at these measured checkpoints.

Matched-total-RMS controls at target ratios $0.5, 1, 2$ do not identify ratio one as a unique loss optimum: both neighboring NLL contrast intervals include zero (Table 18). Direct-2 reduces residual imbalance but does not improve the early loss decrease in either configuration (Table 24). Matching total RMS fixes update length, not its direction relative to the loss gradient. Changing the mixture of the two paths can therefore change the loss response even at equal RMS. Balance supplies a scale-setting default; task scores and held-out NLL evaluate the subsequent optimization outcome.

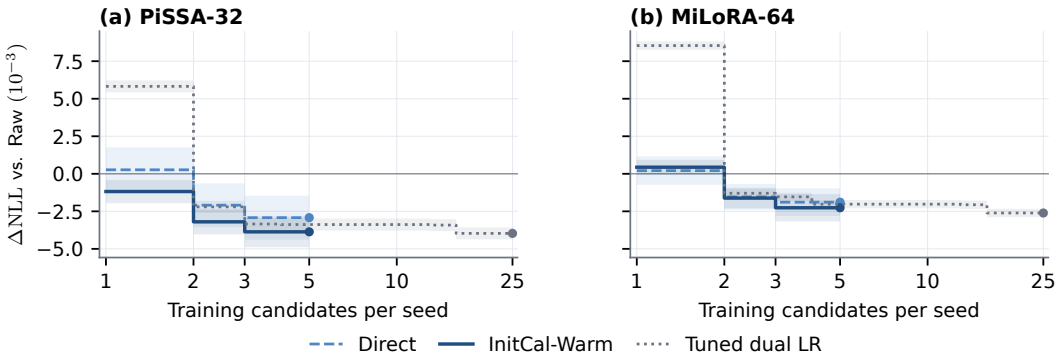

Figure 5: **Validation selection at increasing candidate budgets.** Held-out NLL difference from paired, selected Raw; lower is better. Lines and shading show five-seed mean and sample SD, conditional on the selected candidate. Budgets use logarithmic spacing. Direct and Warm end at five settings; the independent dual-rate grid has 25.

Performance at matched search budgets. Figure 5 uses a fixed candidate order and validation-only prefix selection. Warm is competitive at five settings, while the 25-setting dual-rate search reaches lower held-out NLL in both configurations. Direct does not uniformly beat the five-setting grid. The comparison exposes a budget–performance trade-off; candidate count is not wall-clock cost, and five-versus-five saves no complete-training candidates.

Calibration supplies layer-specific ratios before the common-rate search, whereas the independent grid jointly searches a global ratio and common rate. Thus, the 25-setting reference measures what additional complete-training trials can recover beyond the smaller candidate budget, not an intrinsic advantage of either implementation. The curve and Table 1 also measure different outcomes: reporting-split NLL and the four-task average, respectively. Both are evaluated after validation-only selection. Their rankings need not coincide, so a lower held-out NLL for the larger search is compatible with a higher task average for Warm.

5.4 Downstream breadth and calibration cost

Table 2: **Downstream task scores at rank 64.** Training-seed mean ± SD across three seeds, in percent. Avg. weights all twelve model–task means equally. Calibration uses Warm for structured initializers and two direct corrections for Init[AB]. Paired reference intervals appear in Table 6.

Method	Qwen3-1.7B				Qwen3-4B				Llama-3.2-3B				Avg.
	GSM8K	MATH 500	MMLU Pro	BBH	GSM8K	MATH 500	MMLU Pro	BBH	GSM8K	MATH 500	MMLU Pro	BBH	
Standard LoRA	72.5±0.9	31.7±1.7	34.1±0.5	40.0±0.4	84.5±0.3	41.1±1.3	47.0±0.3	52.4±0.5	33.2±1.2	7.8±0.0	24.6±0.2	33.2±0.3	41.9
PiSSA	68.3±0.3	35.1±1.4	35.3±0.6	41.9±0.3	86.0±0.1	42.7±0.2	48.2±0.3	52.6±0.6	32.8±0.4	9.6±1.0	24.1±0.2	33.5±0.9	42.5
+ Calibration	69.0±0.8	35.2±0.3	35.6±0.4	42.6±0.3	86.2±0.4	43.1±0.7	49.0±0.2	53.4±0.3	34.0±0.5	10.4±0.6	24.5±0.4	34.6±0.1	43.1
MiLoRA	67.2±2.1	34.6±0.3	35.2±0.5	42.5±0.6	85.2±0.6	42.9±0.5	46.8±0.1	53.6±0.5	33.4±0.9	8.3±0.9	25.3±0.5	32.8±0.9	42.3
+ Calibration	68.6±2.2	35.8±1.5	36.3±0.1	43.4±0.8	85.0±0.4	43.7±0.5	47.3±0.3	54.8±0.3	34.5±0.7	9.4±0.9	25.5±0.3	34.3±0.6	43.2
CorDA-IPM	71.8±0.2	31.6±1.4	34.7±0.1	41.1±0.3	79.8±0.8	41.3±0.8	46.7±0.1	53.1±0.4	31.2±0.3	9.1±0.7	22.1±0.3	34.5±0.7	41.4
+ Calibration	72.7±0.4	34.0±1.3	35.5±0.2	42.1±0.3	81.2±0.4	42.3±0.8	47.6±0.2	53.5±1.4	32.1±0.6	9.7±0.3	22.8±0.2	34.8±0.1	42.4
Random Init[AB]	73.4±0.6	31.6±1.2	34.4±0.7	39.7±0.2	83.8±0.4	42.4±1.5	47.0±0.4	52.9±0.1	34.5±0.5	7.2±0.4	24.5±0.2	32.6±0.3	42.0
+ Calibration	73.2±0.7	31.9±1.4	34.3±0.4	40.5±0.3	84.3±0.4	43.0±0.9	47.4±0.2	53.7±0.4	34.1±0.7	8.1±0.3	25.2±0.1	33.2±0.3	42.4

All nine structured model–initializer pairs improve their four-task mean, averaging $+0.82$ points; 35 of 36 task cells improve. The exception is Qwen3-4B/MiLoRA on GSM8K, $85.2 \rightarrow 85.0$. Improving an initializer does not imply beating Standard LoRA: calibrated Qwen3-1.7B/PiSSA reaches 69.0 on GSM8K versus 72.5. Init[AB] improves by $+0.40$ points on average with task-level regressions. Six of nine structured paired reference intervals include zero; tasks are not independent replicates. Complete curves retain heterogeneous near-optimal windows (Figure 8).

Task-level effects retain the initializer's trade-offs. The full scores in Table 2 show why aggregate gains alone are insufficient. On Qwen3-1.7B, PiSSA changes MATH-500 only from 35.1 to 35.2, whereas MiLoRA changes it from 34.6 to 35.8 and CorDA from 31.6 to 34.0. On Qwen3-4B, calibrated PiSSA exceeds Standard LoRA on all four task means, but calibrated CorDA still trails it on GSM8K while exceeding it on the other three. Calibration thus improves how a chosen adapter is used without erasing the task-dependent consequences of choosing that adapter. The random nonzero control also remains distinct: a positive overall change coexists with regressions on individual tasks.

On one RTX A6000 (48 GB), complete Warm calibration takes 14.3 seconds, or 0.42% of a 57.3-minute fine-tuning run, with reported peak memory of 11.6 GiB in either complete process (Table 8). This measures full Warm under the stated configuration, not Direct or a decomposed per-probe cost.

6 Related Work

Structured initialization and feature-learning scales. Nonzero factors can encode weight, data, or gradient information. PiSSA and MiLoRA select principal and minor weight components, respectively (Meng et al., 2024; Wang et al., 2025a), while CorDA uses context-dependent decompositions (Yang et al., 2024). LoRA-GA and LoRA-One construct factors using gradient information (Wang et al., 2024a; Zhang et al., 2025b); OLoRA uses QR decomposition (Büyükakyüz, 2024). For PiSSA, MiLoRA, and CorDA-IPM, our scalar-rescaling analysis retains the selected initial matrix; Appendix E.5 treats additional factor-level objectives separately. Infinite-width feature learning provides criteria for updates that remain active as models grow (Yang & Hu, 2021). LoRA initialization analyses and LoRA+ apply this perspective to factor initialization and unequal learning rates (Hayou et al., 2024a;b); Beyond Zero connects nonzero initialization to learning-rate robustness (Li et al., 2025). μA extends scaling analysis to jointly growing width and rank (Chen et al., 2026). Other interventions control adapter magnitude through rank scaling (Kalajdzievski, 2023), early factor shrinkage (Wu et al., 2026), or separation of weight magnitude and direction (Liu et al., 2024). Our analysis links two-factor allocation to absolute update scale and interaction feasibility within each reciprocal scalar family.

Initialization calibration and factor-aware optimization. Function-preserving, data-dependent initialization balances relative layerwise update rates (Krähenbühl et al., 2016); GradInit chooses parameter-block scales by evaluating virtual optimizer updates (Zhu et al., 2021). Function-space learning-rate matching (FLeRM) measures parameter blocks' final-output changes for hyperparameter transfer (Milsom et al., 2025). InitCal instead allocates the two local contributions at fixed adapter product, rather than matching combined-update magnitude. ScaledGD uses Gram preconditioning to address ill-conditioning in low-rank estimation (Tong et al., 2021). LoRA-Pro studies how factor updates approximate full-weight gradients, including a symmetric-contribution solution (Wang et al., 2025b). LoRA-RITE and LoRA-Muon pursue reparameterization-invariant optimization (Yen et al., 2025; Cesista et al., 2026); RefLoRA refactors or preconditions updates (Zhang et al., 2025a). BaLoRA balances factor Gram matrices while preserving the adapter, motivated by conditioning at minimizers (Castin et al., 2026); SDS-LoRA addresses within-factor spectral anisotropy (Oh et al., 2026). InitCal targets initial output-contribution allocation through reciprocal scalar rescaling, followed by ordinary AdamW training.

7 Conclusion

Structured initialization selects an adapter, but leaves its update allocation undetermined. Even at fixed shape and rank, spectral and signal statistics can change the required ratio. Our analysis separates correctable imbalance from scalar-family feasibility. Spectrum-full and Forward-moment turn these predictions into scale choices; InitCal complements them with optimizer responses. Numerical checks and paired controls assess allocation, layer correspondence, and downstream utility. Both routes preserve the initializer's selected adapter while using its structure and signal statistics or the optimizer's measured response to select factor scales.

Reproducibility Statement

The analysis assumptions are stated in Section 3 and Appendix A; algebraic proofs and conditional scaling derivations appear in Appendices B to E. Calibration and state restoration are specified in Section 4 and Appendix F. Appendix G records the data, training, and statistical protocol, and the numerical checks are collected in Appendix I.

AI Use Statement

Generative AI tools were used in three ways: to aid and polish the writing so that the language is consistent and standard; to survey the existing literature and identify the theoretical gap this paper addresses; and to assist with the experiments and code. The authors are responsible for the scientific claims, mathematical arguments, implementation, and final reported results.

References

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SIGNSGD: compressed optimisation for non-convex problems. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 559–568. PMLR, 2018. URL http://proceedings.mlr.press/v80/bernstein18a.html.

Kerim Büyükakyüz. Olora: Orthonormal low-rank adaptation of large language models. *CoRR*, abs/2406.01775, 2024. doi: 10.48550/ARXIV.2406.01775. URL https://doi.org/10.48550/arXiv.2406.01775.

Valérie Castin, Kimia Nadjahi, Pierre Ablin, and Gabriel Peyré. Balanced LoRA: Removing parameter invariance to accelerate convergence. *arXiv preprint arXiv:2605.31484*, 2026. URL https://arxiv.org/abs/2605.31484.

Franz Louis Cesista, Katherine Crowson, Cédric Simal, and Stella Biderman. Lora-muon: Spectral steepest descent on the low-rank manifold. *CoRR*, abs/2606.12921, 2026. doi: 10.48550/ARXIV.2606.12921. URL https://doi.org/10.48550/arXiv.2606.12921.

Nan Chen, Soledad Villar, and Soufiane Hayou. Learning rate scaling across LoRA ranks and transfer to full finetuning. *arXiv preprint arXiv:2602.06204*, 2026. URL https://arxiv.org/abs/2602.06204.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 382–393, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/fe131d7f5a6b38b23cc967316c13dae2-Abstract.html.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d4387c37b3b06e55f86eccdb8cd1f829-Abstract-Conference.html.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806. PMLR, 2024b. URL https://proceedings.mlr.press/v235/hayou24a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1026–1034. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.123. URL https://doi.org/10.1109/ICCV.2015.123.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023. URL https://arxiv.org/abs/2312.03732.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent initializations of convolutional neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06856.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024. doi: 10.48550/ARXIV.2411.15124. URL https://doi.org/10.48550/arXiv.2411.15124.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Shiwei Li, Xiandi Luo, Xing Tang, Haozhao Wang, Hao Chen, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Beyond zero initialization: Investigating the impact of non-zero initialization on LoRA fine-tuning dynamics. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35519–35535. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25bm.html.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=LzPWWPAdY4.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

Edward Milsom, Ben Anson, and Laurence Aitchison. Function-space learning rates. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44225–44251. PMLR, 2025. URL https://proceedings.mlr.press/v267/milsom25a.html.

Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Sds-lora: Overcoming anisotropic gradient scaling in low-rank adaptation. *CoRR*, abs/2606.16454, 2026. doi: 10.48550/ARXIV.2606.16454. URL https://doi.org/10.48550/arXiv.2606.16454.

Fabian Paischer, Lukas Hauzenberger, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, and Sepp Hochreiter. One initialization to rule them all: Fine-tuning via explained variance adaptation. *arXiv preprint arXiv:2410.07170*, 2024. URL https://arxiv.org/abs/2410.07170.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.824. URL https://doi.org/10.18653/v1/2023.findings-acl.824.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.*, 22:150:1–150:63, 2021. URL https://jmlr.org/papers/v22/20-1067.html.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4823–4836. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.naacl-long.248. URL https://aclanthology.org/2025.naacl-long.248/.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/62c4718cc334f6a0a62fb81c4a2095a1-Abstract-Conference.html.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Lora-pro: Are low-rank adapters properly optimized? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=gTwRMU3lJ5.

Yize Wu, Ke Gao, Ling Li, and Yanjun Wu. Stable-LoRA: Stabilizing feature learning of low-rank adaptation. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10006643.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/83f95bb0ac5046338ea2afe3390e9f4b-Abstract-Conference.html.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Venkata Sai Surya Subramanyam Duvvuri, Inderjit Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. LoRA done RITE: Robust invariant transformation equilibration for LoRA optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bcbc0f660d2dde42f9d1d0ecb14a6f9a-Abstract-Conference.html.

Yilang Zhang, Bingcong Li, and Georgios B. Giannakis. Reflora: Refactored low-rank adaptation for efficient fine-tuning of large models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/85c1ee0b3cdf49b84f14b163b01f2711-Abstract-Conference.html.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. Lora-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/zhang25ax.html.

Zicheng Zhang, Haoran Li, Yifeng Zhang, Guoqiang Gong, Jiaxing Wang, Junxing Hu, Pengzhang Liu, and Qixia Jiang. The primacy of magnitude in low-rank adaptation. In *Advances in Neural Information Processing Systems*, vol-

ume 38, 2025c. URL https://papers.nips.cc/paper_files/paper/2025/hash/0010665e949927b74faf6e3ada6d7f72-Abstract-Conference.html.

Chen Zhu, Renkun Ni, Zheng Xu, Kezhi Kong, W. Ronny Huang, and Tom Goldstein. Gradinit: Learning to initialize neural networks for stable and efficient training. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 16410–16422, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/88ae6372cfdc5df69a976e893f4d554b-Abstract.html.

A Notation and Scaling Assumptions

This appendix records the assumptions used in Section 3. The setup follows earlier large-width analyses of LoRA initialization and factor-specific learning rates, together with joint width–rank analysis (Hayou et al., 2024a;b; Li et al., 2025; Chen et al., 2026). Listing the conditions in one place makes clear which conclusions are inherited from that framework and which require additional structure.

A.1 Layer, features, and update convention

Use the dimensions, residual decomposition and paths of Equations (1) and (2). Write $Z_A^t = A_t x$, $Z_B^t = sB_tA_tx$, and $\widetilde{Z}_B^t = s(B_tA_t - B_0A_0)x$. The update is $A_t = A_{t-1} - \eta_A g_A^t$, $B_t = B_{t-1} - \eta_B g_B^t$; square layers have $d_{\text{in}} = d_{\text{out}} = n$.

A.2 Meaning of asymptotic order

For a vector $v \in \mathbb{R}^d$, $v = \Theta(a)$ refers to a typical coordinate or coordinate-RMS scale unless a norm is written explicitly. For a matrix, it refers to entrywise RMS scale. The number of analyzed training steps T remains fixed as $n, r \to \infty$, following the fixed-horizon width–rank setup (Chen et al., 2026). Constants hidden by $\Theta(\cdot)$ may depend on the data distribution, layer type, fixed optimizer hyperparameters, and finite training horizon, but not on n or r.

A.3 Signal assumptions

Assumption 3 (Bounded forward and backward signals)**.** *For every analyzed fixed $t \le T$, the input and backpropagated output-gradient coordinates have bounded, non-vanishing normalized RMS and sufficient moments for the stated aggregation rules:*

$$x = \Theta(1), \qquad \mathrm{d}h_t = \Theta(1). \tag{11}$$

This condition concerns normalized signal moments, not unnormalized factor energies: S_A^{act} and S_B may scale with width and rank. A first-step result assumes this condition only at initialization; a fixed-horizon extension must assume or prove it at every analyzed step.

Assumption 4 (Coordinate-normalized update)**.** *For scale analysis, we use the SignSGD proxy (Bernstein et al., 2018) for Adam-style coordinate normalization (Kingma & Ba, 2015), as in the optimizer abstraction of μA (Chen et al., 2026):*

$$g_A^t = \operatorname{sign}(\nabla_A \mathcal{L}_t), \qquad g_B^t = \operatorname{sign}(\nabla_B \mathcal{L}_t). \tag{12}$$

A positive global rescaling of a factor gradient therefore leaves the processed update unchanged. Deviations caused by finite AdamW, weight decay, clipping, or precision are measured by the virtual probe rather than absorbed into this assumption.

Assumption 5 (No asymptotic cancellation)**.** *When an iterate is the sum of an initialization term and finitely many update terms, the terms of the largest order do not cancel systematically, including when several terms have the same order.*

A.4 RANK AGGREGATION

The joint width–rank calculation follows the rank-aggregation distinction in μA (Chen et al., 2026) and depends on how sums over the rank dimension combine.

Assumption 6 (Random-factor aggregation)**.** *For independent centered random factors with bounded moments, incoherent signs add in RMS as $\sqrt{r}$, whereas sign-aligned inner products add as r. Under the coordinate-normalized update, this gives the square-layer orders*

$$D_A^t = \Theta(s\eta_A n\sqrt{r})\,\Theta(B_{t-1}), \quad D_B^t = \Theta(s\eta_B r)\,\Theta(Z_A^{t-1}), \quad D_{AB}^t = \Theta(s\eta_A\eta_B n\sqrt{r}). \tag{13}$$

Assumption 7 (Structured-factor aggregation)**.** *For fixed base factors $B_0 \in \mathbb{R}^{d_{\text{out}}\times r}$ and $A_0 \in \mathbb{R}^{r\times d_{\text{in}}}$, define*

$$S_B := \|B_0\|_F^2, \qquad S_A^{\text{act}} := \mathbb{E}_x\|A_0x\|_2^2. \tag{14}$$

We assume that (i) $\|x\|_1 = \Theta(d_{\text{in}})$; (ii) the coordinates of A_0x are not asymptotically concentrated in a vanishing subset, so $\|A_0x\|_1 = \Theta(\sqrt{rS_A^{\text{act}}})$; and (iii)

$$\frac{\|B_0\,\text{sign}(B_0^\top g)\|_2}{\sqrt{d_{\text{out}}}} = \Theta\left(\sqrt{\frac{S_B}{d_{\text{out}}}}\right). \tag{15}$$

For the interaction term, we additionally assume

$$\chi_r := \left(\mathbb{E}_{x,g}\left|\text{sign}(A_0x)^\top \text{sign}(B_0^\top g)\right|^2\right)^{1/2} = \Theta(\sqrt{r}). \tag{16}$$

This is an explicit structured sign-aggregation assumption, motivated by normalized-update width–rank analysis (Chen et al., 2026). Centered rank contributions with sufficiently weak correlations provide one sufficient setting. Conditions (i)–(iii) alone do not imply it. With this additional condition,

$$K_A = d_{\text{in}}\sqrt{S_B/d_{\text{out}}}, \qquad K_B = \sqrt{rS_A^{\text{act}}}, \qquad K_{AB} = d_{\text{in}}\sqrt{r}. \tag{17}$$

For structured factors, the interaction condition requires separate justification from the factor-energy conditions. In general the interaction coefficient is of order $d_{\text{in}}\chi_r$; aligned signs can give $\chi_r = \Theta(r)$. The analytic $\sqrt{r}$ specialization is conditional on Equation (16). The finite-width probe measures the interaction directly and does not substitute this specialization. These conditions are imposed on the states being analyzed, rather than inferred for all training steps from initialization alone.

The coefficient derivation in Appendix E uses the single-sample simplification of earlier LoRA analyses (Hayou et al., 2024a; Li et al., 2025), with a normalized gradient and population RMS over signals. Applying its orders to token-batch RMS additionally requires concentration and that normalization of the batch-aggregated gradient preserves the stated orders. Independence of samples alone does not establish that property. The exact first-step reciprocal identities hold for a fixed batch without this extra assumption; they do not by themselves imply the structured coefficient formulas.

In the population-RMS convention, aggregation bounds apply to the joint second moments entering each update. In particular, the interaction coefficient $d_{\text{in}}\chi_r$ requires

$$J = \text{sign}(A_0x)^\top \text{sign}(B_0^\top g), \tag{18}$$

$$\mathbb{E}[\|x\|_1^2 J^2] = \Theta(d_{\text{in}}^2\mathbb{E}[J^2]). \tag{19}$$

Marginal RMS orders alone do not imply this mixed-moment bound. The Gaussian specialization in Proposition 12 and Appendix I evaluates the relevant moments directly.

A.5 LEARNING CRITERION

We retain the two-active-linear-term focus of earlier LoRA analyses (Hayou et al., 2024b; Li et al., 2025). Our componentwise condition permits a vanishing interaction; it is not the three-active-term efficiency definition of μA (Chen et al., 2026).

Definition 8 (Two-sided non-degenerate update over a fixed horizon)**.** A parameterization has two non-degenerate linear contributions and a bounded interaction over a specified fixed set of training steps when

$$D_A^t = \Theta(1), \qquad D_B^t = \Theta(1), \qquad D_{AB}^t = O(1) \tag{20}$$

for every step in that set, with uniform constants. This extends the per-step criterion in Definition 1.

The definition is componentwise, not a lower bound on the combined vector. To infer a non-vanishing combined update, a sufficient additional condition is

$$\|\delta_A^t + \delta_B^t + \delta_{AB}^t\|_{\mathrm{RMS}} \geq \kappa(D_A^t + D_B^t + D_{AB}^t), \qquad \kappa > 0, \tag{21}$$

uniformly in width and rank. Termwise boundedness implies bounded relative feature changes over a fixed horizon by telescoping; non-degeneration of the combined change additionally uses this condition. It is not needed for reciprocal scaling or the componentwise feasibility corollary. Exact finite equality is the symmetric calibration representative, not a descent guarantee. Fixed-horizon extensions require the signal and aggregation conditions at every analyzed state.

B Algebraic Invariance and Exact First-Step Scaling

B.1 Product and function invariance

Proposition 9 (Product and function invariance)**.** *Fix the base factors A_0, B_0. For any $c > 0$,*

$$B_0(c) := cB_0, \quad A_0(c) := c^{-1}A_0, \quad sB_0(c)A_0(c) = P_0, \quad R + sB_0(c)A_0(c) = W_0. \tag{22}$$

The rank, parameter count, row space of A_0, column space of B_0, adapter matrix, and initial function are unchanged in exact arithmetic.

Proof. $s(cB_0)(c^{-1}A_0) = P_0$, so adding the unchanged residual gives W_0 for every input. Nonzero scalar multiplication preserves ranks, spaces and parameter shapes. It also preserves the tangent set $\{B_0H_A + H_BA_0\}$: substitute $\widetilde{H}_A = cH_A$ and $\widetilde{H}_B = c^{-1}H_B$. The set is unchanged, but its optimizer-weighted contributions need not be. □

B.2 Exact first-step law under coordinate-normalized updates

Let $G = \partial\mathcal{L}/\partial h$ be the loss gradient with respect to the layer output. For a batch, G and x denote the corresponding matrices. At initialization, the factor gradients, their reciprocal rescaling, and the resulting processed updates are

$$\nabla_A\mathcal{L} = sB_0^\top Gx^\top, \qquad \nabla_B\mathcal{L} = sG(A_0x)^\top, \tag{23}$$

$$\nabla_A\mathcal{L}(c) = c\nabla_A\mathcal{L}(1), \qquad \nabla_B\mathcal{L}(c) = c^{-1}\nabla_B\mathcal{L}(1), \tag{24}$$

$$\Delta A_1(c) = \Delta A_1(1), \qquad \Delta B_1(c) = \Delta B_1(1). \tag{25}$$

The first equality uses the fact that reciprocal initialization leaves the model output, and therefore G, unchanged. The last equality holds for SignSGD, or any first-step rule that is exactly invariant to a positive global gradient scale.

Substitution into the three terms of Equation (2) gives

$$\delta_A^1(c) = sB_0(c)\Delta A_1(c)x = c\delta_A^1(1), \tag{26}$$

$$\delta_B^1(c) = s\Delta B_1(c)A_0(c)x = c^{-1}\delta_B^1(1), \tag{27}$$

$$\delta_{AB}^1(c) = s\Delta B_1(c)\Delta A_1(c)x = \delta_{AB}^1(1). \tag{28}$$

Taking the normalized RMS proves Equation (6).

For a single sample, the processed increments can also be written as

$$\Delta A_1x = -\eta_A \operatorname{sign}(B_0^\top G)\|x\|_1, \tag{29}$$

$$\Delta B_1A_0x = -\eta_B \operatorname{sign}(G)\|A_0x\|_1. \tag{30}$$

The scale c leaves the signs unchanged, multiplies the old B_0 factor by c, and multiplies $\|A_0x\|_1$ by c^{-1}.

Finite AdamW. At the first Adam step with zero moments, the coordinate-normalized part of the update is $g/(|g| + \epsilon)$ (Kingma & Ba, 2015). Positive scalar invariance is exact when $\epsilon = 0$ and approximate when $|g| \gg \epsilon$. Decoupled weight decay (Loshchilov & Hutter, 2019), non-covariant clipping histories, and finite precision can introduce deviations. Moment accumulation alone preserves the correspondence when gradients and moments obey the reciprocal scaling relation (Proposition 15). Rather than assume these effects away, the calibration step measures the realized update terms under the actual AdamW configuration.

B.3 SAME-PRODUCT SCALE COMPARISON

In the same-product comparison (Figure 6), independently tuned minima are 0.759, 0.698, and 0.729 at scale ratios $1/4, 1, 4$. These distinct responses after learning-rate retuning motivate factor-scale control, without establishing equal contributions as loss-optimal.

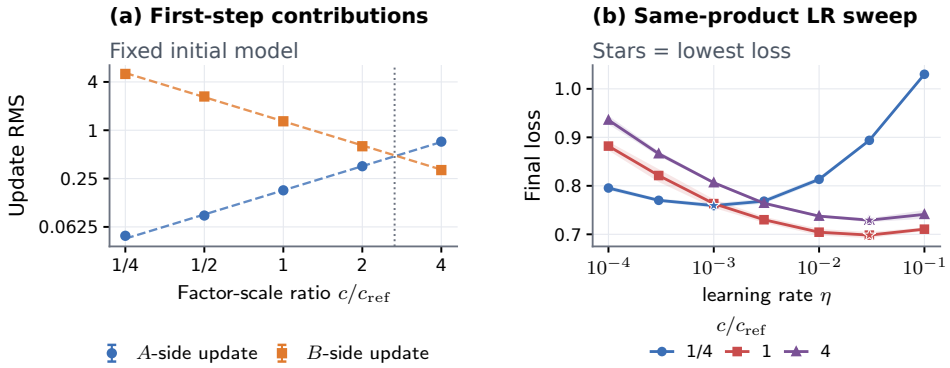

Figure 6: **Same starting model, different factor scales.** PiSSA on MNIST-to-FashionMNIST adaptation, $n = 1024, r = 32$; adapter and residual are fixed. **(a)** First-step A/B-side update contributions D_A, D_B are shown versus c/c_{ref}, relative to the original split; dashed lines are reference slopes $+1/-1$, not additional measurements. **(b)** Each scale receives its own learning-rate sweep; stars mark sampled minima and lower loss is better. The same starting model can have different optimization responses. Bars and bands show three-seed SD.

B.4 FEASIBILITY AT THE BALANCED SCALE

Feasibility is restricted to $(cB_0, c^{-1}A_0)$ for a fixed base factorization. General invertible changes $(B_0D, D^{-1}A_0)$ can alter factor geometry and sign aggregation even with the same product.

Corollary 10 (Two-sided feasibility under reciprocal scalar rescaling)**.** *For a fixed base factorization, assume the uniform two-sided orders in Theorem 2, $K_A, K_B > 0$, and $0 < \rho_{\min} \le \rho \le \rho_{\max} < \infty$, independently of width and rank. Define*

$$\Lambda_{\text{lin}} := s\sqrt{K_A K_B}, \qquad \Lambda_{\text{quad}} := \sqrt{sK_{AB}}. \tag{31}$$

After Equation (41)*,*

$$D_A, D_B = \Theta(\bar{\eta}\Lambda_{\text{lin}}), \qquad D_{AB} = \Theta((\bar{\eta}\Lambda_{\text{quad}})^2). \tag{32}$$

Thus the two linear contributions are $\Theta(1)$ when $\bar{\eta} = \Theta(\Lambda_{\text{lin}}^{-1})$, and the interaction is then $O(1)$ if and only if

$$\Lambda_{\text{quad}} = O(\Lambda_{\text{lin}}). \tag{33}$$

We prove Corollary 10 within the reciprocal scalar family of the fixed base factors.

Proof. Substitute $\eta_A = \bar{\eta}/\sqrt{\rho}$, $\eta_B = \bar{\eta}\sqrt{\rho}$, and Equation (41) into Equation (5). The interaction becomes $\Theta((\Lambda_{\text{quad}}/\Lambda_{\text{lin}})^2)$ at the non-degenerate linear learning-rate scale. □

C INITIAL SCALING AND FIXED-HORIZON DYNAMICS

C.1 GENERAL COEFFICIENT FORM

Fix a base factorization (A_0, B_0) and learning rates η_A, η_B. For the initial update, define K_A, K_B, and K_{AB} by separating the multiplier and learning rates from the remaining coefficients:

$$D_A(1) = \Theta(s\eta_A K_A), \qquad D_B(1) = \Theta(s\eta_B K_B), \qquad D_{AB}(1) = \Theta(s\eta_A\eta_B K_{AB}). \tag{34}$$

At initialization, reciprocal scaling multiplies the old B factor by c and the old A feature by c^{-1}. Positive scaling of a raw gradient leaves its SignSGD update unchanged. Hence

$$K_A(c) = cK_A(1), \qquad K_B(c) = c^{-1}K_B(1), \tag{35}$$

$$D_A(c) = \Theta(s\eta_A cK_A), \qquad D_B(c) = \Theta(s\eta_B c^{-1}K_B), \qquad D_{AB}(c) = \Theta(s\eta_A\eta_B K_{AB}). \tag{36}$$

Together with the relevant aggregation conditions, this proves the initial asymptotic statement in Theorem 2. The exact identities are proved in Appendix B.2.

C.2 Fixed-horizon scale recursions

The multi-step analysis follows the scale recursions of LoRA+, Beyond Zero Initialization, and μA (Hayou et al., 2024b; Li et al., 2025; Chen et al., 2026). Fix the layer input and a finite horizon T independent of width and rank. Under the normalized-update and no-cancellation assumptions, let b_0 and z_0 denote the RMS scales of B_0 and A_0x. For nondegenerate updates and fixed $1 \le t \le T$, the recursions give

$$\mathrm{RMS}(B_t(c)) = \Theta(\max\{cb_0, \eta_B\}), \tag{37}$$

$$\mathrm{RMS}(A_t(c)x) = \Theta\big(\max\{c^{-1}z_0, \eta_A d_{\mathrm{in}}\}\big). \tag{38}$$

The second relation uses the processed-gradient alignment $\mathrm{RMS}(g_A^t x) = \Theta(d_{\mathrm{in}})$ for the single-sample normalized update. Each recurrence compares the initial contribution with finitely many increments; the largest order determines the scale when systematic cancellation is excluded.

These relations do not require the trained factors to remain reciprocal copies. Explicit c or c^{-1} dependence persists only in the corresponding initialization-dominated branch. In an update-dominated branch, the learning rate enters the leading factor or feature scale. Later coefficients therefore cannot generally be treated as independent of both learning rates. Appendix D derives the random dual-nonzero branch and checks its self-consistency; structured aggregation assumptions must hold at any later state to which they are applied.

C.3 Overall scale and relative allocation

Multiplying and dividing the initial linear terms gives

$$M := \sqrt{D_A D_B} = \Theta\Big(s\sqrt{\eta_A\eta_B}\sqrt{K_A K_B}\Big), \tag{39}$$

$$Q := \frac{D_A}{D_B} = \Theta\left(c^2\frac{\eta_A}{\eta_B}\frac{K_A}{K_B}\right). \tag{40}$$

$$c_\star^2 = \Theta\left(\rho\frac{K_B}{K_A}\right). \tag{41}$$

The reciprocal scale changes the ratio but cancels from the geometric mean. The latter is not the combined RMS: the linear sum has squared RMS $M^2(Q + Q^{-1} + 2\gamma_{AB})$, where γ_{AB} is its cosine similarity; the full sum also contains the interaction.

C.4 Symmetric finite-scale allocation

At finite width, suppose the ideal scalar dependence is $D_A(c) = ac$ and $D_B(c) = bc^{-1}$ for constants $a, b > 0$.

Lemma 11 (Symmetric minimax representative)**.** *The unique positive scale $c_{\mathrm{sym}} = \sqrt{b/a}$ equalizes the two terms and minimizes both $\max\{D_A(c), D_B(c)\}$ and $D_A(c) + D_B(c)$ over $c > 0$.*

Proof. The equation $ac = b/c$ has the unique positive solution $c = \sqrt{b/a}$. Below this point, b/c is the larger term and decreases as c increases; above it, ac is the larger term and increases. The crossing therefore minimizes the maximum. For the sum,

$$ac + \frac{b}{c} \ge 2\sqrt{ab}, \tag{42}$$

with equality only at the same scale. □

The lemma selects a parameter-free representative of the asymptotic order class. It does not claim that exact equality minimizes downstream loss. That empirical question is evaluated with complete learning-rate curves.

D Random Dual-Nonzero Initialization

D.1 Conditional joint width–rank branch

Let A_0, B_0 have independent centered entries with bounded moments and variances σ_A^2, σ_B^2. Take the fan-in convention $\sigma_A^2 = \Theta(n^{-1})$, giving $Z_A^0 = A_0 x = \Theta(1)$. Under the signal, no-cancellation and random-aggregation assumptions, fixed-horizon post-step orders are

$$Z_A^t = \max\{\Theta(Z_A^0), \Theta(\eta_A n)\}, \tag{43}$$
$$B_t = \max\{\Theta(B_0), \Theta(\eta_B)\}. \tag{44}$$

The old factors entering an update have these post-step orders only for $2 \le t \le T$. Consequently,

$$D_A^t = \Theta(s\eta_A n\sqrt{r}) \max\{\sigma_B, \eta_B\}, \tag{45}$$
$$D_B^t = \Theta(s\eta_B r) \max\{1, \eta_A n\}, \tag{46}$$
$$D_{AB}^t = \Theta(s\eta_A\eta_B n\sqrt{r}). \tag{47}$$

On the branch $Z_A^t = \Theta(\eta_A n)$, $B_t = \Theta(\sigma_B)$, $D_B^t = \Theta(1)$ requires $\eta_A\eta_B = \Theta((snr)^{-1})$. Writing $\rho = \eta_B/\eta_A$ determines the two rates; substituting into $D_A^t = \Theta(1)$ determines σ_B. The resulting family is

$$\sigma_B^2 = \Theta\Big(\frac{\rho}{sn}\Big), \qquad \eta_A = \Theta\Big((snr\rho)^{-1/2}\Big), \qquad \eta_B = \Theta\bigg(\Big(\frac{\rho}{snr}\Big)^{1/2}\bigg), \tag{48}$$

$$\|A_t x\|_{\mathrm{rms}} = \Theta\bigg(\sqrt{\frac{n}{sr\rho}}\bigg), \qquad D_{AB} = \Theta(r^{-1/2}). \tag{49}$$

These expressions follow directly by substitution into the three path orders. Self-consistency requires $\sigma_B = \Omega(\eta_B)$, whose ratio is $\Theta(\sqrt{r})$, and $\eta_A n \gtrsim 1$, equivalently

$$\rho = O(n/(sr)). \tag{50}$$

The latter ratio may grow with width; the bounded-ρ assumption of Corollary 10 is confined to that first-step corollary. With $\rho = 1$, equal rates are $\Theta((snr)^{-1/2})$, $\sigma_B^2 = \Theta((sn)^{-1})$ and $Z_A^t = \Theta(\sqrt{n/(sr)})$, recovering the fixed-rank, $s = 1$ regime of Li et al. (2025). Order-one intermediate features instead require $\rho = \Theta(n/(sr))$, giving $(\eta_A, \eta_B) = \Theta(n^{-1}, (sr)^{-1})$ and $\sigma_B^2 = \Theta((s^2 r)^{-1})$, consistent with the factor-rate exponent of Hayou et al. (2024b).

Thus when $sr = o(n)$, a common rate, two nondegenerate linear contributions at $t \ge 2$, and an order-one intermediate feature cannot all hold on this branch. This is not a global impossibility result excluding other dominance branches.

D.2 The first step and its distinct feasibility boundary

Before any accumulated update, the paths instead have orders

$$D_A^1 = \Theta(s\eta_A n\sqrt{r}\,\sigma_B), \qquad D_B^1 = \Theta(s\eta_B r), \qquad D_{AB}^1 = \Theta(s\eta_A\eta_B n\sqrt{r}). \tag{51}$$

The family above gives $D_A^1 = \Theta(1)$, $D_B^1 = \Theta(\sqrt{sr\rho/n})$ and $D_{AB}^1 = \Theta(r^{-1/2})$. If $sr\rho = o(n)$, the first B contribution vanishes; extending this branch to the first step requires $sr\rho = \Theta(n)$.

For the equal-variance reference $s = 1$, $\sigma_A^2 = \sigma_B^2 = \Theta(n^{-1})$, the coefficients are $K_A = \Theta(\sqrt{nr})$, $K_B = \Theta(r)$ and $K_{AB} = \Theta(n\sqrt{r})$. Hence $\Lambda_{\mathrm{lin}} = \Theta(n^{1/4} r^{3/4})$ and $\Lambda_{\mathrm{quad}}/\Lambda_{\mathrm{lin}} = \Theta((n/r^2)^{1/4})$. At the calibrated linear rate, $D_{AB}^1 = \Theta(\sqrt{n}/r)$: the simultaneous first-step criterion is infeasible for $r = o(\sqrt{n})$, but compatible with $r = \Omega(\sqrt{n})$ under the assumptions. This is distinct from degeneration of the uncalibrated first B contribution.

D.3 Residual compensation and the experimental Kaiming baseline

The frozen residual $R = W_0 - sB_0A_0$ preserves the initial function, as already used by Li et al. (2025). Its equal-variance theoretical reference differs from independent fan-in Kaiming-uniform factors with $a = \sqrt{5}$: $A_{ij} \sim \mathcal{U}[-d_{\text{in}}^{-1/2}, d_{\text{in}}^{-1/2}]$ and $B_{ij} \sim \mathcal{U}[-r^{-1/2}, r^{-1/2}]$. For square layers their variances are $1/(3n)$ and $1/(3r)$, not equal.

$$K_A = \Theta(n), \qquad K_B = \Theta(r), \qquad K_{AB} = \Theta(n\sqrt{r}). \tag{52}$$

At $s = 1$, $c_\star = \Theta(\sqrt{r/n})$ and common rate $\Theta((nr)^{-1/2})$ make both first-step linear contributions order one, while $D^1_{AB} = \Theta(r^{-1/2})$ is bounded. This conditional random-family result is neither the structured spectral rule nor a later-step AdamW guarantee.

E Structured Initializations with a Fixed Product

E.1 General rectangular-layer derivation

Let an upstream construction use information $\mathcal{S}$—for example the pretrained weight or an activation covariance—to produce a selected low-rank matrix P_r, a frozen residual R_r, and factors $\bar{A}_0, \bar{B}_0$:

$$\mathcal{I}_r(W_0, \mathcal{S}) := (P_r, R_r, \bar{B}_0, \bar{A}_0), \qquad P_r = s\bar{B}_0\bar{A}_0, \qquad W_0 = R_r + P_r. \tag{53}$$

Fixing the selected base factors, we study the positive-scalar family $B_0(c) = c\bar{B}_0$ and $A_0(c) = c^{-1}\bar{A}_0$, $c > 0$.

For a single sample, the first coordinate-normalized update gives

$$\delta_A = -s\eta_A \|x\|_1 B_0 \operatorname{sign}(B_0^\top g), \tag{54}$$

$$\delta_B = -s\eta_B \|A_0 x\|_1 \operatorname{sign}(g). \tag{55}$$

Under assumption 7, $\|x\|_1 = \Theta(d_{\text{in}})$, $\|A_0x\|_1 = \Theta(\sqrt{rS_A^{\text{act}}})$, and the output-coordinate RMS of $B_0 \operatorname{sign}(B_0^\top g)$ is $\Theta(\sqrt{S_B/d_{\text{out}}})$. Hence

$$D_A(c) = \Theta\left(s\eta_A c\, d_{\text{in}} \sqrt{\frac{S_B}{d_{\text{out}}}}\right), \tag{56}$$

$$D_B(c) = \Theta\left(s\eta_B c^{-1} \sqrt{rS_A^{\text{act}}}\right). \tag{57}$$

$$K_A = d_{\text{in}}\sqrt{S_B/d_{\text{out}}}, \qquad K_B = \sqrt{rS_A^{\text{act}}}, \qquad K_{AB} = d_{\text{in}}\chi_r, \tag{58}$$

With $\rho = \eta_B/\eta_A$, equating their orders gives

$$\boxed{c_\star^4 = \Theta\left[\rho^2 \frac{r\, d_{\text{out}} S_A^{\text{act}}}{d_{\text{in}}^2 S_B}\right].} \tag{59}$$

The common-learning-rate case uses $\rho = 1$. The activation moment can be written as

$$S_A^{\text{act}} = \mathbb{E}_x \|A_0x\|_2^2 = \operatorname{tr}(A_0 C_x A_0^\top), \qquad C_x := \mathbb{E}[xx^\top]. \tag{60}$$

The data-free starting point replaces C_x by the identity and therefore uses $S_A^{\text{act}} \approx \|A_0\|_F^2$.

The interaction requires an additional aggregation condition. For a single sample under SignSGD,

$$\delta_{AB} = s\eta_A\eta_B \|x\|_1 \left[\operatorname{sign}(A_0x)^\top \operatorname{sign}(B_0^\top g)\right] \operatorname{sign}(g). \tag{61}$$

Thus K_{AB} has scale $d_{\text{in}}\chi_r$, with χ_r defined in Equation (16). The specialization $K_{AB} = d_{\text{in}}\sqrt{r}$ uses that condition. It is not implied by the two factor energies or by product preservation. The reciprocal first-step identity still holds for any fixed sign alignment.

Effective-update feasibility. Combining these coefficients with Corollary 10 gives

$$\left(\frac{\Lambda_{\mathrm{quad}}}{\Lambda_{\mathrm{lin}}}\right)^4 = \Theta\left(\frac{d_{\mathrm{out}}\chi_r^2}{s^2 r S_A^{\mathrm{act}} S_B}\right). \tag{62}$$

Under $\chi_r = \Theta(\sqrt{r})$, simultaneous non-degenerate linear terms and bounded interaction require $s^2 S_A^{\mathrm{act}} S_B = \Omega(d_{\mathrm{out}})$. This constraint is unchanged by reciprocal scalar rescaling of these factors. For the square-root spectral factors below, writing $T_r = \mathrm{tr}(\Sigma_I)$ and $A_r = \mathrm{tr}(\Sigma_I V_I^\top C_x V_I)$, it becomes $A_r T_r = \Omega(d_{\mathrm{out}})$. For a weak selected spectrum, balance alone therefore does not establish effective first updates within this family.

E.2 Initializers based on selected singular components

PiSSA and MiLoRA initialize from principal and minor singular components, respectively (Meng et al., 2024; Wang et al., 2025a). Their spectral choices correspond to different index sets I in the following matrix and square-root factorization:

$$P_r = U_I \Sigma_I V_I^\top, \tag{63}$$

$$\bar{B}_0 = s^{-1/2} U_I \Sigma_I^{1/2}, \qquad \bar{A}_0 = s^{-1/2} \Sigma_I^{1/2} V_I^\top, \tag{64}$$

so $s\bar{B}_0\bar{A}_0 = P_r$. Since U_I has orthonormal columns,

$$S_B = s^{-1}\,\mathrm{tr}(\Sigma_I), \tag{65}$$

$$S_A^{\mathrm{act}} = s^{-1}\,\mathrm{tr}(\Sigma_I V_I^\top C_x V_I). \tag{66}$$

Substitution gives

$$\boxed{c_\star^4 = \Theta\left[\rho^2 \frac{r\, d_{\mathrm{out}}}{d_{\mathrm{in}}^2} \frac{\mathrm{tr}(\Sigma_I V_I^\top C_x V_I)}{\mathrm{tr}(\Sigma_I)}\right].} \tag{67}$$

For a rectangular layer, common learning rate, isotropic input covariance, and the rank-coordinate non-concentration condition above, this gives $c_\star^4 = \Theta(r d_{\mathrm{out}}/d_{\mathrm{in}}^2)$. The finite order-based reference $c_{\mathrm{ref}} := (r d_{\mathrm{out}}/d_{\mathrm{in}}^2)^{1/4}$ sets the unspecified order constant to one; a square layer recovers $(r/n)^{1/4}$. The next result retains spectral concentration and finite-dimensional constants instead of imposing that condition.

Retaining the selected spectral structure. Fixing P_r and R_r retains which pretrained components initialize the trainable branch and which remain frozen. Reciprocal scalar rescaling preserves the selected singular values and directions, including the directions of the individual factors. MiLoRA explicitly uses initialization to guide optimization rather than imposing a persistent orthogonality constraint during training (Wang et al., 2025a). This motivates preserving the initial structural choice while studying its subsequent use by the optimizer.

Reciprocal scaling preserves P_r, its singular values, its row and column spaces, and the residual decomposition. It does, however, change the factor Gram matrices:

$$B_0(c)^\top B_0(c) = c^2 B_0^\top B_0, \qquad A_0(c) A_0(c)^\top = c^{-2} A_0 A_0^\top. \tag{68}$$

Gram balance and update balance. The square-root split in Equation (64) already satisfies $\bar{B}_0^\top \bar{B}_0 = \bar{A}_0 \bar{A}_0^\top = s^{-1}\Sigma_I$. This Gram balance need not equalize D_A and D_B, whose scales also depend on input statistics, dimensions, and the update rule. BaLoRA studies Gram balancing through the conditioning of loss minimizers (Castin et al., 2026); our calibration instead targets the initial output contributions. Thus the two balance criteria address different optimization questions even for the same selected matrix.

E.3 Exact Gaussian spectral coefficients for rectangular layers

Proposition 12 (Population allocation with a selected spectrum)**.** *Let $x \sim \mathcal{N}(0, I_{d_{\mathrm{in}}})$ and let $A_0 \in \mathbb{R}^{r\times d_{\mathrm{in}}}$ and $B_0 \in \mathbb{R}^{d_{\mathrm{out}}\times r}$ be the square-root factors in Equation* (64)*, with $r \le \min\{d_{\mathrm{in}}, d_{\mathrm{out}}\}$ positive selected singular values and column-orthonormal U_I, V_I. Consider fresh, single-sample*

SignSGD steps; assume the coordinates of g and $B_0^\top g$ are nonzero almost surely. Define output-coordinate RMS by $D_j^2 = \mathbb{E}_{x,g}\|\delta_j\|_2^2/d_{\text{out}}$, $T_r = \sum_{i\in I}\sigma_i$, and $a = 2/\pi$. At the unscaled base factors, write $D_A = s\eta_A K_A$ and $D_B = s\eta_B K_B$. Their exact coefficients satisfy

$$K_A^2 = \frac{T_r}{s}\frac{d_{\text{in}}}{d_{\text{out}}}[1 + a(d_{\text{in}} - 1)], \tag{69}$$

$$K_B^2 = \frac{1}{s}\left[(1-a)T_r + a\left(\sum_{i\in I}\sqrt{\sigma_i}\right)^2\right]. \tag{70}$$

Consequently, under $B_0(c) = cB_0$, $A_0(c) = c^{-1}A_0$, the population-balancing scale is

$$c_{\text{pop}}^4 = \rho^2\frac{d_{\text{out}}}{d_{\text{in}}}\frac{(1-a)T_r + a(\sum_{i\in I}\sqrt{\sigma_i})^2}{T_r[1 + a(d_{\text{in}} - 1)]}. \tag{71}$$

Writing $q_\sigma = (\sum_{i\in I}\sqrt{\sigma_i})^2/T_r \in [1, r]$, this gives $c_{\text{pop}}^4 = \Theta(\rho^2 d_{\text{out}} q_\sigma/d_{\text{in}}^2)$. For $d_{\text{in}} = d_{\text{out}} = n$, the exact expression reduces to the square-layer formula and its order in Equation (74)*.*

Proof. The single-sample identities above give $D_A^2 = (s\eta_A)^2\mathbb{E}\|x\|_1^2 T_r/(sd_{\text{out}})$, since orthogonality of U_I yields $\|B_0\,\text{sign}(B_0^\top g)\|_2^2 = T_r/s$. For independent standard normal coordinates,

$$\mathbb{E}\|x\|_1^2 = d_{\text{in}} + ad_{\text{in}}(d_{\text{in}} - 1). \tag{72}$$

Also $A_0 x$ has independent Gaussian coordinates of variances σ_i/s, because V_I has orthonormal columns. Expanding its squared ℓ_1 norm gives

$$\mathbb{E}\|A_0 x\|_1^2 = \frac{1}{s}\left[\sum_i \sigma_i + a\sum_{i\neq j}\sqrt{\sigma_i\sigma_j}\right]. \tag{73}$$

Since $\|\,\text{sign}(g)\|_2^2 = d_{\text{out}}$, this cancels the output-RMS denominator for D_B and proves the two coefficients. Solving $\eta_A c K_A = \eta_B c^{-1} K_B$ gives Equation (71). Finally $T_r \le (\sum_i\sqrt{\sigma_i})^2 \le rT_r$, and $0 < a < 1$ gives the stated order. No independence between x and g is needed for these two coefficients beyond the specified Gaussian marginal for x and the nonzero-coordinate assumptions. □

For a flat spectrum, the finite expression is $c_{\text{pop}}^4 = \rho^2(d_{\text{out}}/d_{\text{in}})[1 + a(r-1)]/[1 + a(d_{\text{in}} - 1)]$. At $\rho = 1$ its order is $(rd_{\text{out}}/d_{\text{in}}^2)^{1/4}$ for c_{pop}, reducing to $(r/n)^{1/4}$ only for square layers. If q_σ remains bounded, the rank factor is absent and the order becomes $(d_{\text{out}}/d_{\text{in}}^2)^{1/4}$. The LoRA multiplier s cancels from the balancing scale, although not from the individual coefficients.

This proposition concerns population RMS of fresh single-sample SignSGD updates, not normalization of a batch-averaged gradient or an exact AdamW formula. It does not assume weak interaction signs: allocation and interaction feasibility are distinct. Appendix I checks the finite rectangular formula without fitting its constants to measured update ratios.

$$c_\star^4 = \Theta\left(\frac{d_{\text{out}}(\sum_{i\in I}\sqrt{\sigma_i})^2}{d_{\text{in}}^2\sum_{i\in I}\sigma_i}\right). \tag{74}$$

E.4 Context-oriented decomposition

A context-oriented initializer uses an activation covariance to transform the right-side directions before selecting components (Yang et al., 2024). Our core comparison uses instruction-previewed mode (CorDA-IPM), not a claim covering every mode. Abstractly, write

$$P_r = U_I\Sigma_I\widetilde{V}_I^\top, \tag{75}$$

$$\bar{B}_0 = s^{-1/2}U_I\Sigma_I^{1/2}, \qquad \bar{A}_0 = s^{-1/2}\Sigma_I^{1/2}\widetilde{V}_I^\top. \tag{76}$$

For a fixed covariance construction and index set I, reciprocal scalar rescaling retains this context-selected matrix, its transformed directions, and the frozen residual; it does not reselect components.

Algorithm 1 Scale calibration with the initial structure fixed

Require: Function-preserving (R, A_0, B_0), calibration batch, AdamW configuration, starting scales $c_{0,\ell} > 0$, correction count K

1: Apply $B_\ell \leftarrow c_{0,\ell} B_\ell$, $A_\ell \leftarrow c_{0,\ell}^{-1} A_\ell$
2: **for** $k = 1, \ldots, K$ **do**
3: Save factors and all state touched by the virtual step
4: Simulate one AdamW step; measure $D_{A,\ell}, D_{B,\ell}, D_{AB,\ell}$
5: Restore parameters, gradients, optimizer, scheduler, scaler, and RNG states
6: Apply Equation (10) to resolved channels; flag others; verify product and forward invariance
7: **end for**
8: **return** calibrated factors; start formal training with a fresh optimizer

In the implementation based on a decomposition of $W_0 C$, the transformed directions satisfy $\widetilde{V}_I^\top = V_I^\top C^{-1}$ after covariance regularization. The resulting moments and factor scale are

$$S_B = s^{-1} \operatorname{tr}(\Sigma_I), \tag{77}$$

$$S_A^{\mathrm{act}} = s^{-1} \operatorname{tr}(\Sigma_I \widetilde{V}_I^\top C_x \widetilde{V}_I), \tag{78}$$

$$c_\star^4 = \Theta\left[\rho^2 \frac{r\, d_{\mathrm{out}}}{d_{\mathrm{in}}^2} \frac{\operatorname{tr}(\Sigma_I \widetilde{V}_I^\top C_x \widetilde{V}_I)}{\operatorname{tr}(\Sigma_I)}\right]. \tag{79}$$

A complete theorem for this construction additionally requires bounded conditioning of the context transform, an explicit covariance regularization rule, and rank-sum control for the transformed directions. An empirical check must record these quantities and use the AdamW probe to assess their finite-width effect.

E.5 Initializers with additional factor-level constraints

Distinct initialization objectives. Activation- and quantization-based constructions provide further context. EVA selects initialization directions from activation variance and can redistribute the rank budget (Paischer et al., 2024); LoftQ jointly chooses quantized weights and a low-rank initialization to reduce quantization discrepancy (Li et al., 2024). These are different design objectives, not additional evaluated baselines or automatic corollaries of our scalar-rescaling analysis.

Orthogonal or QR-based initialization. OLoRA obtains orthonormal initialization through QR decomposition (Büyükakyüz, 2024). Reciprocal scaling preserves the initialized product and the exact first-step scalar law. It does not preserve a particular orthonormal or QR normalization assigned to the factors. Applying it to this extension therefore requires reporting both product invariance and the change in factor normalization.

Gradient-informed initialization. LoRA-One initializes both factors from a single full-gradient SVD with normalized singular values (Zhang et al., 2025b). LoRA-GA instead chooses factors to approximate the full-model gradient through the first low-rank update (Wang et al., 2024a). Its product update contains factor-level terms of the form

$$BB^\top G \qquad \text{and} \qquad GA^\top A. \tag{80}$$

Reciprocal scaling maps them to $c^2 BB^\top G$ and $c^{-2} GA^\top A$. The product and tangent-space set remain fixed, but the two terms in the original gradient-approximation objective do not. This is a boundary extension requiring gradient cosine alignment and update-norm-ratio checks before and after calibration, rather than a case where the entire upstream objective is invariant.

F Calibration and Optimizer Equivalence

State and measurements. Each virtual step uses the training loss mask, precision, clipping, decay and optimizer configuration. Snapshot and restore all touched parameters, gradients, moments/counters, parameter-group metadata, scheduler, scaler and CPU/accelerator RNG states. Corrections occur outside the optimizer; formal training starts with fresh moments. Output paths use the

pre-step activations and the exact factor increments in Equation (2); signed-Gram accumulation is specified once in Appendix J.2.

Multiple calibration batches.

$$q_\ell = \left[\frac{B^{-1} \sum_{b=1}^{B} D_{B,\ell,b}^2}{B^{-1} \sum_{b=1}^{B} D_{A,\ell,b}^2} \right]^{1/4}. \tag{81}$$

This convention equally weights batch-level squared RMS and reduces to $\sqrt{D_B/D_A}$ for one batch. It is an exact target under the ideal reciprocal law; finite AdamW is remeasured. Token-pooled diagnostics instead use their explicit coordinate weights.

F.1 Numerical safeguards

Factor norms, reciprocal scales, and invariant checks are evaluated in FP32 even when the base model is BF16. The protocol rejects non-finite contributions. Both channels, in the registered normalization, must exceed a predeclared positive detectability threshold ε_D before applying

$$q_\ell = \exp\{\tfrac{1}{2} [\log D_{B,\ell} - \log D_{A,\ell}]\} \tag{82}$$

If either channel is at or below its detectability threshold, set $q_\ell = 1$, retain that layer's factors, and record the unresolved channel. Such layers are not counted as successfully balanced. Adding a floor to two vanishing channels would give a misleading ratio near one. For the focused strong controls, detection uses $D_{j,\ell}/\mathrm{RMS}(W_{\mathrm{pre},\ell} X_\ell)$ and $\varepsilon_D = 10^{-12}$, requiring a positive finite denominator; the shared normalization cancels from q_ℓ. No scale cap is applied. Non-finite or unresolved cases remain flagged, with counts in Table 25. A capped variant would not inherit the exact equalization claim.

After every correction, the product error is checked through

$$\epsilon_{P,\ell} := \frac{\|B_\ell^{\mathrm{new}} A_\ell^{\mathrm{new}} - B_\ell^{\mathrm{old}} A_\ell^{\mathrm{old}}\|_F}{\max(\|B_\ell^{\mathrm{old}} A_\ell^{\mathrm{old}}\|_F, 10^{-12})}, \tag{83}$$

together with the relative fixed-batch forward error, normalized by the pre-correction output norm. The original breadth protocol uses 10^{-6} for the FP32 product check and 10^{-4} for the BF16 forward path. The separately registered focused strong-control protocol uses 10^{-5} for both relative checks (Appendix K.3); these tolerances are not retroactively substituted into the breadth study.

F.2 One correction and its empirical variants

$$c_0 \sqrt{D_B(c_0)/D_A(c_0)} = \sqrt{D_B(1)/D_A(1)}. \tag{84}$$

Corollary 13 (One correction removes the starting scale)**.** *Under Equation* (6)*, fix the batch and learning rates, with $D_A(1), D_B(1) > 0$. From any $c_0 > 0$, the correction $q(c_0) = \sqrt{D_B(c_0)/D_A(c_0)}$ yields Equation* (84)*. The corrected scale is independent of c_0 and equalizes the linear terms; a repeated identical probe gives $q = 1$.*

Proof. Substitute $D_A(c_0) = c_0 D_A(1)$ and $D_B(c_0) = c_0^{-1} D_B(1)$ into $c_0 q(c_0)$. At the resulting scale $\sqrt{D_B(1)/D_A(1)}$, both contributions equal $\sqrt{D_A(1) D_B(1)}$. □

The identity needs neither aggregation assumptions nor a closed-form start. Finite AdamW, unresolved channels, and capped corrections need not satisfy it.

Direct uses $(c_{0,\ell}, K) = (1, 1)$, Direct-2 and random Init[AB] use $(1, 2)$, and Warm uses the coarse start with two corrections. The variants are never relabeled when reporting results. A post-correction diagnostic is not another correction and is charged separately.

F.3 Same-probe factor-specific learning-rate control

Fix the raw factors, calibration batch, optimizer state, and a common probe learning rate η. Suppose the realized increments depend linearly on each factor's learning rate. Define $q_\ell = \sqrt{D_{B,\ell}/D_{A,\ell}}$ from resolved positive channels and keep the raw factors unchanged. The same probe specifies

$$\rho_\ell := D_{A,\ell}/D_{B,\ell}, \qquad \eta_{A,\ell} := \eta/\sqrt{\rho_\ell} = \eta q_\ell, \qquad \eta_{B,\ell} := \eta\sqrt{\rho_\ell} = \eta/q_\ell. \tag{85}$$

Proposition 14 (Same-probe first-step allocation)**.** *Under the fixed-state, learning-rate-linear assumption above, Equation* (85) *gives*

$$\delta'_{A,\ell} = q_\ell \delta_{A,\ell}, \qquad \delta'_{B,\ell} = q_\ell^{-1} \delta_{B,\ell}, \qquad \delta'_{AB,\ell} = \delta_{AB,\ell}. \tag{86}$$

Thus $D'_{A,\ell} = D'_{B,\ell} = \sqrt{D_{A,\ell} D_{B,\ell}}$*, while* M_ℓ *and the interaction vector are unchanged. If the normalized update also satisfies Equation* (25)*, this matches the entire first-step product increment after reciprocal factor rescaling by* q_ℓ*.*

Proof. Learning-rate linearity gives $\Delta A'_\ell = q_\ell \Delta A_\ell$ and $\Delta B'_\ell = q_\ell^{-1} \Delta B_\ell$. Substitution into the three update paths proves Equation (86). For the ideal reciprocal comparison, suppress layer indices and write $q = q_\ell$. The rescaled factors retain the original increments, and

$$(qB + \Delta B)(q^{-1}A + \Delta A) = (B + q^{-1}\Delta B)(A + q\Delta A). \tag{87}$$

Subtracting BA and multiplying by s proves full product-increment equality, including $\Delta B \Delta A$. □

Proposition 15 (Reciprocal trajectory equivalence)**.** *Fix* $q_\ell > 0$ *per layer. Assume the loss depends on each factor pair only through its product, with identical batches and stochastic realizations, exact arithmetic, and no clipping or decay. Compare calibrated factors* $(A_0/q, qB_0)$ *with common rate* η_t *to raw factors* (A_0, B_0) *with rates* $(q\eta_t, \eta_t/q)$*. For SignSGD, or Adam with* $\epsilon = 0$*, zero initial moments, and defined normalized updates, every step satisfies*

$$A_t^{\mathrm{LR}} = qA_t^{\mathrm{Cal}}, \quad B_t^{\mathrm{LR}} = q^{-1}B_t^{\mathrm{Cal}}, \quad B_t^{\mathrm{LR}}A_t^{\mathrm{LR}} = B_t^{\mathrm{Cal}}A_t^{\mathrm{Cal}}. \tag{88}$$

The statements apply layerwise; corresponding nonzero initial moments can replace zero moments.

Proof. The factor relation holds initially. Equal products and matched randomness give the same loss signals, hence $g_A^{\mathrm{Cal}} = q g_A^{\mathrm{LR}}$ and $g_B^{\mathrm{Cal}} = q^{-1} g_B^{\mathrm{LR}}$. Sign updates are unchanged by these positive scales. For Adam, induction on the moment recursions gives

$$m_A^{\mathrm{Cal}} = q m_A^{\mathrm{LR}}, \qquad v_A^{\mathrm{Cal}} = q^2 v_A^{\mathrm{LR}}, \tag{89}$$

$$m_B^{\mathrm{Cal}} = q^{-1} m_B^{\mathrm{LR}}, \qquad v_B^{\mathrm{Cal}} = q^{-2} v_B^{\mathrm{LR}}. \tag{90}$$

Bias correction preserves these identities. With $\epsilon = 0$, normalized directions are equal, and the stated rates preserve the factor relation after the next update. Induction proves the claim. □

Finite AdamW distinction. Learning-rate linearity holds for a fixed exact-arithmetic AdamW step when the learning rate only multiplies the increment, including decoupled weight decay. Thus same-probe learning-rate allocation can equalize that measured first step even when factor rescaling does not. For fresh moments, no clipping, and common decay λ, the rescaled-factor increments are, coordinatewise,

$$\Delta A(c) = -\eta \left[\frac{g_A}{|g_A| + \epsilon/c} + \lambda c^{-1} A \right], \tag{91}$$

$$\Delta B(c) = -\eta \left[\frac{g_B}{|g_B| + c\epsilon} + \lambda c B \right], \tag{92}$$

where g_A, g_B are the raw-factor gradients. Nonzero ϵ and decay generally break the comparison when those hyperparameters remain identical. Moment accumulation alone does not: Proposition 15 includes momentum. Global clipping can introduce a time-varying relative gradient scale between the two parameterizations; SignSGD is invariant to a positive clipping multiplier, whereas Adam need not be invariant to its history.

Without clipping, the AdamW correspondence can be restored by assigning the LR-side parameters

$$(\epsilon_A^{\mathrm{LR}}, \epsilon_B^{\mathrm{LR}}) = (\epsilon/q, q\epsilon), \qquad (\lambda_A^{\mathrm{LR}}, \lambda_B^{\mathrm{LR}}) = (\lambda/q, q\lambda). \tag{93}$$

Indeed the normalized denominators and rate–decay products then transform as required. This is a diagnostic identity, not the unchanged-hyperparameter control used in the comparisons. Appendix I checks the exact and finite cases separately.

Complexity. Reciprocal rescaling costs $O(\sum_\ell r_\ell(d_{\text{in},\ell} + d_{\text{out},\ell}))$. A correction requires one shared forward/backward optimizer probe per calibration batch, not one pass per layer. It adds no parameters or persistent training operations; diagnostic work and complete-training candidate counts are accounted for separately.

F.4 Starting-scale variants

Warm: an empirical enhancement. For structured factors satisfying $R + sB_0A_0 = W_0$, Warm uses the data-free starting scale

$$c_{\text{cf}} := \left[\rho^2 \frac{r\, d_{\text{out}} \|A_0\|_F^2}{d_{\text{in}}^2 \|B_0\|_F^2}\right]^{1/4}, \tag{94}$$

$$B_0 \leftarrow c_{\text{cf}} B_0, \qquad A_0 \leftarrow c_{\text{cf}}^{-1} A_0. \tag{95}$$

We use $\rho = 1$ and $S_A^{\text{act}} \approx \|A_0\|_F^2$. Equal-norm square-root factors in $n \times n$ layers give $c_{\text{cf}} = (r/n)^{1/4}$. Unlike Spectrum-full, this coarse rule omits spectral concentration. In Algorithm 1, Direct uses $(c_{0,\ell}, K) = (1, 1)$, Direct-2 uses $(1, 2)$, and Warm uses $(c_{\text{cf},\ell}, 2)$. Analytic-only uses $(c_{\text{cf},\ell}, 0)$ without probes. These controls isolate the start and extra correction.

G Data, Training, and Statistical Protocol

G.1 Data roles and training

Qwen3 (Team, 2025) and Llama-3.2-3B use a fixed revision of Tulu-3 (Lambert et al., 2024). Raw conversations are split with seed 42 before model-specific chat formatting. Training (80,000), selection validation (32,000), reporting NLL (4,096), and diagnostics (512) are pairwise disjoint. Calibration is a training subset, not a fifth independent split. NLL pools assistant-token loss sums and valid-token counts.

Table 3: **Data roles.** Counts are raw conversations; none is a replication count.

Role	Count	Permitted use
Training	80,000	Formal optimizer updates
Selection validation	32,000	Final-step NLL selects rates
Reporting NLL	4,096	Selected checkpoint evaluation only
Calibration	4	Scale selection; contained in training
Diagnostics	512	Independent initialization/update probes

Five nonoverlapping 128-conversation training pools define calibration identities; the default four-conversation subset is shared within seed and across correction rounds. Sensitivity uses nested sizes $1, 2, 4, 8, 16$ and batch counts $1, 2, 4, 8$. Diagnostics use 16 batches of 32; the representative-layer subset fixes four sequences of 256 valid tokens. Frozen model/data revisions and the complete records accompany the source package.

G.2 Initializers, selection, and candidate budgets

The structured downstream matrix contains the following observed configurations: Llama-3.2-3B, Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B: $r \in \{16, 32, 64\}$, CorDA (raw/calibrated), MiLoRA (raw/calibrated), PiSSA (raw/calibrated), Standard LoRA (standard). The random Init[AB] control is reported separately below. Each breadth configuration scans $\{1, 3, 5, 8\} \times 10^{-6}$, $\{1, 3, 5, 8\} \times 10^{-5}$, $\{1, 3, 5, 8\} \times 10^{-4}$ and 10^{-3}. Structured seeds are 42, 2025 and 2026; additional/random seeds are 101, 202 and 303. Each raw/calibrated pair shares factors, residual and minibatch order. Structured calibration uses Warm; random Init[AB] uses two direct corrections. Its independent fan-in Kaiming-uniform factors use $a = \sqrt{5}$, as specified in Appendix D.3, not the no-argument $a = 0$ convention (He et al., 2015; Hu et al., 2022).

Focused configurations are Qwen3-0.6B/PiSSA-32 and Qwen3-1.7B/MiLoRA-64, each with 112 attention projections in 28 blocks. For PiSSA, $(d_{\text{out}}, d_{\text{in}})$ is $(2048, 1024)$ for q, $(1024, 1024)$ for k/v

Table 4: Language-model training configuration.

Item	Setting
Sequence length	1024, dynamic padding, length bucketing
Training budget	80k samples, 1 epoch, 2,500 optimizer steps
Effective batch size	32 samples
Optimizer	AdamW, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$
Weight decay	0.01
Schedule	5% warmup, cosine decay to 0.1 of peak LR
Gradient clipping	Global norm 1.0
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`
LoRA multiplier	`lora_alpha`=r, hence actual $s = 1$
Dropout / bias	0 / none
Precision	BF16 base, FP32 non-zero-adapter target path
Checkpoint selection	Final step only; no early stopping

and $(1024, 2048)$ for o; for MiLoRA it is $(2048, 2048)$ for q/o and $(1024, 2048)$ for k/v. Five paired seeds are $101, 202, 303, 404, 505$. The ordered grids are

$$\mathcal{E} = (3, 5, 8, 13, 20) \times 10^{-5}, \qquad \mathcal{R} = (1/16, 1/4, 1, 4, 16).$$

Five-seed mean validation NLL selects one setting per method, not one setting per seed. Tuned dual LR uses $\eta_A = \eta/\sqrt{\rho}$, $\eta_B = \eta\sqrt{\rho}$. Its five-setting subset uses zero-based grid pairs $(0, 0), (1, 4), (2, 2), (3, 3), (4, 1)$; the 25-grid appends remaining pairs lexicographically. Budget prefixes are fixed before reporting outcomes. Five/25 candidates imply 25/125 candidate–seed trainings; shared candidates count once. Counts do not establish equal wall-clock costs.

Table 5: **Focused variant rules.** Probes are virtual calibration steps per candidate and seed; B counts complete-training candidates per seed. Inherited settings are not a new search.

Method	Information / intervention	Probes	B
Raw	Original factors	0	5
Analytic-only	Coarse energy start	0	5
Spectrum-full	All eligible rectangular/square spectra	0	5
Spectrum-hybrid	Square spectral / rectangular coarse	0	5
Forward-moment	Raw activation squared moments	0	5
Direct / Direct-2	One / two corrections from Raw	1 / 2	5
Warm	Coarse start and two corrections	2	5
Same-probe LR	Direct's response mapped to factor rates	1	5
Total / Layer norm	Match Warm's pooled / module RMS	3	5
Type-shared / shuffle	Reassign Direct's fixed scales	1	5
Tuned dual LR	Independent global ratio grid	0	5 / 25
Warm-mapped LR	Warm's cumulative scale; inherited rate	2	1
Stage decay	Direct/Warm crossed with two decay stages	1 / 2	1

G.3 Reporting and independent replication units

Within each training seed, average the four task-native percentage scores before subtracting paired methods. Report the mean $\bar{d}$, sample SD s_d, and descriptive interval $\bar{d} \pm t_{0.975,n-1} s_d/\sqrt{n}$; $n = 5$ for focused comparisons and $n = 3$ for breadth, with t multipliers 2.776 and 4.303. The groups remain separate. These intervals assume independent approximately normal seed differences, condition on selected settings, omit selection uncertainty and are not multiplicity-adjusted. Layers, tokens, tasks, checkpoints and inference repeats do not enlarge n; a zero-containing interval does not demonstrate equivalence.

External evaluation uses lm-eval 0.4.8 and vLLM 0.8.5 (Kwon et al., 2023): GSM8K 8-shot answer matching (Cobbe et al., 2021), MATH-500 mathematical verification (Hendrycks et al., 2021; Lightman et al., 2024), MMLU-Pro exact multiple-choice accuracy (Wang et al., 2024b), and BBH under the locked zero-shot chain-of-thought harness task (Suzgun et al., 2023). For breadth, five fixed-checkpoint inference repeats (temperature 0.3, top-p 0.95, seeds 2026080501–05) are averaged

within training seed before across-seed statistics. Focused records already have seed-level scores. No test metric selects rates or checkpoints.

The breadth near-optimal window uses validation threshold $L_{\min} + 0.0015$ and the contiguous accepted grid component containing the minimum; disconnected components are not bridged. Its width is $\log_2(\eta_{\max}/\eta_{\min})$, not a generalization metric. Final checkpoints alone are evaluated; there is no early stopping.

G.4 Initialization and controlled adaptation diagnostics

Initialization diagnostics use Qwen3-0.6B, ranks $16, 32, 64$, four initializers and 16 held-out batches. Each batch starts from the same initial parameters and fresh AdamW state, probes at 10^{-4} with training precision/decay/clipping, and restores state. Pool squared RMS within layer before ratios. The ratio D_A/D_B measures allocation, $|\log_2(D_A/D_B)|$ imbalance, and $\sqrt{D_A D_B}$ geometric magnitude; neither is combined RMS. Magnitudes are normalized by the same pretrained-output RMS. Layer medians/IQR are descriptive; across-run SD uses run medians. Unresolved or nonfinite channels are flagged, not omitted.

The same-product adaptation follows the controlled architecture of Li et al. (2025): a $784 \to n \to n \to 10$ ReLU MLP is pretrained on MNIST (LeCun et al., 1998) for 2,000 AdamW steps (batch 64, rate 10^{-3}), then its middle layer adapts to FashionMNIST (Xiao et al., 2017) for 100 steps without warmup, clipping or dropout. Widths are $256, 1024, 4096$, ranks $8, 32$, reciprocal scale ratios $1/4, 1/2, 1, 2, 4$, and seeds 42–44; data order, adapter and residual are paired. Numerical coefficient checks remain separate in Appendix I.

H Breadth, Initialization Diagnostics, and Cost

H.1 Initialization and complete loss responses

At $r = 32$, held-out median imbalance falls from 1.223 to 0.202 for PiSSA, 0.853 to 0.167 for MiLoRA, 0.620 to 0.169 for CorDA-IPM, and 1.705 to 0.219 for Init[AB]. However, 28, 44, 54 and 12 of 336 layer pairs respectively become more imbalanced. Median magnitude ratios span 0.914–1.090 rather than uniformly increasing. Thus the aggregate shift does not imply improvement at every layer. Run-level reductions remain positive at ranks 16, 32 and 64 without a shared monotonic rank trend. The complete loss responses retain flat minima, asymmetric deterioration and heterogeneous

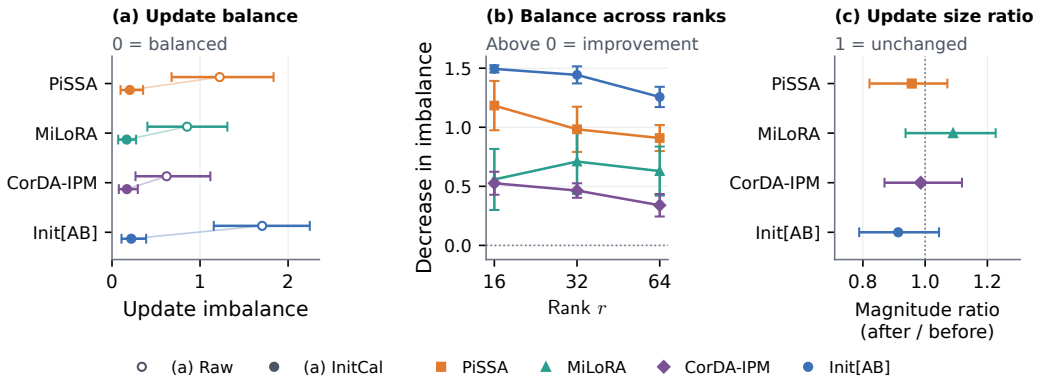

Figure 7: **Held-out initialization diagnostics.** Qwen3-0.6B, probe rate 10^{-4}. Imbalance is $|\log_2(D_A/D_B)|$; magnitude is $\sqrt{D_A D_B}$. In (a,c), points and bars summarize 336 layer pairs per method at $r = 32$ by medians and interquartile ranges, not confidence intervals. In (b), bars show three-run SD around the mean of run-level medians of paired imbalance reductions. All 336 paired layers per method remain in the source records.

window-width changes. Each raw/calibrated pair has its own validation-only rate selection. These curves are optimization diagnostics; task gains and seed-paired uncertainty are reported separately below.

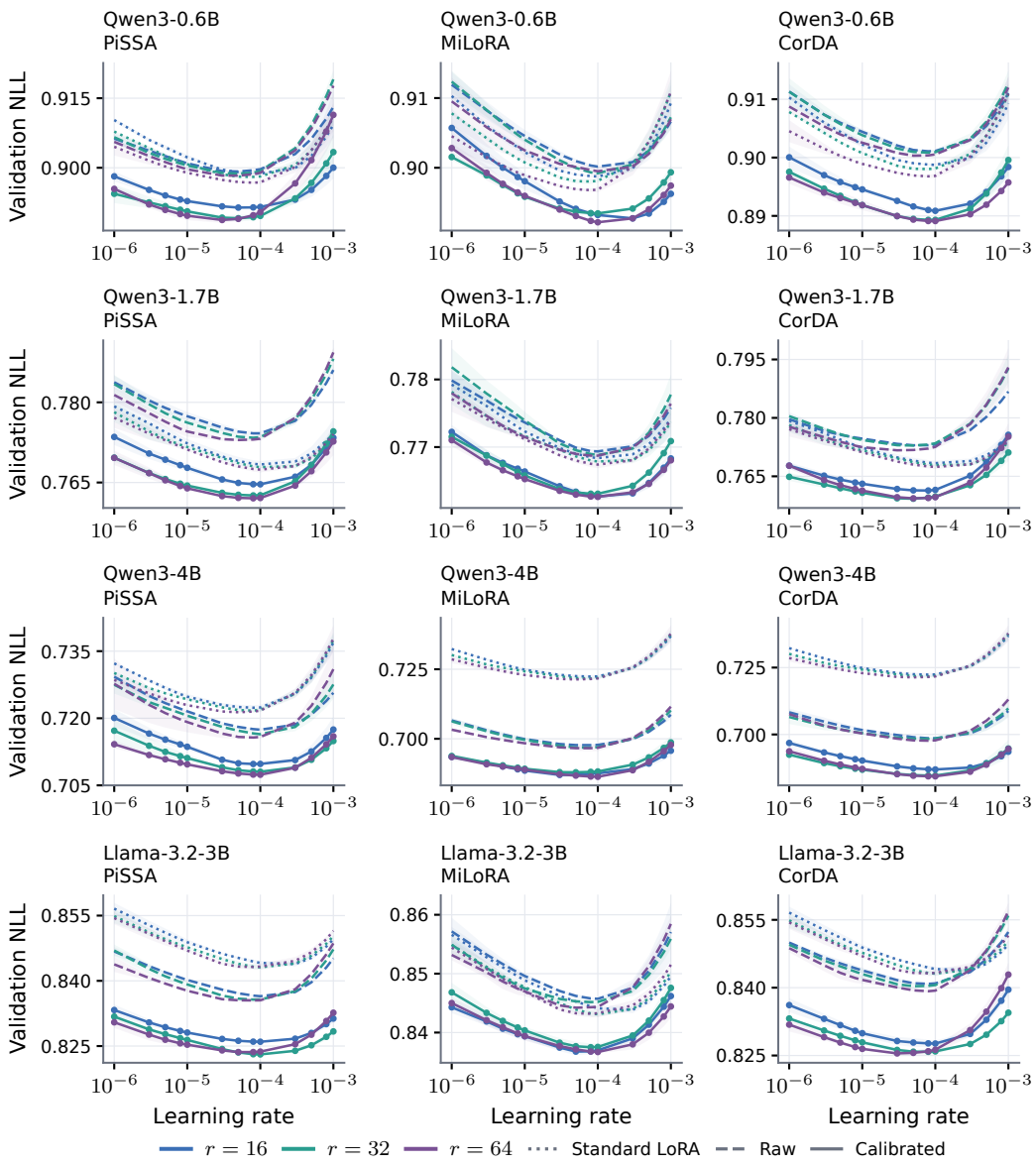

Figure 8: Complete validation-NLL curves across all four models and three structured initializers. Color denotes rank $(16, 32, 64)$; dotted, dashed, and solid lines denote Standard LoRA, raw initialization, and calibration. Bands show SD. Each Standard LoRA curve is repeated across the three initializer panels for comparison.

H.2 Task scores and paired changes

Task-level means and training-seed standard deviations are reported in Table 2; here we report paired changes and the complete breadth matrix. The full task table preserves the GSM8K regression for Qwen3-4B/MiLoRA and the cases below Standard LoRA. At rank 64, six of nine structured paired intervals include zero (eight of twelve including Init[AB]); the mean gain is not universal superiority. Across all model/rank combinations, the compact matrix below retains every structured pair rather than repeating four task columns and selection diagnostics for every row. Unaggregated task scores, seed records and all thirteen-rate values remain in the companion source data.

Table 6: **Paired four-task changes at rank 64.** Percentage points; $n = 3$ paired seeds. Each d_j averages four task changes within one seed. The descriptive reference interval is $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$, not significance evidence. Seed order is 42, 2025, 2026 for structured initializers and 101, 202, 303 for Init[AB]. Structured calibration is Warm; random calibration is direct $K = 2$.

Model	Initializer	d_1	d_2	d_3	Mean	SD	95% t reference interval
Qwen3-1.7B	PiSSA	+0.15	+0.77	+0.46	+0.46	0.31	[-0.32, +1.24]
Qwen3-1.7B	MiLoRA	+1.63	+1.14	+0.68	+1.15	0.47	[-0.03, +2.33]
Qwen3-1.7B	CorDA-IPM	+1.31	+1.40	+1.17	+1.29	0.11	[+1.01, +1.58]
Qwen3-4B	PiSSA	+0.70	+0.74	+0.11	+0.52	0.35	[-0.36, +1.40]
Qwen3-4B	MiLoRA	+0.69	+0.24	+0.81	+0.58	0.30	[-0.18, +1.34]
Qwen3-4B	CorDA-IPM	+1.07	+1.00	+0.70	+0.92	0.19	[+0.44, +1.40]
Llama-3.2-3B	PiSSA	+1.18	+0.93	+0.46	+0.85	0.37	[-0.06, +1.77]
Llama-3.2-3B	MiLoRA	+1.26	+0.63	+1.02	+0.97	0.32	[+0.17, +1.76]
Llama-3.2-3B	CorDA-IPM	+0.73	+0.81	+0.23	+0.59	0.31	[-0.19, +1.37]
Qwen3-1.7B	Random Init[AB]	+0.39	-0.16	+0.23	+0.15	0.28	[-0.55, +0.86]
Qwen3-4B	Random Init[AB]	-0.13	+0.84	+0.99	+0.56	0.61	[-0.94, +2.07]
Llama-3.2-3B	Random Init[AB]	+0.57	+0.41	+0.46	+0.48	0.08	[+0.28, +0.69]

Table 7: **Complete structured breadth matrix.** Four-task gain over paired Raw, mean ± sample SD across three training seeds. Every registered model–rank–initializer cell is retained.

Model	Rank	PiSSA	MiLoRA	CorDA-IPM
Qwen3-0.6B	16	+0.75±0.33	+0.66±0.31	+1.33±0.13
Qwen3-0.6B	32	+0.67±0.35	+1.23±0.33	+1.27±0.24
Qwen3-0.6B	64	+1.05±0.44	+0.71±0.54	+1.12±0.25
Qwen3-1.7B	16	+0.98±0.07	+0.32±0.59	+1.09±0.61
Qwen3-1.7B	32	+0.79±0.21	+0.55±0.42	+1.37±0.15
Qwen3-1.7B	64	+0.46±0.31	+1.15±0.47	+1.29±0.11
Qwen3-4B	16	+0.89±0.27	+0.50±0.23	+0.81±0.29
Qwen3-4B	32	+0.61±0.40	+0.84±0.13	+0.92±0.54
Qwen3-4B	64	+0.52±0.35	+0.58±0.30	+0.92±0.19
Llama-3.2-3B	16	+0.56±0.08	+0.90±0.25	+0.97±0.11
Llama-3.2-3B	32	+0.78±0.17	+0.91±0.07	+1.03±0.36
Llama-3.2-3B	64	+0.85±0.37	+0.97±0.32	+0.59±0.31

Table 8: Complete-Warm calibration measurement and its training reference. Peak memory is for the corresponding complete process, not an additive calibration component.

Measurement	Time	Peak memory	Fraction of training
Original fine-tuning run	57.3 min	11.6 GiB	100%
One-time complete Warm calibration	14.3 s	—	0.42%
Complete run including Warm calibration	—	11.6 GiB	—

H.3 Calibration sensitivity and complete-process cost

Sensitivity is not monotonic in calibration batch count. The timing uses Qwen3-0.6B/PiSSA-32, one RTX A6000 (48 GB), eager attention and gradient checkpointing. It covers the full Warm procedure, not independent start/correction/restoration components or Direct. Peak memory describes complete processes, not additive components. Numerical safeguards and batch definitions are shared with Appendices F and G.

I Executed Numerical Checks

Rectangular Gaussian prediction. The fixed shape/rank grid is $(48, 96, 8), (96, 48, 8), (64, 64, 8), (32, 128, 4), (128, 32, 16)$. Four spectra are flat (1), power-

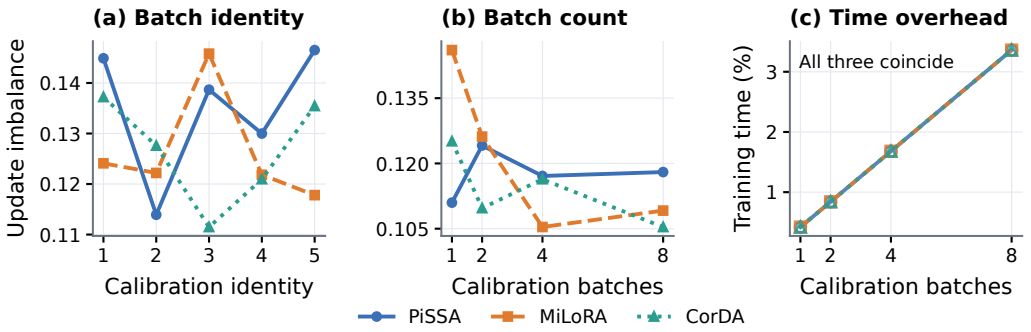

Figure 9: **Calibration data and cost.** **(a)** Held-out update imbalance for five calibration identities. **(b)** Imbalance versus batch count. **(c)** Time overhead. All three structured initializers are retained; coincident cost traces are identified. Update imbalance is $|\log_2(D_A/D_B)|$. Sensitivity values are point summaries.

law (i^{-2}), concentrated (i^{-4}), and spiked ($\sigma_1 = 1$, others 10^{-4}), with seeds 601–603. Reduced QR gives dense column-orthonormal factors, $s = \rho = 1$. Each of 60 configurations uses 16,384 independent Gaussian input/gradient pairs. We explicitly form each sample's gradients, take signs and multiply the two output paths; predictions are not fitted to observations. Squared output norms are pooled before RMS and scale estimation.

Median/maximum relative scale errors are 0.11%/0.51%, maximum K_A/K_B errors are 0.15%/1.13%, and the largest predicted-scale ratio deviation from one is 1.02%. The separate square-grid check uses widths $256, 1024, 4096$, ranks $8, 32, 128$, flat/concentrated spectra, seeds 401–403 and 2,048 Gaussian signals per setting: median/maximum errors are 0.14%/1.22%. Neither check establishes an identity for batch AdamW.

Table 9: **Finite rectangular scale prediction.** Direct Gaussian single-sample SignSGD checks, $r = 8$. Observed c: mean $\pm$ SD over three seeds, each pooling 16,384 samples. The last column is the largest relative c error across those seeds.

d_{in}	d_{out}	Spectrum	Predicted c	Observed c	Max. error
48	96	Flat	0.7708	0.7715 ± 0.0011	0.21%
48	96	Concentrated	0.5789	0.5799 ± 0.0010	0.35%
96	48	Flat	0.4590	0.4591 ± 0.0004	0.08%
96	48	Concentrated	0.3447	0.3449 ± 0.0010	0.28%
64	64	Flat	0.6036	0.6038 ± 0.0006	0.14%
64	64	Concentrated	0.4533	0.4532 ± 0.0004	0.12%

Balance versus interaction. Using the same measured unit-rate paths, evaluate independently predicted $\eta_{\text{lin}} = (K_A K_B)^{-1/2}$ and $\eta_{\text{safe}} = 1/\max\{\sqrt{K_A K_B}, \sqrt{K_{AB}}\}$ at $c = c_{\text{pop}}$. Here $K_{AB}^2 = r[n + (2/\pi)n(n-1)]$ follows from independent Gaussian output signs; linear and interaction paths scale as η and η^2. The interaction increases from 2.403 to 9.641 while both linear contributions stay near one. The safe rate keeps interaction near one but attenuates the linear contribution from 0.643 to 0.321. This finite grid illustrates the theorem's trade-off, not an asymptotic proof.

Whole-trajectory correspondence. The executed float64 check has $A \in \mathbb{R}^{3\times 7}$, $B \in \mathbb{R}^{5\times 3}$, eleven Gaussian input/target pairs (seed 19), and loss $\|BAX - Y\|_F^2/22$. For 60 steps, $(A/3, 3B)$ at rate .002 is paired with (A, B) at rates $(.006, .002/3)$. Adam uses moments $(.9, .999)$ initialized at zero. Ideal conditions set $\epsilon = \lambda = 0$ with no clipping; clipping-only uses threshold .3; finite AdamW uses $\epsilon = 10^{-8}, \lambda = .01$; mapped AdamW uses calibrated-side $\epsilon = 10^{-3}, \lambda = .1$ and Equation (93) on the rate side. Ideal products agree to numerical precision, including accumulated moments. Unchanged finite hyperparameters break correspondence, while the specified covariant mapping restores it. This

Table 10: **Balance does not control interaction.** Flat spectrum, $r = 8$, three-seed means of the Gaussian path measurements. The first three columns use η_{lin}; the last two use η_{safe}. Complete rank and spectrum combinations remain in the source records.

Width n	D_A	D_B	D_{AB}	Safe D_A	Safe D_{AB}
256	1.001	0.999	2.403	0.643	0.992
1024	1.000	0.995	4.825	0.455	0.996
4096	1.000	0.994	9.641	0.321	0.995

Table 11: **Whole-trajectory numerical correspondence.** Maximum relative Frobenius difference between the two adapter products over 60 executed steps. Finite-case differences describe this matrix problem, not downstream performance.

Optimizer configuration	Maximum product error
SignSGD	3.65×10^{-15}
Adam, ideal	7.09×10^{-16}
Adam, clipped	4.92×10^{-3}
AdamW	1.85×10^{-3}
AdamW, mapped	9.86×10^{-16}

is a representation check, not evidence of a downstream advantage. Joint nonlinear-network and raw-moment checks appear in Appendix N.

J Mechanism and Measurement Controls

J.1 Components and a second configuration

The seven PiSSA-32 variants follow Table 5. Initial diagnostics, reporting NLL and task scores use the same validation-selected setting. Total RMS is divided by the paired Warm value at Warm's selected rate; Total-norm match additionally uses its seed-specific multiplier. Analytic-only is the coarse energy start, not Spectrum-full. Direct/Same-probe LR share a Raw probe (Appendix F.3); Direct-2 changes only correction count, while Warm also changes the starting scale. The companion

Table 12: **Probe value under matched selection.** Five-seed mean ± SD. Initial diagnostics, NLL, and task scores use each variant's validation-selected setting, including the calibration multiplier for Total-norm match. RMS is normalized by paired Warm; Avg. equally weights four task-native percentage scores. Probes count virtual calibration/matching probes, not search candidates.

Variant	Probes	Held-out D_A/D_B	Combined RMS/Warm	Held-out NLL	Four-task Avg.
Raw	0	2.31±0.56	0.70±0.05	0.8967±0.0025	33.49±0.25
Analytic-only	0	0.41±0.10	1.16±0.08	0.8956±0.0019	33.59±0.37
Direct	1	0.93±0.20	1.00±0.01	0.8938±0.0022	34.19±0.36
Direct-2	2	0.91±0.17	1.00±0.00	0.8939±0.0020	34.21±0.29
InitCal-Warm	2	0.90±0.15	1.00±0.00	0.8929±0.0021	34.49±0.34
Same-probe LR	1	0.96±0.21	1.00±0.01	0.8938±0.0021	34.29±0.38
Total-norm match	3	2.31±0.56	0.97±0.05	0.8959±0.0025	33.53±0.30

seed-paired NLL ledger retains all seven methods. This table reports their selected absolute NLL alongside path diagnostics; task-level outcomes and key paired comparisons below retain the distinct replication units. The probe-informed variants improve on Raw and the coarse start, not on every probe-free alternative. A second direct correction does not give monotonic gains. Analytic-only

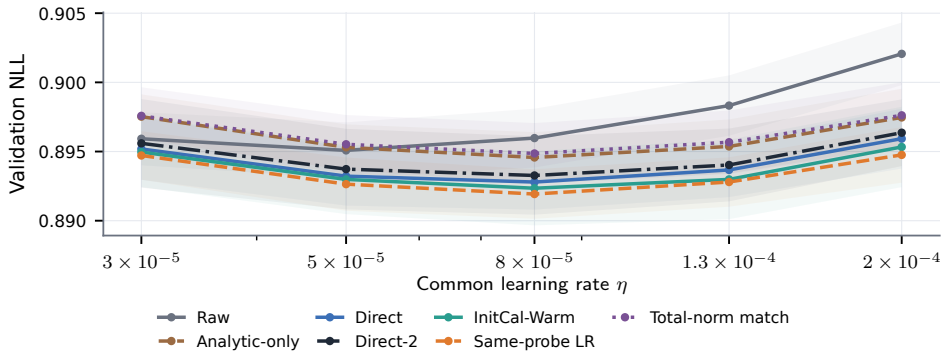

Figure 10: **Complete nominal-rate responses.** All five validation-NLL candidates for the seven variants; mean and SD over five paired seeds. Selection uses validation means, whereas Table 12 reports held-out NLL and task scores.

Table 13: **Four-task comparison under the unified selection protocol.** Qwen3-0.6B/PiSSA, $r = 32$; five-seed mean $\pm$ SD of task-native percentage scores. Avg. equally weights four tasks within each seed; the final column reports seed-paired average gain over Raw in percentage points.

Method	GSM8K	MATH-500	MMLU-Pro	BBH	Avg.	Paired gain vs. Raw
Raw	55.88±0.39	22.52±0.50	22.61±0.37	32.94±0.31	33.49±0.25	+0.00±0.00
Analytic-only	55.97±0.58	22.88±0.92	22.62±0.37	32.88±0.35	33.59±0.37	+0.10±0.17
Direct	56.71±0.44	23.36±0.75	23.24±0.40	33.46±0.39	34.19±0.36	+0.70±0.14
Direct-2	56.50±0.54	23.56±0.74	23.25±0.53	33.51±0.57	34.21±0.29	+0.72±0.07
InitCal-Warm	56.82±0.35	23.80±0.66	23.44±0.51	33.90±0.36	34.49±0.34	+1.00±0.21
Same-probe LR	56.63±0.58	23.68±0.73	23.19±0.52	33.66±0.28	34.29±0.38	+0.80±0.14
Total-norm match	55.95±0.52	22.68±0.54	22.57±0.34	32.94±0.55	33.53±0.30	+0.05±0.11

and Total-norm match retain negative seed–task changes; average improvement is not a per-task guarantee.

MiLoRA-64 repeats six variants with the same five paired seeds and five candidates, excluding Total-norm match: 150 candidate–seed settings and 30 selected records. No PiSSA and MiLoRA seeds are pooled. The paired comparisons distinguish extra correction, starting scale and rate-based implementation. Of the 5 key task contrasts, 2 wholly positive, 0 wholly negative, and 3 zero-

Table 14: **Second-configuration summary.** Qwen3-1.7B/MiLoRA, $r = 64$; each variant uses its own validation-selected nominal rate. NLL, four-task average, and seed-paired gain over Raw are five-seed mean $\pm$ SD. Task scores are percentages and gains are percentage points.

Variant	Selected η	Held-out NLL	Four-task Avg.	Paired gain vs. Raw
Raw	1.3×10^{-4}	0.7684±0.0007	45.06±0.33	+0.00±0.00
Analytic-only	8×10^{-5}	0.7679±0.0009	45.24±0.46	+0.18±0.15
Direct	8×10^{-5}	0.7665±0.0011	45.62±0.30	+0.55±0.06
Direct-2	8×10^{-5}	0.7662±0.0013	45.62±0.43	+0.56±0.12
InitCal-Warm	8×10^{-5}	0.7661±0.0010	45.76±0.34	+0.70±0.06
Same-probe LR	8×10^{-5}	0.7658±0.0008	45.65±0.33	+0.59±0.05

containing reference intervals are obtained from the unrounded paired differences. Wholly positive: Direct − Analytic-only, Warm − Analytic-only. Containing zero: Warm − Direct-2, Direct-2 − Direct, Same-probe LR − Direct.

Table 15: **Full task breakdown for the second configuration.** Five-seed mean ± SD of task-native percentage scores. Avg. equally weights tasks within seed before summarizing across seeds.

Variant	GSM8K	MATH-500	MMLU-Pro	BBH	Avg.
Raw	67.66±0.75	34.56±0.62	35.33±0.55	42.69±0.45	45.06±0.33
Analytic-only	67.92±0.78	34.68±0.70	35.38±0.67	42.98±0.41	45.24±0.46
Direct	68.28±0.79	35.08±0.61	35.82±0.47	43.28±0.39	45.62±0.30
Direct-2	68.22±0.96	35.04±0.79	35.85±0.52	43.36±0.40	45.62±0.43
InitCal-Warm	68.32±0.75	35.80±0.68	35.85±0.87	43.08±0.67	45.76±0.34
Same-probe LR	68.43±0.69	35.08±0.58	35.95±0.65	43.13±0.33	45.65±0.33

Table 16: **Key paired task comparisons in the second configuration.** Differences are first minus second, in percentage points. Reference intervals summarize five paired seeds at the selected settings.

Contrast	Paired task gain	95% reference interval
Direct − Analytic-only	+0.375±0.202	[+0.125, +0.626]
Warm − Analytic-only	+0.525±0.154	[+0.334, +0.716]
Warm − Direct-2	+0.148±0.125	[-0.007, +0.303]
Direct-2 − Direct	+0.001±0.174	[-0.215, +0.218]
Same-probe LR − Direct	+0.032±0.035	[-0.012, +0.076]

J.2 Exact output-space measurement

At each checkpoint, a method retains its own factors, activations and optimizer moments. Virtual steps use its fixed validation-selected peak rate, including factor-rate or norm-matching multipliers, on one shared held-out batch; they preserve moments and do not advance the scheduler. Training itself follows warmup/cosine decay. These are evolving-state diagnostics under fixed probe rates, not scheduled training increments.

The three paths in Equation (2) are tensors in $\mathbb{R}^{N_{\text{valid}} \times d_{\text{out}}}$, not rank-dimensional vectors. Sum their signed pairwise inner products over token chunks, then divide by the total coordinate count:

$$G_{ij} = \frac{\langle \delta_i, \delta_j \rangle_F}{N_{\text{valid}} d_{\text{out}}}, \qquad i, j \in \{A, B, AB\}. \tag{96}$$

This streaming reduction is exact up to arithmetic precision and preserves all cross terms without projection. For the representative 28 k-projections, pool sums and coordinate counts across layers before forming $D_i = \sqrt{G_{ii}}$, total RMS $T = \sqrt{\mathbf{1}^\top G \mathbf{1}}$ and D_{AB}/T. Each layer uses 1,024 valid tokens and output coordinates. Then summarize five seed-level diagnostics; do not average layer ratios. Padding is excluded. The geometric mean is not total RMS, and cancellation can make $D_{AB}/T > 1$. All virtual state is restored as in Appendix F.

The eight indices $0, 1, 5, 10, 25, 50, 100, 200$ yield 280 method–step–seed records; Figure 4 shows five variants, with Analytic-only/Direct-2 retained in the source data. Fixed-state common-rate multiplication leaves the linear ratio unchanged, but different training rates alter later states. Consequently, initially matching norms does not force later curves to coincide.

J.3 Norm matching and neighboring targets

Total-norm match leaves Raw factors fixed and solves a common-rate multiplier from cached paths:

$$G_{AB,AB}\lambda^4 + 2(G_{A,AB} + G_{B,AB})\lambda^3 + (G_{A,A} + 2G_{A,B} + G_{B,B})\lambda^2 - T_{\text{Warm}}^2 = 0.$$

Signed cross terms can make the norm nonmonotonic. Enumerate positive roots in $[1/16, 16]$, choose the one minimizing $|\log \lambda|$, and verify recomputed RMS; unresolved/failed matches remain flagged.

Warm's two target probes plus one Raw probe cost three virtual steps, while root solves reuse cached paths.

For target ratios $q = .5, 1, 2$, initialize $c_q = \sqrt{q/Q_{\mathrm{Raw}}}$, then rescale, reprobe finite AdamW and correct the achieved ratio before matching total RMS. Apply accepted constants unchanged to independent held-out tokens. These are fixed-reference-rate controls, not a new rate search. Both target-one-minus-neighbor NLL intervals include zero, so they do not identify a unique optimum.

Table 17: **Target-ratio control.** Five-seed mean ± SD. Calibration determines the factor scale and common-rate multiplier; held-out total RMS is normalized by paired Warm.

Target D_A/D_B	Held-out D_A/D_B	Total RMS/Warm	Held-out NLL
0.5	0.46±0.08	1.02±0.04	0.8940±0.0029
1	0.92±0.17	1.00±0.00	0.8934±0.0021
2	1.84±0.34	0.97±0.05	0.8944±0.0020

Table 18: **Paired target-allocation contrasts.** Held-out NLL difference, target one minus the stated neighboring target. Mean ± SD and descriptive 95% reference interval over five paired seeds.

Contrast	Paired ΔNLL	95% reference interval
$1 - 0.5$	-0.00057±0.00091	[-0.00170, +0.00055]
$1 - 2$	-0.00098±0.00096	[-0.00217, +0.00022]

J.4 Structure under single-example and batch updates

For each of PiSSA and MiLoRA, the comparison uses 28 k-projections in Qwen3-0.6B ($n = 1024, r = 32$) and a shared four-sequence, 1,024-valid-token batch. Each seed first takes the layer median of $|\log_2(c_{\mathrm{pred}}/c_{\mathrm{response}})|$ and post-probe imbalance; only then are seeds summarized. The figure instead displays each layer's five-seed median response scale.

The held-out response scale is an evaluation-only root of $\log_2[D_A(c)/D_B(c)]$. Recompute the layer gradient/step at each candidate, restore state, bracket in $[1/16, 16]$, refine roots and choose the acceptable root nearest one in log scale. Unbracketed, unresolved or inaccurate cases remain flagged. This search never chooses training scales/rates and is separate from reported calibration cost. Taking each example's gradient sign before pooling is distinct from signing the batch-averaged gradient; finite AdamW adds its actual denominator, clipping and decay. Spectrum-aware error is lower in

Table 19: **Structural prediction across update rules.** Five-seed mean ± SD of the 28-layer median absolute $\log_2$ scale error and post-probe imbalance, separately for PiSSA and MiLoRA. The same factors and diagnostic batch define the three rule comparisons.

Family	Update rule	Width–rank error	Spectrum-aware error	Post-probe imbalance
PiSSA	Single-example SignSGD	0.201±0.032	0.086±0.026	0.049±0.007
PiSSA	Batch SignSGD	0.227±0.045	0.141±0.029	0.073±0.014
PiSSA	Batch AdamW	0.218±0.042	0.156±0.035	0.094±0.009
MiLoRA	Single-example SignSGD	0.307±0.032	0.094±0.014	0.062±0.006
MiLoRA	Batch SignSGD	0.248±0.048	0.124±0.021	0.072±0.007
MiLoRA	Batch AdamW	0.242±0.057	0.190±0.058	0.119±0.017

all six family–rule rows, but rises under both batch rules. The prediction remains useful without becoming a batch-AdamW identity.

J.5 FINITE-OPTIMIZER SENSITIVITY

The shared baseline is AdamW with $\epsilon = 10^{-8}$, global clipping 1, decay .01, BF16 backbone and FP32 target paths. Separately set $\epsilon = 0$, disable clipping, disable decay, or make the backbone FP32. Factors, seeds and batches remain paired. Per seed, take the layer median imbalance and form Warm minus Direct; this is not a task-gain contrast. The baseline and zero-ϵ residual-gap intervals are

Table 20: **One-factor sensitivity of residual allocation.** All comparisons use the shared batch-AdamW baseline except for the named toggle. Direct and Warm columns report five-seed mean $\pm$ SD of within-seed layer-median imbalance. The final column gives the paired Warm-minus-Direct 95% reference interval.

Setting	Direct imbalance	Warm imbalance	Paired Warm$-$Direct interval
AdamW baseline	0.192$\pm$0.030	0.124$\pm$0.031	[-0.088, -0.047]
$\epsilon = 0$	0.183$\pm$0.019	0.127$\pm$0.038	[-0.083, -0.030]
No clipping	0.169$\pm$0.041	0.130$\pm$0.035	[-0.081, +0.004]
No weight decay	0.156$\pm$0.045	0.150$\pm$0.038	[-0.046, +0.034]
FP32	0.183$\pm$0.025	0.165$\pm$0.036	[-0.076, +0.040]

negative; other toggles contain zero. These conditional sensitivities do not decompose interacting causes of downstream gains. The separate calibration/training decay crossing in Appendix L tests final outcomes rather than treating local imbalance as their explanation.

K STRONG ALTERNATIVES AND LAYERWISE MAGNITUDE

K.1 SELECTED ALTERNATIVES AND FIXED CANDIDATE BUDGETS

Both focused configurations use all 112 attention projections, with shapes, grids and selection rules in Appendix G. Spectrum-full applies the rectangular spectral formula to every eligible module; Spectrum-hybrid keeps the historical square-spectral/rectangular-coarse split. Eligibility requires orthogonal square-root factors and resolved positive spectral quantities; unresolved modules are recorded rather than silently assigned another rule. Independent dual-rate search uses

$$\eta_A = \eta/\sqrt{\rho}, \qquad \eta_B = \eta\sqrt{\rho}. \tag{97}$$

Direct $-$ Spectrum-full: PiSSA-32 -0.091 points, $[-0.413, +0.231]$ (zero-containing); MiLoRA-64 $+0.015$ points, $[-0.145, +0.175]$ (zero-containing).

Warm $-$ Spectrum-full: PiSSA-32 $+0.206$ points, $[+0.019, +0.393]$ (wholly positive); MiLoRA-64 $+0.164$ points, $[-0.038, +0.367]$ (zero-containing).

Warm $-$ Tuned dual LR (5): PiSSA-32 $+0.342$ points, $[+0.107, +0.577]$ (wholly positive); MiLoRA-64 $+0.297$ points, $[+0.045, +0.549]$ (wholly positive).

Warm $-$ Tuned dual LR (25): PiSSA-32 $+0.297$ points, $[+0.061, +0.534]$ (wholly positive); MiLoRA-64 $+0.238$ points, $[-0.024, +0.501]$ (zero-containing).

The fixed candidate prefixes in Figure 5 select solely by five-seed validation means. Direct and Warm stop at five settings; the dual-rate grid continues to 25. The five-setting Warm NLL values are .89286/.76613, versus .89335/.76637 for dual-rate and .89276/.76578 for its full grid (PiSSA/MiLoRA). More candidates need not monotonically improve held-out NLL. Candidate order, grid geometry and calibration work differ, so these curves neither prove optimal search nor compare equal wall-clock budgets; the complete prefix ledger stays in source data.

K.2 PER-MODULE AND FULL-NETWORK DIAGNOSTICS

Layer-norm match keeps Raw factors and equal A/B rates within each layer. At each nominal-rate candidate, independently solve the signed-Gram norm equation for all 112 modules, targeting Warm's same-candidate combined RMS. Choose the root nearest one in log scale and require calibration

Table 21: **Validation-selected strong controls.** NLL is evaluated on the independent held-out partition. Avg. is the equal-weight four-task percentage score; gain is relative to paired Raw in percentage points. Results are five-seed mean ± SD. The ratio column gives the searched global ρ only for Tuned dual LR; a dash denotes no searched global ratio, not equal per-layer effective rates.

Configuration	Method	η	ρ	NLL	Avg.	Gain
PiSSA-32	Direct	8×10^{-5}	–	0.8938±0.0022	34.19±0.36	+0.70±0.14
PiSSA-32	InitCal-Warm	8×10^{-5}	–	0.8929±0.0021	34.49±0.34	+1.00±0.21
PiSSA-32	Spectrum-full	8×10^{-5}	–	0.8930±0.0023	34.28±0.31	+0.80±0.16
PiSSA-32	Tuned dual LR (5)	1.3×10^{-4}	4	0.8933±0.0021	34.15±0.26	+0.66±0.04
PiSSA-32	Tuned dual LR (25)	8×10^{-5}	4	0.8928±0.0021	34.19±0.26	+0.70±0.04
PiSSA-32	Layer-norm match	8×10^{-5}	–	0.8954±0.0027	33.75±0.25	+0.26±0.17
PiSSA-32	Spectrum-hybrid	8×10^{-5}	–	0.8947±0.0024	33.72±0.28	+0.24±0.09
MiLoRA-64	Direct	8×10^{-5}	–	0.7665±0.0011	45.62±0.30	+0.55±0.06
MiLoRA-64	InitCal-Warm	8×10^{-5}	–	0.7661±0.0010	45.76±0.34	+0.70±0.06
MiLoRA-64	Spectrum-full	8×10^{-5}	–	0.7660±0.0009	45.60±0.38	+0.54±0.14
MiLoRA-64	Tuned dual LR (5)	1.3×10^{-4}	4	0.7664±0.0007	45.47±0.43	+0.41±0.18
MiLoRA-64	Tuned dual LR (25)	8×10^{-5}	4	0.7658±0.0007	45.53±0.42	+0.47±0.19
MiLoRA-64	Layer-norm match	8×10^{-5}	–	0.7676±0.0008	45.30±0.30	+0.23±0.11
MiLoRA-64	Spectrum-hybrid	8×10^{-5}	–	0.7674±0.0009	45.24±0.40	+0.18±0.13

Table 22: **Paired task contrasts for the strong controls.** First method minus second, in percentage points; five paired training seeds. Intervals are descriptive 95% t reference intervals.

Configuration	Contrast	Paired gain	95% interval
PiSSA-32	Direct − Spectrum-full	-0.091±0.259	[-0.413, +0.231]
PiSSA-32	InitCal-Warm − Spectrum-full	+0.206±0.151	[+0.019, +0.393]
PiSSA-32	Spectrum-full − Spectrum-hybrid	+0.560±0.225	[+0.280, +0.840]
PiSSA-32	InitCal-Warm − Tuned dual LR (5)	+0.342±0.189	[+0.107, +0.577]
PiSSA-32	InitCal-Warm − Tuned dual LR (25)	+0.297±0.190	[+0.061, +0.534]
PiSSA-32	InitCal-Warm − Layer-norm match	+0.740±0.152	[+0.552, +0.929]
MiLoRA-64	Direct − Spectrum-full	+0.015±0.129	[-0.145, +0.175]
MiLoRA-64	InitCal-Warm − Spectrum-full	+0.164±0.163	[-0.038, +0.367]
MiLoRA-64	Spectrum-full − Spectrum-hybrid	+0.357±0.238	[+0.062, +0.653]
MiLoRA-64	InitCal-Warm − Tuned dual LR (5)	+0.297±0.203	[+0.045, +0.549]
MiLoRA-64	InitCal-Warm − Tuned dual LR (25)	+0.238±0.211	[-0.024, +0.501]
MiLoRA-64	InitCal-Warm − Layer-norm match	+0.469±0.135	[+0.301, +0.637]

residual at most 10^{-8}. Fix multipliers for training and held-out evaluation; do not rematch held-out batches.

Layer error is the within-seed median of $100|T_\ell/T_{\ell,\mathrm{Warm}} - 1|$, followed by five-seed statistics. Reporting uses each method's selected rate; Warm and the layer-matched method select the same rate in both configurations. A separate disposable full-network step measures logits-change RMS relative to paired Warm and $\Delta L_{\mathrm{virt}} = L(z^{\mathrm{virt}}) - L(z^0)$ on shared valid tokens. These are actual network-output definitions, not quantities inferred from local Grams. Small held-out layer errors and near-Warm logits RMS coexist with smaller task gains for the matched control. Conversely, the 25-setting dual-rate search has a larger local NLL decrease on MiLoRA. Initial amplitude therefore does not explain all outcomes, but these mixed observations do not identify balance as their sole cause.

K.3 Starting scales, correction count, and safeguards

Final $\log_2 c$ and residual imbalance first take seedwise medians over 112 layers. Early loss decrease is $L(0) - L(20)$ under scheduled training on a fixed held-out batch, not a fixed-rate virtual step. Direct-2 lowers residual imbalance without improving early loss in either configuration; Warm's early loss is nearly unchanged from Direct on MiLoRA. Thus a smaller residual does not make the

Table 23: **Layerwise magnitude and full-network diagnostics at selected settings.** Layer-norm residuals use held-out batches with no rematching. Logits RMS is normalized by paired Warm; virtual-step ΔNLL is after minus before the disposable step.

Configuration	Method	Layer-norm residual (%)	Logits RMS /Warm	Virtual-step ΔNLL
PiSSA-32	Raw	11.97±0.21	0.85±0.03	-0.00028±0.00008
PiSSA-32	Direct	0.29±0.06	1.01±0.03	-0.00110±0.00017
PiSSA-32	InitCal-Warm	0.00±0.00	1.00±0.00	-0.00112±0.00012
PiSSA-32	Spectrum-full	0.44±0.15	0.99±0.05	-0.00105±0.00017
PiSSA-32	Tuned dual LR (5)	67.11±0.91	1.68±0.04	-0.00156±0.00022
PiSSA-32	Tuned dual LR (25)	2.48±0.50	1.03±0.03	-0.00102±0.00014
PiSSA-32	Layer-norm match	2.49±0.25	1.03±0.03	-0.00066±0.00012
PiSSA-32	Spectrum-hybrid	0.41±0.13	1.10±0.03	-0.00076±0.00005
MiLoRA-64	Raw	175.15±1.71	2.28±0.11	-0.00041±0.00018
MiLoRA-64	Direct	0.30±0.06	1.01±0.07	-0.00086±0.00011
MiLoRA-64	InitCal-Warm	0.00±0.00	1.00±0.00	-0.00085±0.00007
MiLoRA-64	Spectrum-full	0.46±0.04	0.97±0.05	-0.00091±0.00009
MiLoRA-64	Tuned dual LR (5)	70.70±0.34	1.66±0.08	-0.00149±0.00017
MiLoRA-64	Tuned dual LR (25)	4.64±0.21	1.02±0.05	-0.00097±0.00010
MiLoRA-64	Layer-norm match	2.83±0.24	1.07±0.04	-0.00051±0.00010
MiLoRA-64	Spectrum-hybrid	0.42±0.06	1.07±0.03	-0.00067±0.00014

coarse start or second correction theoretically necessary. Unresolved channels use $q_\ell = 1$ and remain

Table 24: **Starting-point and correction-count diagnostics.** Checkpoint zero is after calibration and before training. The NLL decrease uses scheduled training updates; final Avg. is the four-task percentage score.

Configuration	Method	Final $\log_2 c$	Residual imbalance	20-step NLL decrease	Final Avg.
PiSSA-32	Direct	-1.251±0.023	0.159±0.021	0.0093±0.0003	34.19±0.36
PiSSA-32	Direct-2	-1.250±0.029	0.149±0.028	0.0091±0.0013	34.21±0.29
PiSSA-32	InitCal-Warm	-1.235±0.019	0.123±0.011	0.0103±0.0005	34.49±0.34
MiLoRA-64	Direct	-1.352±0.013	0.155±0.014	0.0076±0.0009	45.62±0.30
MiLoRA-64	Direct-2	-1.363±0.020	0.149±0.013	0.0070±0.0005	45.62±0.43
MiLoRA-64	InitCal-Warm	-1.350±0.014	0.125±0.010	0.0076±0.0004	45.76±0.34

Table 25: **Safeguard ledger for the focused controls.** Unresolved count is the maximum count among five seeds, each with 112 layers; these layers are not successful balancing observations. Error columns are maxima over seeds and layers, checked against 10^{-5}.

Configuration	Method	Unresolved layers	Maximum product error	Maximum forward error
PiSSA-32	Direct	0	1.6×10^{-7}	3.6×10^{-7}
PiSSA-32	Direct-2	0	1.9×10^{-7}	3.7×10^{-7}
PiSSA-32	InitCal-Warm	0	1.9×10^{-7}	3.8×10^{-7}
MiLoRA-64	Direct	0	2×10^{-7}	3.5×10^{-7}
MiLoRA-64	Direct-2	0	2×10^{-7}	2.5×10^{-7}
MiLoRA-64	InitCal-Warm	0	1.9×10^{-7}	3.4×10^{-7}

counted. The focused detectability threshold is 10^{-12} after common pretrained-output normalization, with no scale cap; product and forward relative errors must be at most 10^{-5}. These do not replace the separate breadth tolerances in Appendix F.

L Cumulative-Scale Mapping and Decay Stages

Warm-mapped LR restores Raw factors and residuals, then maps Warm's complete accumulated scale, not merely its last correction:

$$c_\ell = c_{\mathrm{cf},\ell} q_{\ell,1} q_{\ell,2}, \qquad \eta_{A,\ell,t} = \eta_t c_\ell, \qquad \eta_{B,\ell,t} = \eta_t / c_\ell. \tag{98}$$

It inherits Warm's selected peak rate 8×10^{-5}, two probes and schedule, without another search. Initialization, calibration information, seeds and batch order remain paired. Both start fresh optimizers; the mapped branch leaves ϵ, decay and clipping untransformed, unlike Equation (93). Same-probe LR instead maps Direct's single Raw probe. Finite-AdamW identity is therefore not assumed.

Executed replay check. For $A \in \mathbb{R}^{4\times 12}$, $B \in \mathbb{R}^{16\times 4}$ and seeds $71, 82, 93, 104, 115$, compare sequential scales $(.52, 1.07, .97)$ (discarding and restoring each virtual probe/RNG) with their one-shot product. Fresh AdamW branches share 20 minibatches. Maximum initial factor and trajectory-product discrepancies are 8.88×10^{-16} and 1.78×10^{-15}. This tests replay with no retained probe history, not language-model scale selection or a benefit from the sequence itself.

Calibration and training decay. PiSSA-32 crosses Direct/Warm with calibration decay $c \in \{0, .01\}$ and training decay $t \in \{0, .01\}$: eight groups at fixed rate 8×10^{-5}, with five paired seeds and no reselection. Only c affects virtual probes; after state restoration, fresh training uses t. The Warm row reused by the mapping comparison is not a new replicate. For seed j, write $Y_{j,m,c,t}$ for

Table 26: **Final outcomes of inherited-scale mapping and stage-specific decay controls.** All rows use nominal peak rate 8×10^{-5} with no new selection. The first four rows compare Warm with its raw-factor two-rate implementation at decay 0.01 in both stages; the remaining eight cross method with $(\lambda_{\mathrm{cal}}, \lambda_{\mathrm{train}})$. NLL is held out; Avg. is the equal-weight four-task percentage score. Entries are five-seed mean ± SD. PiSSA-32 Warm at $(0.01, 0.01)$ appears in both comparisons, not as independent evidence.

Configuration	Comparison	Method	Held-out NLL	Four-task Avg.
PiSSA-32	Mapping	InitCal-Warm	0.89286±0.00205	34.49±0.34
PiSSA-32	Mapping	Warm-mapped LR	0.89284±0.00204	34.49±0.33
MiLoRA-64	Mapping	InitCal-Warm	0.76613±0.00105	45.76±0.34
MiLoRA-64	Mapping	Warm-mapped LR	0.76615±0.00131	45.77±0.34
PiSSA-32	$(0, 0)$	Direct	0.89383±0.00236	34.25±0.34
PiSSA-32	$(0, 0)$	InitCal-Warm	0.89300±0.00237	34.42±0.33
PiSSA-32	$(0, 0.01)$	Direct	0.89364±0.00226	34.21±0.38
PiSSA-32	$(0, 0.01)$	InitCal-Warm	0.89281±0.00202	34.48±0.36
PiSSA-32	$(0.01, 0)$	Direct	0.89374±0.00239	34.24±0.35
PiSSA-32	$(0.01, 0)$	InitCal-Warm	0.89324±0.00239	34.46±0.36
PiSSA-32	$(0.01, 0.01)$	Direct	0.89381±0.00221	34.19±0.36
PiSSA-32	$(0.01, 0.01)$	InitCal-Warm	0.89286±0.00205	34.49±0.34

either four-task average or held-out NLL. Define method gaps, stage-averaged changes and their interaction:

$$d_j(c,t) = Y_{j,\mathrm{Warm},c,t} - Y_{j,\mathrm{Direct},c,t}, \qquad h_j = d_j(0.01, 0.01) - d_j(0, 0). \tag{99}$$

$$\begin{aligned} C_j &= \tfrac{1}{2}\{d_j(0.01, 0) - d_j(0, 0) + d_j(0.01, 0.01) - d_j(0, 0.01)\}, \\ T_j &= \tfrac{1}{2}\{d_j(0, 0.01) - d_j(0, 0) + d_j(0.01, 0.01) - d_j(0.01, 0)\}, \\ I_j &= d_j(0.01, 0.01) - d_j(0.01, 0) - d_j(0, 0.01) + d_j(0, 0). \end{aligned} \tag{100}$$

Here $C_j + T_j = h_j$, while I_j tests whether one stage's effect on the method gap depends on the other. The metrics and five-seed interval convention follow Appendix G.3; positive task gaps and negative NLL gaps favor Warm. Warm and Warm-mapped LR have absolute mean task differences below 0.01 percentage points in both configurations; their task and NLL intervals all include zero.

Table 27: **Paired differences for cumulative-scale mapping.** Five seed-paired contrasts between Warm and Warm-mapped LR in each configuration, reported separately for task Avg. and held-out NLL; mean ± sample SD and descriptive 95% reference interval. Avg. differences are percentage points.

Configuration	Contrast	Metric	Paired difference	95% interval
PiSSA-32	Warm − Warm-mapped LR	Avg.	+0.000±0.056	[-0.069, +0.070]
PiSSA-32	Warm − Warm-mapped LR	NLL	+0.00003±0.00040	[-0.00047, +0.00052]
MiLoRA-64	Warm − Warm-mapped LR	Avg.	-0.004±0.105	[-0.135, +0.126]
MiLoRA-64	Warm − Warm-mapped LR	NLL	-0.00002±0.00029	[-0.00039, +0.00035]

Table 28: **Stage-specific paired contrasts for PiSSA-32.** The four Warm–Direct gaps $d(c,t)$, joint toggle h, stage-averaged changes C, T, and stage interaction I are defined in Equations (99) and (100). Each uses five paired seeds; entries are mean ± sample SD and descriptive 95% reference intervals. Avg. differences are percentage points.

Contrast	Avg. difference	95% interval	NLL difference	95% interval
$d(0,0)$	+0.172±0.349	[-0.261, +0.606]	-0.00083±0.00174	[-0.00299, +0.00133]
$d(0,0.01)$	+0.267±0.376	[-0.200, +0.734]	-0.00083±0.00117	[-0.00228, +0.00063]
$d(0.01,0)$	+0.210±0.352	[-0.227, +0.648]	-0.00049±0.00172	[-0.00263, +0.00165]
$d(0.01,0.01)$	+0.297±0.338	[-0.123, +0.717]	-0.00095±0.00114	[-0.00236, +0.00047]
Joint toggle	+0.124±0.071	[+0.037, +0.212]	-0.00011±0.00075	[-0.00104, +0.00082]
Calibration gap change C	+0.034±0.072	[-0.056, +0.123]	+0.00011±0.00013	[-0.00005, +0.00027]
Training gap change T	+0.091±0.087	[-0.018, +0.199]	-0.00022±0.00077	[-0.00117, +0.00073]
Stage interaction	-0.008±0.070	[-0.095, +0.079]	-0.00046±0.00039	[-0.00095, +0.00003]

Close point estimates and failure to exclude zero are not equivalence tests: no equivalence margin was specified. These outcomes do not establish an intrinsic advantage of factor rescaling over the mapped implementation.

For PiSSA-32, the joint toggle increases the mean Warm–Direct task gap by 0.124 percentage points (95% reference interval $[0.037, 0.212]$). The stage-averaged changes are 0.034 for calibration ($[-0.056, 0.123]$) and 0.091 for training ($[-0.018, 0.199]$), while the stage interaction is -0.008 ($[-0.095, 0.079]$). All four simple-gap intervals and all NLL contrast intervals include zero. The positive joint-toggle interval is compatible with the zero-containing stage intervals: their sum retains the seed-paired covariance and answers a different question. At this fixed rate, neither stage is individually resolved as the source of the task-gap change, and the zero-containing interaction interval does not establish additivity. The contrasts do not show that decay explains all Warm gains or that local residual balance determines final performance.

M CorDA-IPM and the Covariance Reference

The Llama-3.2-3B/CorDA-IPM comparison fixes rank 64, the selected adapter and frozen residual. For the regularized invertible construction transform C_{reg}, decompose $W_{\mathrm{pre}}C_{\mathrm{reg}} = U\Sigma V^\top$ and use the IPM components:

$$B = s^{-1/2}U_I\Sigma_I^{1/2}, \qquad A = s^{-1/2}\Sigma_I^{1/2}V_I^\top C_{\mathrm{reg}}^{-1}. \tag{101}$$

Thus $B^\top B = \Sigma_I/s$, but $AA^\top$ need not be diagonal. This applies to the stated covariance-transform implementation, not every CorDA variant. Construction transform C_{reg} and input covariance C_x are distinct.

For $x \sim \mathcal{N}(0, C_x)$, define the absolute-value moment

$$G(C) := \mathbb{E}\|\mathcal{N}(0,C)\|_1^2 \tag{102}$$

$$= \operatorname{tr} C + \frac{2}{\pi}\sum_{i\neq j}\sqrt{C_{ii}C_{jj}}\left(\sqrt{1-r_{ij}^2} + r_{ij}\arcsin r_{ij}\right), \qquad r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}. \tag{103}$$

The sum uses ordered off-diagonal pairs; zero-variance terms vanish. Under nonzero signal coordinates/projections and orthogonal nonzero B columns, the same-input SignSGD energies are $E_A = \|B\|_F^2 G(C_x)$ and $E_B = d_{\text{out}} G(AC_xA^\top)$ by Equation (106). Hence

$$c_{\text{cov}}^4 = \rho^2 \frac{d_{\text{out}} G(AC_xA^\top)}{\|B\|_F^2 G(C_x)}. \tag{104}$$

No input–gradient independence is needed. With isotropic inputs and orthogonal square-root factors, this recovers the exact spectral rule. Gaussianity is required for this covariance-based moment formula, not for the empirical raw-moment rule in Appendix N; neither is a finite-batch AdamW identity.

All methods share the same 256-conversation training subset for CorDA construction. Covariance-aware reuses its centered sufficient statistics $(N, \sum x, \sum xx^\top)$ without a new backward pass. Direct and Warm-Cov use the same contained four-conversation calibration batch, with one/two corrections. Warm-Cov uses the covariance start, not the coarse-start Warm of the breadth matrix. Five candidates and five paired seeds follow Table 5; Direct/Warm-Cov require 25/50 virtual calibration steps over the search, separate from complete-training counts. Held-out NLL and tasks do not enter selection.

Table 29: **CorDA-IPM focused comparison.** Five-seed mean ± SD. Avg. is the equal-weight four-task score; gain is paired to Raw in percentage points. NLL is independent reporting NLL. ρ is displayed only for the searched global ratio.

Method	η	ρ	NLL	Avg.	Gain
Raw	8×10^{-5}	–	0.8395 ± 0.0024	24.26 ± 0.36	0.00 ± 0.00
Covariance-aware	8×10^{-5}	–	0.8370 ± 0.0017	24.51 ± 0.33	0.24 ± 0.27
Direct	8×10^{-5}	–	0.8359 ± 0.0014	25.18 ± 0.63	0.92 ± 0.31
Warm-Cov	8×10^{-5}	–	0.8353 ± 0.0012	25.05 ± 0.28	0.79 ± 0.23
Tuned dual LR (5)	1.3×10^{-4}	4	0.8361 ± 0.0012	24.79 ± 0.58	0.52 ± 0.35

Table 30: **CorDA seed-paired contrasts.** First method minus second, in percentage points. The descriptive 95% t reference intervals condition on selected candidates and do not include selection uncertainty or multiplicity correction.

Contrast	Mean ± SD	95% interval
Covariance-aware − Raw	$+0.245 \pm 0.269$	$[-0.089, +0.579]$
Direct − Covariance-aware	$+0.676 \pm 0.502$	$[+0.053, +1.299]$
Warm-Cov − Covariance-aware	$+0.544 \pm 0.313$	$[+0.156, +0.933]$
Warm-Cov − Direct	-0.131 ± 0.467	$[-0.712, +0.449]$
Warm-Cov − Tuned dual LR (5)	$+0.266 \pm 0.401$	$[-0.232, +0.764]$
Direct − Tuned dual LR (5)	$+0.398 \pm 0.394$	$[-0.092, +0.887]$

Warm-Cov minus Direct is -0.131 points, with reference interval $[-0.712, +0.449]$. These selected-setting contrasts do not establish that Warm-Cov must outperform Direct or the independent rate search. Warm-Cov is the covariance-start variant, not the coarse-start Warm in the breadth tables. These comparisons fix the subspace, not choose a better one. Forward-moment below separates real activation-distribution information from the Gaussian approximation, so a gain over covariance alone cannot isolate the value of backward information.

N Raw Forward Moments and an Executed Ideal-Condition Check

Distribution-free same-sample identity. Consider $y = Rx + sBAx$ with $s > 0$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$. Let $\delta = \nabla_y \mathcal{L}$ be the output signal for this same sample. Assume that the nonzero columns of B are mutually orthogonal, $\|B\|_F > 0$, and, almost surely, every coordinate of δ and

every projection $b_j^\top \delta$ for a nonzero column b_j is nonzero. Zero columns of B, and zero coordinates of x or Ax, are allowed. The two raw moments below must be finite and positive. Neither Gaussian inputs nor independence between x and δ is required.

For single-sample SignSGD, the actual outer-product gradients are $\nabla_A \mathcal{L} = sB^\top \delta x^\top$ and $\nabla_B \mathcal{L} = s\delta(Ax)^\top$. With $\operatorname{sign}(0) = 0$, the two same-input linear paths, omitting their learning rates and common factor $-s$, are

$$p_A = B \operatorname{sign}(B^\top \delta)\|x\|_1, \qquad p_B = \operatorname{sign}(\delta)\|Ax\|_1, \tag{105}$$

$$\|p_A\|_2^2 = \|B\|_F^2 \|x\|_1^2, \qquad \|p_B\|_2^2 = d_{\text{out}} \|Ax\|_1^2. \tag{106}$$

The first equality of norms follows pointwise from the diagonal Gram matrix: every nonzero-column sign has squared value one. The second follows from the d_{out} nonzero signal coordinates. If these signal conditions fail, the sign-support factors can depend on x and cannot simply be replaced by $\|B\|_F^2$ and d_{out}.

Write $\mu_x = \mathbb{E}\|x\|_1^2$ and $\mu_{Ax} = \mathbb{E}\|Ax\|_1^2$. Under $B \leftarrow cB$, $A \leftarrow A/c$, the two expected linear-output energies are $s^2\eta_A^2 c^2 \|B\|_F^2 \mu_x$ and $s^2 \eta_B^2 c^{-2} d_{\text{out}} \mu_{Ax}$. Thus, for $\rho = \eta_B/\eta_A$,

$$c_{\text{fwd}}^4 = \rho^2 \frac{d_{\text{out}} \mu_{Ax}}{\|B\|_F^2 \mu_x}, \qquad \widehat{c}_{\text{fwd}}^4 = \rho^2 \frac{d_{\text{out}} \overline{\|Ax\|_1^2}}{\|B\|_F^2 \, \overline{\|x\|_1^2}}. \tag{107}$$

The empirical formula exactly balances the average of these *per-sample* energies over a valid sample set; it is not an expectation of per-sample ratios. The population rule holds whenever the stated moments exist. A representative calibration sample is an additional requirement for using its empirical estimate on another input distribution. For zero-mean Gaussian inputs, the two moments become $G(C_x)$ and $G(AC_xA^\top)$ in Equation (103); the raw-moment rule itself also permits nonzero means and non-Gaussian inputs. It is not an identity for the sign of a batch-averaged gradient, finite-ϵ AdamW, or loss-optimal allocation.

Streaming statistic, without centering. For a fixed snapshot of A, a binary validity mask selects tokens before accumulation. Each chunk contributes only $(N, \sum_i \|x_i\|_1^2, \sum_i \|Ax_i\|_1^2)$; divide the two sums by the same effective token count. Unequal chunks are not given equal weight. Neither $\overline{\|x\|_1}^2$ nor a moment of $x - \bar{x}$ is the required statistic. No activation, covariance matrix, or gradient is retained. The implementation rejects incompatible shapes, nonfinite valid tokens or sums, empty effective samples, nonpositive moments, and nonorthogonal nonzero columns of B; masked padding is excluded before finite-value checks. The additional storage is two scalar sums and a count, apart from the fixed factor snapshot and current chunk.

Joint ideal calibration in a nonlinear network. The loss depends on the factors through the network function, without a separate factor-level regularizer. This is a separate identity: it does not require column orthogonality or single-sample gradients. With all adapter products and residuals fixed, identical inputs and stochastic state preserve every layer's activations and upstream signals when arbitrary positive layerwise c_ℓ are applied jointly. Hence the full-batch gradients transform as $g_{A,\ell} \mapsto c_\ell g_{A,\ell}$ and $g_{B,\ell} \mapsto c_\ell^{-1} g_{B,\ell}$. For a fresh, bias-corrected AdamW first step with $\epsilon = 0$, zero decay, and clipping disabled, the update is $-\eta \operatorname{sign}(g)$, including zero-gradient entries. Therefore $D_{A,\ell}(c_\ell) = c_\ell D_{A,\ell}(1)$ and $D_{B,\ell}(c_\ell) = D_{B,\ell}(1)/c_\ell$. Positive channels give

$$c_{0,\ell} \sqrt{D_{B,\ell}(c_{0,\ell})/D_{A,\ell}(c_{0,\ell})} = \sqrt{D_{B,\ell}(1)/D_{A,\ell}(1)}. \tag{108}$$

All layers reach the same cumulative scale as Direct after one correction; a second correction is one. Equality concerns the restored-and-rescaled initialization, not an unconstrained multistep training trajectory.

Executed scope and results. The NumPy float64 numerical checks are recorded in `paper_data/information_controls/math_checks.json`. First, 38 masked tokens drawn from shifted exponential distributions and an input-dependent δ verify Equation (106) by explicitly forming each outer-product gradient. Maximum normalized error is 7.1×10^{-16}; four-chunk versus full-batch moment error is 1.3×10^{-16}. Here $\overline{\|x\|_1^2} = 57.815$, whereas $\overline{\|x\|_1}^2 = 53.382$ and centering first gives 16.296.

Second, a $5 \to 7 \to 6 \to 4$ network with ranks $(2, 3, 2)$, two tanh layers, 15% dropout, and a fixed 12-sample batch executes joint calibration. Explicit reverse-mode derivatives of the actual squared loss are audited at all 83 factor entries by central differences; maximum absolute error is 1.2×10^{-10}. We evaluate the fresh AdamW first-step closed form, not a persistent optimizer implementation. Four nonuniform starting-scale vectors, spanning $1/32$ to 16, each receive two corrections; their cumulative scales differ from one-round Direct by at most 2.3×10^{-16} relatively, and second corrections differ from one by at most 1.2×10^{-16}. Maximum absolute product and final-forward discrepancies are 8.4×10^{-17} and 1.2×10^{-16}. Probes are pure functions: parameters, residuals, batches and global RNG remain unchanged; repeated probes are bitwise identical. No scheduler or mixed-precision scaler exists in this NumPy check.

The same network also executes boundary controls with $(\epsilon, \lambda) = (0.03, 0), (0, 0.2), (0.03, 0.2)$, where λ is decay. For the recorded nonuniform start, two-round cumulative-scale differences from the corresponding one-round Direct are 0.714, 0.00694, and 0.297 in maximum relative error. These are counterexamples to extending the ideal identity unchanged, not downstream rankings or explanations of Warm's observed gains. This test runs no Qwen/Llama model, training campaign, or real-model calibration; it verifies the mathematical paths and small-network implementation only.

O Information and Layer-Assignment Controls

O.1 Forward moments and matched access

Forward-moment and Direct use the same four calibration conversations, valid-token mask, factors and residuals. The former accumulates N, $\sum \|x\|_1^2$ and $\sum \|A_0 x\|_1^2$ in one shared forward pass per batch, without centering or squaring mean norms; Direct adds the actual backward/optimizer response. Both fix scales before training and use the same five nominal-rate candidates and five seeds. A fixed-state common-rate multiplier cancels from the linear ratio, allowing initial scale reuse; later trajectories are not equated. Dual-Kaiming Init[AB] is excluded from the forward analytic comparison because it does not guarantee column orthogonality.

CorDA's covariance reference retains its larger 256-conversation construction statistics, not a weakened four-conversation substitute. Forward-moment and Direct access the same four-conversation subset. Forward/backward pass counts are not wall-clock or memory measurements. Forward-moment gains 0.63/0.64/0.78 points on PiSSA/MiLoRA/CorDA. All three Direct-minus-Forward intervals include zero (Table 32); paired information, not overlapping marginal bars, defines this comparison. The three Direct–Forward intervals do not establish equivalence because no equivalence margin was prespecified. The result compares usable information routes, not a universal ranking or a requirement for backward computation.

O.2 Layer correspondence at fixed information

Groups are defined by model, training seed, module type and input/output shape. Four attention types are never mixed; each Qwen group contains 28 layers. Starting from Direct's measured scales,

$$\bar{c}_g = \exp\left(\frac{1}{|g|}\sum_{\ell \in g} \log c_\ell\right), \qquad c_\ell^{\text{shuffle}} = c_{\pi_g(\ell)}, \tag{109}$$

Each within-group permutation π_g is drawn once from a fixed stream before evaluating outcomes, without requiring a derangement or choosing a favorable permutation. Shuffling preserves the full group multiset; sharing preserves its geometric mean. Every positive assigned scale acts reciprocally on that layer's Raw factors and preserves its product/residual. Both controls independently select among the same five rates. One permutation per training seed defines the reported protocol, not all possible assignments; final scores are evaluated separately from scale statistics. Direct-minus-sharing has a positive interval in both Qwen configurations; Direct-minus-shuffling does so only in MiLoRA. Sharing removes within-type variation; shuffling retains it but breaks layer correspondence. Neither identifies a globally optimal ratio or proves all layers need distinct scales.

Table 31: **Selected information and layer-allocation controls.** Five paired seeds; mean ± sample SD. Avg. is the equal-weight four-task percentage score; gain is relative to paired Raw. Reporting NLL and task scores are excluded from selection.

Configuration	Method	Selected η	Reporting NLL	Avg.	Gain
PiSSA-32	Raw	5×10^{-5}	0.89673±0.00247	33.49±0.25	+0.00±0.00
PiSSA-32	Spectrum-full	8×10^{-5}	0.89303±0.00233	34.28±0.31	+0.80±0.16
PiSSA-32	Forward-moment	8×10^{-5}	0.89376±0.00143	34.12±0.52	+0.63±0.35
PiSSA-32	Direct	8×10^{-5}	0.89381±0.00221	34.19±0.36	+0.70±0.14
PiSSA-32	Type-shared	8×10^{-5}	0.89456±0.00213	33.87±0.22	+0.38±0.13
PiSSA-32	Within-type shuffle	8×10^{-5}	0.89491±0.00168	34.00±0.32	+0.51±0.14
MiLoRA-64	Raw	13×10^{-5}	0.76839±0.00071	45.06±0.33	+0.00±0.00
MiLoRA-64	Spectrum-full	8×10^{-5}	0.76602±0.00093	45.60±0.38	+0.54±0.14
MiLoRA-64	Forward-moment	8×10^{-5}	0.76766±0.00090	45.70±0.36	+0.64±0.05
MiLoRA-64	Direct	8×10^{-5}	0.76650±0.00112	45.62±0.30	+0.55±0.06
MiLoRA-64	Type-shared	8×10^{-5}	0.76880±0.00145	45.34±0.42	+0.28±0.15
MiLoRA-64	Within-type shuffle	8×10^{-5}	0.76843±0.00130	45.15±0.37	+0.09±0.30
CorDA-64	Raw	8×10^{-5}	0.83948±0.00241	24.26±0.36	+0.00±0.00
CorDA-64	Covariance-aware	8×10^{-5}	0.83696±0.00170	24.51±0.33	+0.24±0.27
CorDA-64	Forward-moment	8×10^{-5}	0.83626±0.00152	25.04±0.47	+0.78±0.28
CorDA-64	Direct	8×10^{-5}	0.83593±0.00143	25.18±0.63	+0.92±0.31

Table 32: **Paired information and assignment contrasts.** First method minus second in task percentage points. Five paired training seeds and the reference-interval convention in Appendix G.3; settings are selected independently by validation.

Configuration	Contrast	Paired difference	95% interval
PiSSA-32	Direct − Forward-moment	$+0.073 \pm 0.393$	$[-0.415, +0.561]$
PiSSA-32	Forward-moment − Spectrum-full	-0.164 ± 0.373	$[-0.628, +0.299]$
PiSSA-32	Direct − Type-shared	$+0.326 \pm 0.219$	$[+0.054, +0.597]$
PiSSA-32	Direct − Within-type shuffle	$+0.193 \pm 0.207$	$[-0.064, +0.450]$
MiLoRA-64	Direct − Forward-moment	-0.081 ± 0.107	$[-0.214, +0.052]$
MiLoRA-64	Forward-moment − Spectrum-full	$+0.096 \pm 0.160$	$[-0.103, +0.295]$
MiLoRA-64	Direct − Type-shared	$+0.272 \pm 0.153$	$[+0.082, +0.462]$
MiLoRA-64	Direct − Within-type shuffle	$+0.468 \pm 0.331$	$[+0.057, +0.880]$
CorDA-64	Direct − Forward-moment	$+0.140 \pm 0.259$	$[-0.182, +0.461]$
CorDA-64	Forward-moment − Covariance-aware	$+0.536 \pm 0.378$	$[+0.067, +1.005]$

O.3 Frozen settings and a new paired seed group

Before this comparison, freeze each selected nominal rate and the 25-candidate dual-rate method's selected ratio. New seeds $611, 722, 833, 944, 1055$ are disjoint from the original group. No rate is reselected and no seed dropped. Calibration remains seed-specific; freezing hyperparameters does not freeze measured scales. Initialization and minibatch order are paired within seed. Report this group separately, never as a pooled post-selection ten-seed interval; it checks selected settings, not uncertainty of the full selection procedure or cross-dataset transfer. 4 of 4 Warm-minus-Spectrum-full or Warm-minus-dual-rate intervals include zero in this fixed-setting group. The earlier selected-setting ordering is not uniformly resolved by these new seeds. The stronger reference is the validation-selected 25-setting dual-rate grid, not whichever baseline scores best on tasks. Ideal-limit checks in Appendix N establish correction consistency, not a cause of downstream differences.

Table 33: **Fixed-setting confirmation group.** Five new paired seeds, with no new search. Reporting NLL and task gains over the new group's paired Raw are mean ± sample SD.

Configuration	Method	Reporting NLL	Paired task gain
PiSSA-32	Spectrum-full	0.89249±0.00125	+0.70±0.35
PiSSA-32	Forward-moment	0.89311±0.00224	+0.73±0.15
PiSSA-32	Direct	0.89364±0.00109	+0.88±0.10
PiSSA-32	Warm	0.89228±0.00099	+0.76±0.39
PiSSA-32	Tuned dual LR (25)	0.89244±0.00204	+0.77±0.30
MiLoRA-64	Spectrum-full	0.76514±0.00148	+0.52±0.35
MiLoRA-64	Forward-moment	0.76650±0.00149	+0.36±0.30
MiLoRA-64	Direct	0.76459±0.00111	+0.55±0.23
MiLoRA-64	Warm	0.76430±0.00100	+0.63±0.24
MiLoRA-64	Tuned dual LR (25)	0.76496±0.00139	+0.56±0.23

Table 34: **Fixed-setting paired contrasts.** First minus second, in task percentage points. The five new seeds are the replication unit; reference intervals are not multiplicity-adjusted.

Configuration	Contrast	Paired difference	95% interval
PiSSA-32	Direct − Forward-moment	$+0.152 \pm 0.100$	$[+0.028, +0.277]$
PiSSA-32	Warm − Spectrum-full	$+0.062 \pm 0.632$	$[-0.723, +0.848]$
PiSSA-32	Warm − Tuned dual LR (25)	-0.005 ± 0.359	$[-0.451, +0.441]$
MiLoRA-64	Direct − Forward-moment	$+0.190 \pm 0.375$	$[-0.275, +0.656]$
MiLoRA-64	Warm − Spectrum-full	$+0.109 \pm 0.323$	$[-0.292, +0.510]$
MiLoRA-64	Warm − Tuned dual LR (25)	$+0.070 \pm 0.433$	$[-0.468, +0.608]$