RETHINKING SUBSPACE SIMILARITY IN LOW-RANK ADAPTATION: WHAT GOVERNS LOCAL UPDATES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation is usually understood as optimization under a low-rank constraint, but its update map is not a projection: it weights a gradient by the spectrum the factors were built from. We reframe it as optimization under a spectrum-weighted map, and identify spectral allocation—how the spectrum is distributed across the two factors—as the variable governing early training, not the rank or the selected subspace. We test this reframing with paired upstream updates, shared-gradient probes, and natural-magnitude component interventions. A modified initialization can preserve its selected subspaces almost exactly while changing the weighted update map appreciably—a divergence we call the *subspace–update gap*. The distinction reaches training: two factorizations of one weight with identical initial effective weight, identical selected subspaces, and a shared learning rate produce ten-step updates lying $77.5°$–$88.0°$ apart, so spectral weighting changes the direction of adaptation, not only its speed. The gap is not an artifact of one spectral choice: cross-subspace coupling reproduces the response under principal-component selection but not under minor-component selection, where changes within the selected block dominate. The criterion is measurable with the same interventions and applies to any modification of the reference weights, including quantization and pruning.

1 INTRODUCTION

Low-rank adaptation makes fine-tuning efficient by optimizing factor matrices (Hu et al., 2022), and its parameterization determines how a gradient becomes a weight update (Hayou et al., 2024b; Wang et al., 2025c; Yen et al., 2025): two initializations can represent the same effective weights while mapping the same gradient to different updates. This is not a technicality. Low-rank adaptation is usually understood as optimization under a low-rank constraint. We reframe it as optimization under a spectrum-weighted map: the factorization is not a projection but weights the gradient by the spectrum it was built from before an update is formed, so the variable governing early training is how the spectrum is allocated across the two factors—not the rank and not the selected subspace alone. For weight-based spectral initialization, changes to the reference weights before initialization can therefore alter subsequent learning even when the initial function is preserved.

A natural explanation focuses on the selected singular subspaces: perturbations move these subspaces, changing the directions available to the low-rank update, as captured by classical subspace perturbation bounds (Wedin, 1972; Vu et al., 2021). Yet for balanced spectral factors under a common factor-SGD learning rate, the update map also contains spectral weighting within the selected subspaces. A change in this internal structure can alter the response without moving the subspaces themselves, so two adapters spanning the same subspace need not optimize alike. The unresolved question is therefore: *under input-induced weight perturbations, when does cross-boundary subspace motion approximately explain the local update response, and when do changes within the selected subspaces also matter?*

We investigate this question by separating the perturbation delivered to the weights from its effect on the initialized update map, decomposing it into selected, complementary, and cross blocks and testing their contributions under a shared gradient. In the principal-selection instance, PiSSA(Meng et al., 2024), cross retention approximately reproduces the full response difference despite its small share of perturbation energy; removal sharply reduces the response along the full-change direction.

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.

Is cross sufficiency a general consequence of subspace coupling, or does it depend on the selected components? Minor-component selection, instantiated by MiLoRA(Wang et al., 2025b), reveals a boundary of this explanation. In the evaluated settings, cross retention does not reproduce the weighted response as accurately. Selected-block interventions expose a contribution that subspace displacement alone cannot describe: they preserve the selected subspaces while producing an identifiable response change. Together, the interventions show that a shared cross-boundary coupling structure need not yield the same response attribution. Additional evidence bounds the scope of this account: block-energy-matched random controls reproduce the attribution but not the task-gradient magnitude, delivered perturbation ranks held-out responses far better than a static gap indicator, and a first-order model predicts held-out responses while matched-forward interventions reproduce the (small) ten-step separation. Our contributions are threefold:

- We separate cross-boundary subspace coupling from within-selected changes and characterize their effects on the spectrum-weighted update map.
- We establish contrasting response attributions under principal and minor selection, identifying when cross retention suffices and when selected-block changes dominate.
- We validate the explanation with matched random controls, held-out prediction, and matched-forward short-horizon adaptation.

Subspace similarity is widely treated as evidence that a modification has preserved what matters—after quantization, pruning, merging, or continued training. Our measurements show it is not sufficient: selected-block changes can alter the weighted update response while leaving the selected factor subspaces nearly unchanged—displacement ratios near 10^{-8} alongside responses of tens of percent (Section 4.3). We call this divergence the *subspace–update gap*: the distance between how much of the selected subspace a modification preserves and how much of its update behavior it preserves. The distinction reaches training: two factorizations of one weight with identical initial effective weight, selected subspaces, and learning rate produce cumulative updates that are nearly orthogonal after ten steps (Section 4.6). Measuring the weighted update map is therefore necessary to judge comparable adaptation, and Section 5 outlines how the same measurements order a modified checkpoint's layers.

2 Related Work

From low-rank parameterization to update geometry. LoRA reduces adaptation cost by optimizing a factorized weight increment (Hu et al., 2022), and this factorization also shapes optimization: a gradient with respect to the effective weight is mapped through the factors before producing an update. This mapping is reflected in work on unequal factor learning rates (Hayou et al., 2024b), rank-dependent scaling (Kalajdzievski, 2023), effective-gradient correction (Wang et al., 2025c), and invariance to equivalent factor transformations (Yen et al., 2025). Initialization shapes it too: LoRA-GA aligns the initial update with the full-finetuning gradient (Wang et al., 2024), and which factor starts at zero governs usable learning rates (Hayou et al., 2024a). Together these studies establish that low-rank adaptation depends on how an update is parameterized and optimized, beyond rank alone, but they do not ask how the mapping changes when the reference weight it is initialized from differs—the question we isolate.

From update geometry to spectral selection. Spectral initialization makes the choice of trainable directions explicit: principal and minor singular components provide contrasting choices, instantiated by PiSSA (Meng et al., 2024) and MiLoRA (Wang et al., 2025b). Adjacent designs vary that boundary—knowledge-aware truncation (Wang et al., 2025a), QR-orthonormal bases (Büyükakyüz, 2024), and principal-plus-residual splits (Chen & Tan, 2026)—while SVD-form parameterizations prune singular components by importance (Zhang et al., 2023). Which property drives the gains is contested: DoRA separates magnitude from direction (Liu et al., 2024), and spectral initialization's benefit has been attributed to magnitude amplification rather than the selected subspace (Zhang et al., 2025). The reference structure can also incorporate data—CorDA conditions decomposition on activation covariance (Yang et al., 2024), EVA initializes from activation directions (Paischer et al., 2025), and data-aware sensitivity is argued to select the subspace (Feng et al., 2026). These approaches turn initialization into a choice of structure, taking the reference weight as given, and

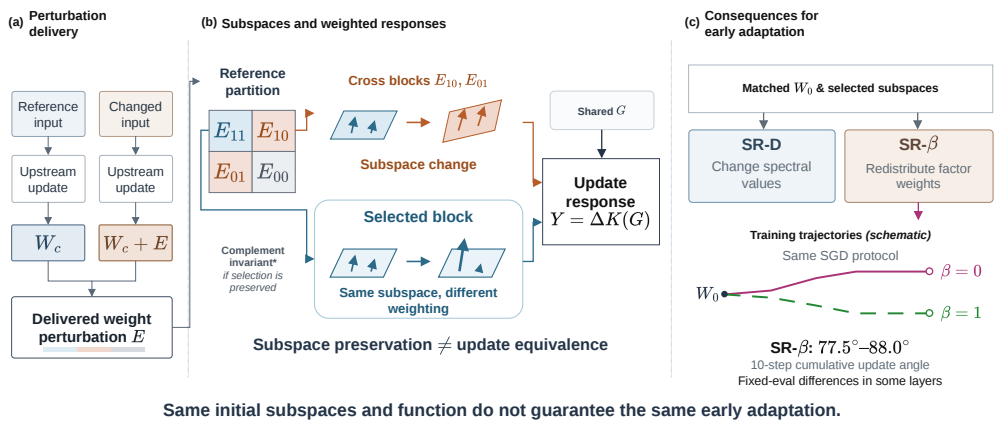

Figure 1: **From spectral perturbations to early adaptation.** **(a)** Paired updates deliver E. **(b)** Cross blocks can move selected subspaces; selected-block changes can alter the weighted response without moving them. **(c)** SR-D and SR-β fix the initial weight and subspaces. Schematic paths compare $\beta = 0$ and $\beta = 1$ within SR-β; angles are layer-wise seed medians of ten-step cumulative updates.

motivate the question of which parts of a changed reference control the update response; principal and minor selection provide concrete instances for testing it.

From spectral perturbation to response attribution. Classical singular-subspace perturbation bounds relate displacement to perturbation size and spectral separation (Wedin, 1972; Davis & Kahan, 1970); they bound displacement rather than attributing a response to a block. Perturbation expansions of truncated SVD characterize local changes and approximation errors (Vu et al., 2021), and gradient-flow analyses show the final error is governed by misalignment between pretrained and target singular spaces, which spectral initialization reduces (Xu et al., 2025). For the balanced factor-SGD update considered here, however, the gradient mapping depends on weighted spectral structure rather than selected projectors alone, so changes within a selected subspace can alter the response without moving it. Our component interventions measure when cross-boundary components reproduce the finite response and when selected-block changes matter; unlike influence-function analyses, which trace predictions to training points (Koh & Liang, 2017), our perturbation is an input to the analysis, not a quantity it estimates.

Where weight perturbations come from. Weight differences of the kind we study arise in several settings, and the object E is the same in each. Quantization: QLoRA uses quantized weights as the substrate (Dettmers et al., 2023), LoftQ fits initialization to the quantization residual (Li et al., 2024), and LQ-LoRA splits matrices into quantized and high-precision parts (Guo et al., 2024). Inputs: bounded embedding perturbations are standard in adversarial training (Miyato et al., 2017; Zhu et al., 2020; Jiang et al., 2020), and trigger-token poisoning survives downstream fine-tuning (Kurita et al., 2020; Yang et al., 2021). Weight differences as objects: task vectors subtract pretrained from fine-tuned weights (Ilharco et al., 2023), and backdoor vectors are the difference between poisoned and clean fine-tunes (Pawlak et al., 2025)—coinciding with our E. The spectral structure of such changes is also studied directly (Si et al., 2025; Shuttleworth et al., 2025; Puertolas Merenciano et al., 2026). These works establish that such perturbations occur and carry spectral structure; they do not ask how a perturbation acts through low-rank initialization—the question we pose.

3 Methodology

We investigate how input-induced weight perturbations change spectral initialization by decomposing the delivered perturbation in a fixed spectral frame, isolating its components through natural-magnitude interventions, and measuring their effects on selected subspaces and local updates (Figure 1); notation is in Appendix A.

3.1 A Fixed Selected–Complement Partition

Paired upstream updates start from the same checkpoint and differ only in the prescribed input perturbation, producing $E = W_\delta - W_c$. For a direct weight-space decomposition, we write $W_c = U\Sigma V^\top$ and let S select r singular components. The corresponding left and right projectors are $P = U_S U_S^\top$ and $Q = V_S V_S^\top$. They define four perturbation blocks:

$$\begin{aligned} E_{11} &= PEQ, & E_{10} &= PE(I-Q), \\ E_{01} &= (I-P)EQ, & E_{00} &= (I-P)E(I-Q). \end{aligned} \tag{1}$$

The indices 1 and 0 denote selected and complementary directions, respectively. These mutually orthogonal blocks sum to E.

We distinguish components that connect the two subspaces from those confined within them: $E_{\text{cross}} = E_{10} + E_{01}$ and $E_{\text{within}} = E_{11} + E_{00}$. All interventions use this fixed reference partition and preserve the original component magnitudes, then recompute the spectral initialization of each intervened matrix. Selecting principal or minor singular components instantiates the construction with PiSSA or MiLoRA, respectively (Meng et al., 2024; Wang et al., 2025b).

Context-dependent spectral coordinates. The CorDA context-dependent construction is specified in Appendix B.6.

3.2 Spectral Initialization and Local Update Response

To compare the construction through a common response object, consider $W_{\text{eff}} = W_{\text{res}} + sBA$, where W_{res} is frozen, $s > 0$, and $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$ are the initialized trainable factors. Under simultaneous factor SGD with common learning rate η, a matrix gradient G induces

$$\Delta W_{\text{eff}} = -\eta s^2 \mathcal{K}_{B,A}(G) + O(\eta^2), \qquad \mathcal{K}_{B,A}(G) = BB^\top G + GA^\top A. \tag{2}$$

Thus the gradient-to-update map depends on the factor Gram matrices, not only on the selected projectors.

For a direct balanced spectral construction, the factors $B(W) = s^{-1/2} U_S \Sigma_S^{1/2}$ and $A(W) = s^{-1/2}\Sigma_S^{1/2} V_S^\top$ give Gram matrices $s^{-1} U_S \Sigma_S U_S^\top$ and $s^{-1} V_S \Sigma_S V_S^\top$ (Appendix B). For the context-dependent construction, we use the actual factors obtained after mapping an intervention back from the coordinates in Equation 19. Equation 2 therefore provides a common response definition without requiring all constructions to share the same factorization coordinates or factor scaling.

For each intervention F, let (B_F, A_F) denote the factors reinitialized from the intervened matrix and let (B_0, A_0) denote the reference factors. We measure $Y_F = \mathcal{K}_{B_F,A_F}(G) - \mathcal{K}_{B_0,A_0}(G)$, with $Y = Y_{\text{full}}$. Reinitializing the factors after each intervention captures the component's effect on the spectral initialization, while holding G fixed isolates the change in how the initialization geometry maps the same gradient into a weight update.

3.3 Subspace Coupling and Within-Subspace Changes

The reference partition exposes two mechanisms that can change the weighted update map, analyzed with a fixed factor construction and scale and spectral separation that preserves selection of the reference block.

Cross-boundary coupling. The cross blocks connect selected and complementary directions in reference spectral coordinates. They can therefore change the selected projectors and the factor Gram matrices governing the update response. By contrast, under a fixed construction and preserved spectral selection, a pure complementary-block perturbation leaves the selected factors' Gram matrices unchanged, $\mathcal{K}_{B_{00},A_{00}} = \mathcal{K}_{B_0,A_0}$, as follows from the block-diagonal structure when the selection is preserved.

Changes within the selected block. A pure selected-block intervention preserves the selected left and right subspaces while changing their internal spectral structure. The resulting changes in the

factor Gram matrices can produce an update response even when the selected projectors remain fixed (Appendix B.5). Thus subspace motion and weighted response capture different aspects of the perturbation.

For the direct balanced spectral construction, the same decomposition yields a reference-state predictor. Differentiating the spectral map in E gives

$$\begin{aligned}\widehat{Y}(E;G) &= D\mathcal{K}_{W_c}[E](G) \\ &= D\mathcal{K}_{W_c}[E_{\mathrm{cross}}](G) + D\mathcal{K}_{W_c}[E_{11}](G).\end{aligned} \tag{3}$$

The complementary-block derivative vanishes. The full predictor includes both terms, whereas the cross-only predictor retains the first. Both use the reference spectrum, the observed perturbation, and the specified gradient, without computing a perturbed SVD.

3.4 Component Interventions and Mechanism Validation

We refer to the decomposition, intervention, and response-measurement protocol used throughout this section as SplitProbe (Subspace–Spectral Split Probing). We compare five interventions—full (E), cross (E_{cross}), within (E_{within}), selected-only (E_{11}), and complement-only (E_{00})— formed in the direct weight-space partition or, for the context-dependent construction, in the decomposition coordinates and mapped back to weight space. Cross retention tests whether boundary coupling reproduces the full response; cross removal measures what remains.

Let P_F, Q_F be the selected projectors obtained from the actual factors after intervention. We measure their displacement separately on the left and right: $d_L(F) = \|P_F - P\|_F$ and $d_R(F) = \|Q_F - Q\|_F$. For identifiable full responses, response attribution is quantified by

$$\begin{aligned} e_{\mathrm{cross}} &= \frac{\|Y_{\mathrm{cross}} - Y\|_F}{\|Y\|_F}, \quad & a_F &= \frac{\|Y_F\|_F}{\|Y\|_F}, \\ p_F &= \frac{\langle Y_F, Y\rangle_F}{\|Y\|_F^2}, & i_{\mathrm{cw}} &= \frac{\|Y - Y_{\mathrm{cross}} - Y_{\mathrm{within}}\|_F}{\|Y\|_F}. \end{aligned} \tag{4}$$

These measure vector reproduction, relative response magnitude, signed projection onto the full response, and finite interaction, respectively.

Small e_{cross} supports approximate cross sufficiency, selected-only tests whether a response can remain with unchanged subspaces, and full versus cross-only prediction measures the selected-block contribution to accuracy.

For short-horizon validation, each frame is initialized with a frozen residual $W_{\mathrm{res}}^{(F)} = M - sB_FA_F$ so all frames share the same initial effective weight M, and their subsequent updates evolve separately (Appendix F).

4 Experimental Evaluation

Our experiments test whether retaining cross components at their observed magnitudes is sufficient to reproduce the full response, and whether selected-block changes produce appreciable responses without appreciable subspace motion; conditional prediction and matched-forward adaptation further assess quantitative accuracy and short-horizon effects.

4.1 Experimental Setup

We use SST-2-adapted DeBERTa-v3-base and RoBERTa-base checkpoints. Paired upstream updates deliver $E = W_\delta - W_c$ from continuous perturbations at the embeddings of inserted trigger tokens. Both branches retain the same trigger, and the reference branch sets the prescribed continuous perturbation to zero. The construction uses random-sign and layer-targeted spectral perturbations; component interventions retain their observed magnitudes. The input construction and optimization protocol are specified in Appendix D.

The main analyses use rank eight and ten seeds: a natural-response panel of two DeBERTa and six RoBERTa layers, a construction-comparison set of ten DeBERTa and nine RoBERTa layers under

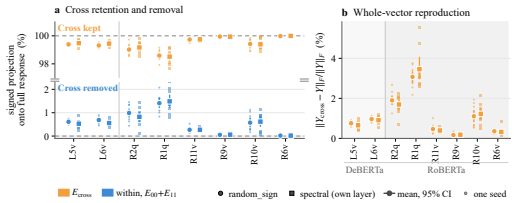

Figure 2: Natural-magnitude component attribution under shared task gradients and principal selection. (a) Signed projection onto the full response after cross retention or removal. (b) Whole-vector error of cross retention. Large markers show group means and 95% confidence intervals.

Table 1: Natural-magnitude component interventions under shared task gradients and principal selection. Entries are ranges of layer–family group means across ten seeds; errors, norm ratios, and interaction residuals are percentages, while signed projections are dimensionless.

	DeBERTa	RoBERTa
e_{cross} (%)	0.658–0.964	0.169–3.458
a_{within} (%)	0.662–0.964	0.165–3.461
p_{cross}	0.9931–0.9948	0.9851–0.9999
p_{within}	0.0053–0.0069	0.0002–0.0149
Cross energy (%)	2.639–3.505	1.189–6.728
i_{cw} (%)	0.0111–0.0877	0.0193–0.1602

isotropic gradients, and a six-layer subset for short-horizon validation, with held-out prediction on eight further layers. The construction comparison evaluates principal and minor selection (PiSSA and MiLoRA) on the same layer set, and we additionally evaluate CorDA context-dependent constructions (IPM and KPM modes, fixed balanced factor scaling).

Ratios are computed per measurement, averaged over repeated draws and gradient probes within each seed, and summarized across seeds; the natural-response and validation results report ranges of group means, and the construction comparison reports layer-wise medians. Range endpoints describe variation across conditions, not confidence intervals, and only identifiable ratios enter the summaries; thresholds, valid seed counts, and bootstrap intervals are in the appendix.

4.2 When Are Cross Components Approximately Sufficient?

We first test whether boundary-crossing components are sufficient at the delivered magnitudes. Under shared task gradients in the principal-selection natural-response panel, retaining only E_{cross} reproduces the full response within 0.658%–0.964% (DeBERTa) and 0.169%–3.458% (RoBERTa) group-mean error, while removing it leaves only 0.662%–0.964% and 0.165%–3.461% of the full response norm (Figure 2; Table 1). The retention–removal contrast thus holds for the response vector and its magnitude, not only its projection. Together, these interventions establish approximate cross sufficiency at the delivered magnitudes: retaining cross reproduces the full response while removing it leaves a remaining response norm a_{within} of at most 3.5% of the full response—a magnitude contrast disproportionate to the cross components' perturbation energy, with a small cross–within interaction. The synthetic-gradient panel reaches a maximum group-mean error of 8.643%. This sufficiency does not follow from complementary-block invariance, because selected-block changes and finite interactions remain possible.

Shared attribution does not fix response magnitude. Block-energy-matched random controls reproduce the broad attribution but not the larger input-induced task responses (Appendix E.4).

4.3 Subspace Motion Does Not Determine Weighted Response

We next ask whether this attribution extends across spectral constructions. Under the isotropic-gradient panel, layer-wise median cross-reproduction errors range from 1.25% to 9.64% under principal selection but 21.25% to 78.75% under minor selection (Table 2). Cross-boundary coupling can therefore be structurally present without being approximately sufficient for the weighted response.

To identify the missing channel, we isolate the selected block: under minor selection its subspace-displacement ratios are no larger than 1.15×10^{-8} while its response norm reaches 21.25%–78.74% of the full response, so selected-block changes can contribute substantially with negligible subspace displacement. Figure 3 contrasts selected-block subspace displacement with its weighted response on the same units. The distinction also appears in the context-dependent CorDA constructions, where selected-block displacement stays small in the actual factor subspaces while the weighted

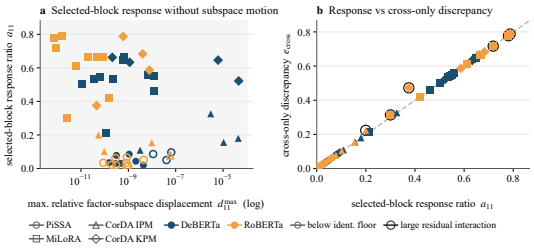

Figure 3: Selected-block response without appreciable subspace motion. (a) Response ratio versus relative factor-subspace displacement. (b) Selected-block response ratio versus cross-only reproduction error under balanced factor scaling.

Table 2: Selected-block displacement and weighted response under random-sign perturbations and isotropic gradients. Displacement entries are the largest layer-wise medians; response entries are ranges of layer-wise medians in percent. IPM and KPM denote the two CorDA context modes.

	PiSSA	MiLoRA	IPM	KPM
r_L^{11}.	4.12×10^{-8}	1.10×10^{-8}	1.59×10^{-8}	3.86×10^{-5}
r_R^{11}.	6.41×10^{-8}	1.15×10^{-8}	4.03×10^{-5}	4.33×10^{-5}
a_{11} (%)	1.10–9.64	21.25–78.74	7.85–32.62	37.49–78.73
e_{cross} (%)	1.25–9.64	21.25–78.75	7.85–32.62	47.25–78.73

response stays appreciable (Table 2); it is therefore not confined to an orthogonal partition of the original weight matrix.

These interventions establish a response channel that does not require appreciable subspace motion, though they do not show that the cross intervention reproduces the direction of the full projector change. Our claim is therefore narrowly interventional: selected-block changes can alter the weighted update response while leaving the selected factor subspaces nearly unchanged.

Finite interactions further determine how the isolated responses combine (Appendix I.3).

4.4 Conditional Prediction of Finite Responses

A mechanism account should constrain response magnitudes and directions, rather than only identify active components. We therefore evaluate the reference-state first-order predictor on eight held-out layers in the principal-selection setting. Given the observed perturbation and task gradient, the full predictor achieves layer–family group-mean relative errors of 0.00014%–0.220% in DeBERTa and 0.0036%–0.106% in RoBERTa, without computing a perturbed SVD.

Cross-only first-order errors are larger: 0.342%–8.13% in DeBERTa and 0.181%–4.70% in RoBERTa. Approximate control of the dominant finite response therefore does not make the remaining first-order contribution dispensable for accurate prediction, and the comparison supports the value of retaining the complete derivative. Details are in Appendix G. Appendix Figure 6 shows the held-out errors for the full and cross-only predictors across both perturbation families.

4.5 Matched-Forward Short-Horizon Validation

We finally test whether the component attribution remains informative once gradients evolve during adaptation. For each target $M \in \{W_c, W_\delta\}$, residual compensation gives the reference, full, cross, and within frames identical initial effective weights, isolating an initialization-geometry difference from a difference in the initial forward function; their factor-SGD updates then evolve separately for ten steps.

Across the six evaluated principal-selection layers, cross retention approximately reproduces the full-intervention trajectory separation, with group-mean reproduction errors below 3.6% at steps one and ten. The separation itself is small: at step ten, its group-mean magnitude relative to the reference update is 0.0076%–0.0609% in DeBERTa and 0.0054%–0.251% in RoBERTa, with no predicted-class disagreements on the fixed 32-example batch. The component explanation thus extends beyond the initial shared-gradient comparison, while the measured effect remains a small trajectory difference. Appendix Figure 5 presents both the trajectory-reproduction error and the intervention magnitude relative to the reference update. Thresholds, record-level filtering, and effective seed counts are reported in Appendix F.

Table 3: Factor reweighting and early adaptation. $T_{10} = \|M_{10}^0 - M_{10}^1\|_F / \|M_{10}^1 - W_0\|_F$; θ is the seed-median angle between the two cumulative updates. $\Delta\ell$ is the mean paired fixed-evaluation loss $\ell(\beta = 0) - \ell(\beta = 1)$ at step 10; negative favors $\beta = 0$ *under this protocol*. Each row contains ten paired seeds. Prediction disagreements sum repeated evaluations of the same 128 samples over the ten seeds and are not counts of independent test examples.

Layer	T_{10}	θ	$\Delta\ell$	Disagreements
DeB L5v	1.539	80.2°	-4.69×10^{-4}	0/1280
DeB L6v	1.250	79.9°	-6.73×10^{-5}	0/1280
RoB R10v	1.256	77.5°	$+1.65 \times 10^{-3}$	0/1280
RoB R1q	2.174	88.0°	-4.39×10^{-2}	10/1280
RoB R2q	2.610	87.8°	-5.69×10^{-2}	99/1280
RoB R6v	2.178	79.5°	-2.81×10^{-2}	54/1280

4.6 Does Spectral Weighting Affect Early Adaptation?

The selected-block interventions above show that nearly identical selected subspaces can induce different weighted update responses. We next ask whether this distinction reaches actual adaptation. We use two controls that hold the initial effective weight and the selected left and right subspaces fixed: one changes the selected singular values, and the other redistributes unchanged singular values between the two trainable factors. Both use the same single-layer training protocol, batches, seeds, and initial task gradient across arms. These experiments isolate two related sources of weighting without equating a change in the spectrum with a change in factor parameterization.

Redistributing spectral weights between factors. For the principal selected components of $W_0 = U\Sigma V^\top$, let

$$B_\beta = s^{-1/2} U_r \Sigma_r^{1-\beta}, \qquad A_\beta = s^{-1/2} \Sigma_r^\beta V_r^\top, \qquad R = W_0 - s B_\beta A_\beta, \tag{5}$$

where $\beta \in \{0, \frac{1}{2}, 1\}$. Thus every arm has the same $B_\beta A_\beta$, frozen residual R, selected singular values, selected subspaces, and effective weight $M_0 = R + sB_\beta A_\beta = W_0$. The balanced initialization is $\beta = \frac{1}{2}$. Only the distribution of spectral weights between the two factors, and consequently the SGD-induced map $\mathcal{K}_\beta(G) = B_\beta B_\beta^\top G + G A_\beta^\top A_\beta$, varies.

With a shared initial gradient G_0 and ordinary factor SGD, the effective-weight difference after one step obeys

$$M_1^\beta - M_1^{\beta'} = -\eta s^2 \big[\mathcal{K}_\beta(G_0) - \mathcal{K}_{\beta'}(G_0)\big]. \tag{6}$$

This is an exact *difference* for this construction, because the finite-step product terms cancel; the algebraic self-check is reported in Appendix I.4.

We train ranks of eight with shared $\eta = 10^{-2}$ for ten steps on DeBERTa (L5v, L6v) and RoBERTa (R10v, R1q, R2q, R6v), using ten paired seeds per layer. Table 3 reports the prespecified $\beta = 0$ versus $\beta = 1$ contrast. Across the six layers, the ten-step distance between effective weights is 1.25–2.61 times the $\beta = 1$ update from W_0. To distinguish an endpoint speed change from a directional change, we also measure the angle between the two cumulative updates $d_\beta = M_{10}^\beta - W_0$. Their norms and pairwise distance determine this angle exactly by

$$\cos\theta = \frac{\|d_0\|_F^2 + \|d_1\|_F^2 - \|d_0 - d_1\|_F^2}{2\|d_0\|_F \|d_1\|_F}. \tag{7}$$

The layer-wise seed medians are 77.5°–88.0°. Consequently, the *observed* ten-step updates are not positive scalar multiples of one another. The best scalar-alignment residuals of their initial update operators, 0.969–0.999 across layers, give a corresponding first-step diagnosis. Neither measurement establishes what would happen after retraining the arms with separately tuned learning rates.

The fixed-evaluation results show a task-level consequence in some of the tested layers, but no universally preferable factor split: the paired loss difference favors $\beta = 1$ at R10v, whereas R6v changes sign across seeds (seven negative, three positive) even though its cumulative update directions remain separated. The differences at R1q and R2q are larger and include prediction disagreements. All evaluations are on the same fixed 128-example batch per run; repeated seeds should not be interpreted as independent test examples.

Changing the selected singular values. The complementary SR-D intervention and its scaling controls are reported in Appendix I.4.

Together, the two interventions establish a training-side consequence of weighted spectral initialization: identical initial functions and selected subspaces need not imply identical early adaptation. They also separate two mechanisms with different effect scales—changing selected singular values and redistributing unchanged values between factors. These experiments use single-layer SGD, ten steps, and one common learning rate. They do not identify an optimal β, establish a benefit after separately tuned learning rates or full-budget training, or validate a layer-recovery rule for quantization.

5 Discussion

Our findings distinguish preserving a selected subspace from preserving how gradients act within it: cross retention approximately explains the response under principal selection, while selected-block interventions reveal appreciable responses with negligible subspace displacement elsewhere. The *subspace–update gap* is thus not closed by preserving the selected subspaces.

Implications. Low-rank adaptation is usually optimized as if the rank were the only spectral variable. Our results suggest otherwise: the spectral allocation across the factors, not the rank or the selected subspace, governs how a gradient becomes an update. This offers a diagnostic for spectral initialization after quantization or pruning: layers can be ranked by how much their weighted update response changes under the modification, and this ranking need not agree with one based on reconstruction error. SplitProbe reuses these measurements without additional training. The input-induced setting also raises a security question— whether manipulated upstream data can alter adaptation geometry—whose use for detection or defense requires attack-specific validation. **Consequences for training.** Two factorizations that share their initial effective weight, selected subspaces, and learning rate still diverge: at rank eight and ten SGD steps, the cumulative updates of the $\beta = 0$ and $\beta = 1$ arms lie $77.5°$–$88.0°$ apart, so they differ in direction rather than only in speed (Section 4.6). Changing the selected singular values while holding the subspaces fixed moves the first step by 0.28%–0.40% of the reference update. **Limitations.** Our evidence concerns two classification-adapted encoders, with the main component analyses at rank eight. Cross-construction comparisons use isotropic gradients and specified factor scalings, so the principal–minor contrast has not been reproduced under task gradients; conditional prediction and ten-step single-layer SGD validation cover the principal-selection setting only, and prediction is conditional on the observed perturbation and gradient. Records below the identifiability floor are excluded from the affected medians without quantifying that exclusion, and the measured trajectory effects remain small, so their consequences for long-term training and task behavior are not established.

6 Conclusion and Future Work

Natural-magnitude interventions show that cross-boundary coupling can approximately explain the local update response of spectral low-rank initialization, but its sufficiency depends on the spectral construction: internal selected-block changes can contribute substantially without appreciable subspace motion. The contribution is a reframing of low-rank adaptation: the update map is spectrum-weighted, and the variable governing early training is how the spectrum is allocated across the factors, not the rank or the selected subspace. Spectral initialization is commonly judged by subspace preservation, yet a modification can leave the selected subspaces nearly unchanged while still altering how a gradient becomes an update—the *subspace–update gap* is not closed by preservation. Where weights are modified—by quantization, pruning, merging, or upstream data—what must be checked is the weighted update map, which SplitProbe makes measurable without additional training. A training-side consequence is established in Section 4.6: two factorizations of one weight with identical effective weight, subspaces, and learning rate produce cumulative updates $77.5°$–$88.0°$ apart after ten SGD steps. Two directions follow: the same measurements can rank the layers of a quantized checkpoint by weighted-response change, to be compared with reconstruction error before recovery, and extending the analysis to multi-layer adaptation and optimizer-dependent maps would show whether local response differences accumulate into task-level effects.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing of the manuscript and for drafting portions of the experimental-protocol documentation. We have not used generative AI tools to generate experimental data, analyses, figures, or quantitative claims, and no LLM-generated experimental artifacts appear in this manuscript. We reviewed all AI-assisted text for correctness and consistency with our code and results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, new datasets, or deployed systems.

REPRODUCIBILITY STATEMENT

Code and configuration for reproducing all experiments will be released as an anonymous repository upon acceptance.

REFERENCES

Kerim Büyükakyüz. OLoRA: Orthonormal low-rank adaptation of large language models. *arXiv preprint arXiv:2406.01775*, 2024. URL https://arxiv.org/abs/2406.01775.

Lixian Chen and Jianhong Tan. HiP-LoRA: Budgeted spectral plasticity for robust low-rank adaptation. *arXiv preprint arXiv:2604.17751*, 2026. URL https://arxiv.org/abs/2604.17751.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.14314.

Zhi-Quan Feng, Ying-Jia Lin, and Hung-Yu Kao. Learning in the fisher subspace: A guided initialization for LoRA fine-tuning. *arXiv preprint arXiv:2605.01046*, 2026. URL https://arxiv.org/abs/2605.01046.

Han Guo, Philip Greengard, Eric P. Xing, and Yoon Kim. LQ-LoRA: Low-rank plus quantized matrix decomposition for efficient language model finetuning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.12023.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, 2024a. URL https://arxiv.org/abs/2406.08447.

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.

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://arxiv.org/abs/2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://aclanthology.org/2020.acl-main.197/.

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.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2793–2806, 2020. URL https://aclanthology.org/2020.acl-main.249/.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.08659.

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 *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3d5CIRG1n2.

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://arxiv.org/abs/2404.02948.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1605.07725.

Fabian Paischer, Lukas Hauzenberger, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, and Sepp Hochreiter. Parameter efficient fine-tuning via explained variance adaptation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.07170.

Stanisław Pawlak, Jan Dubiński, Daniel Marczak, and Bartłomiej Twardowski. Backdoor vectors: a task arithmetic view on backdoor attacks and defenses. *arXiv preprint arXiv:2510.08016*, 2025. URL https://arxiv.org/abs/2510.08016.

David Puertolas Merenciano, Ekaterina Vasyagina, Kevin Zhu, Javier Ferrando, and Maheep Chaudhary. Weight space detection of backdoors in LoRA adapters. *arXiv preprint arXiv:2602.15195*, 2026. URL https://arxiv.org/abs/2602.15195.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.21228.

Chongjie Si, Xuankun Yang, Muqing Liu, Yadao Wang, Xiaokang Yang, Wenbo Su, Bo Zheng, and Wei Shen. Weight spectra induced efficient model adaptation. *arXiv preprint arXiv:2505.23099*, 2025. URL https://arxiv.org/abs/2505.23099.

Bart Vandereycken. Low-rank matrix completion by Riemannian optimization—extended version. *arXiv preprint arXiv:1209.3834*, 2012. URL https://arxiv.org/abs/1209.3834.

Trung Vu, Evgenia Chunikhina, and Raviv Raich. Perturbation expansions and error bounds for the truncated singular value decomposition. *Linear Algebra and its Applications*, 627:94–139, 2021. doi: 10.1016/j.laa.2021.05.020. URL https://arxiv.org/abs/2009.07542.

Fan Wang, Juyong Jiang, Chansung Park, Sunghun Kim, and Jing Tang. KaSA: Knowledge-aware singular-value adaptation of large language models. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2412.06071.

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 (NAACL)*, 2025b. URL https://arxiv.org/abs/2406.09044.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2407.05000.

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*, 2025c. URL https://openreview.net/forum?id=gTwRMU3lJ5.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972. doi: 10.1007/BF01932678.

Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and René Vidal. Understanding the learning dynamics of LoRA: A gradient flow perspective on low-rank adaptation in matrix factorization. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=hphdX8WlcT.

Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2048–2058, 2021. URL https://aclanthology.org/2021.naacl-main.165/.

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://arxiv.org/abs/2406.05223.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Sai Surya Duvvuri, Inderjit S. 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://arxiv.org/abs/2410.20625.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2303.10512.

Zicheng Zhang, Haoran Li, Yifeng Zhang, Guoqiang Gong, Jiaxing Wang, Pengzhang Liu, Qixia Jiang, and Junxing Hu. The primacy of magnitude in low-rank adaptation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.06558.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. FreeLB: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.11764.

A NOTATION

Table 4 collects the symbols used in this paper.

Table 4: Notation. All quantities are evaluated at a single target matrix of a task-adapted checkpoint.

Symbol	Meaning
W_c	weight after the reference upstream update
W_δ	weight after the perturbed upstream update
E	delivered perturbation $W_\delta - W_c$
G	shared matrix gradient
B_0, A_0	reference spectral factors
$\mathcal{K}_{B,A}(G)$	gradient-to-update map induced by factors B, A
Y	response difference under the full perturbation
$E_{11}, E_{10}, E_{01}, E_{00}$	blocks of E in reference spectral coordinates
$E_{\text{cross}}, E_{\text{within}}$	cross and within components of E
U_S, V_S	selected reference singular subspaces; $\perp$ their complements
p_F, e_F	signed projection and reproduction error of intervention F
$G_{\text{iso}}, G_{00}, G_{10}, G_{01}$	probe gradients: isotropic and block-restricted probes
M	common initial target in the matched-forward experiment
q	block-energy-matched random-to-input response ratio
i_{cw}	cross–within interaction residual
s, η	adapter scaling and factor learning rate ($s = 1$ in all experiments)
$S_K(E, G)$	normalized sensitivity $\|Y\|_F/(\|E\|_F\|G\|_F)$
$Z_t^{(j)}$, a_t	trajectory separation and relative effect scale

B Operator and Structural Derivations

For a direct balanced spectral construction $W = U\Sigma V^\top$, the factors are

$$B(W) = s^{-1/2} U_S \Sigma_S^{1/2}, \qquad A(W) = s^{-1/2} \Sigma_S^{1/2} V_S^\top. \tag{8}$$

Their Gram matrices are

$$\begin{aligned} B(W)B(W)^\top &= s^{-1} U_S \Sigma_S U_S^\top, \\ A(W)^\top A(W) &= s^{-1} V_S \Sigma_S V_S^\top. \end{aligned} \tag{9}$$

Throughout this section, $s > 0$ and the selected rank r are fixed. The factors satisfy $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$, with effective weight $W_{\text{eff}} = W_{\text{res}} + sBA$. The residual is frozen.

B.1 Exact Simultaneous Factor-SGD Update

Let $G = \nabla_{W_{\text{eff}}} \mathcal{L}$ be evaluated before the update. By the chain rule,

$$\nabla_B \mathcal{L} = sGA^\top, \qquad \nabla_A \mathcal{L} = sB^\top G.$$

With a common learning rate η, simultaneous SGD updates give

$$B^+ = B - \eta s G A^\top, \qquad A^+ = A - \eta s B^\top G.$$

Consequently,

$$\begin{aligned} W_{\text{eff}}^+ - W_{\text{eff}} &= s(B^+A^+ - BA) \\ &= -\eta s^2 \left(BB^\top G + GA^\top A\right) + \eta^2 s^3 GA^\top B^\top G. \end{aligned} \tag{10}$$

This identity gives the local operator

$$\mathcal{K}_{B,A}(G) = BB^\top G + GA^\top A.$$

The first-order expression in the main text drops only the displayed quadratic term. Both factor gradients are evaluated at the same pre-update state; the identity does not describe sequential factor updates or adaptive optimizers.

B.2 Operator at Balanced Spectral Initialization

Write

$$W_c = U_{c,r}\Sigma_{c,r}V_{c,r}^\top + U_{c,\perp}C_cV_{c,\perp}^\top,$$

where $[U_{c,r}, U_{c,\perp}]$ and $[V_{c,r}, V_{c,\perp}]$ are complete orthogonal bases. The matrix C_c may be rectangular. For the balanced initialization,

$$B_0 = s^{-1/2}U_{c,r}\Sigma_{c,r}^{1/2}, \qquad A_0 = s^{-1/2}\Sigma_{c,r}^{1/2}V_{c,r}^\top.$$

Hence

$$\begin{aligned} B_0B_0^\top &= s^{-1}U_{c,r}\Sigma_{c,r}U_{c,r}^\top, \\ A_0^\top A_0 &= s^{-1}V_{c,r}\Sigma_{c,r}V_{c,r}^\top. \end{aligned} \tag{11}$$

These weighted spectral maps determine the initialized operator. Signs of individual singular vectors do not affect either map. Orthogonal changes of basis within an equal-singular-value cluster also leave the maps unchanged. Substituting into the definition of $\mathcal{K}$,

$$\mathcal{K}_{B,A}(G) = s^{-1}\big(U_{c,r}\Sigma_{c,r}U_{c,r}^\top G + G\,V_{c,r}\Sigma_{c,r}V_{c,r}^\top\big),$$

so $\mathcal{K}_W$ depends on W only through its leading singular triple $(U_{c,r}, \Sigma_{c,r}, V_{c,r})$. This is the premise of the invariance statement below: a perturbation that leaves the leading triple unchanged leaves the operator unchanged as well, and a perturbation that changes it can change the operator.

B.3 Complementary-Block Invariance

Suppose $\sigma_r(W_c) > 0$ and

$$E_{00} = U_{c,\perp}DV_{c,\perp}^\top.$$

If

$$\|C_c + D\|_2 < \sigma_r(W_c), \tag{12}$$

then

$$\mathcal{K}_{W_c+E_{00}}(G) = \mathcal{K}_{W_c}(G) \quad \text{for every } G. \tag{13}$$

Proof. In the reference orthogonal bases,

$$[U_{c,r}, U_{c,\perp}]^\top(W_c + E_{00})[V_{c,r}, V_{c,\perp}] = \begin{bmatrix}\Sigma_{c,r} & 0\\ 0 & C_c + D\end{bmatrix}.$$

The selected block is unchanged. Condition equation 12 ensures that every singular value of the complementary block is smaller than the smallest selected singular value. The leading weighted left and right spectral maps therefore remain those in Equation equation 11, which proves the identity.

A sufficient condition is

$$\|E_{00}\|_2 < \sigma_r(W_c) - \sigma_{r+1}(W_c), \tag{14}$$

because $\|C_c + D\|_2 \le \sigma_{r+1}(W_c) + \|E_{00}\|_2$. If a complementary singular value enters the selected set, this argument no longer applies. The leading triple is then reordered, and both the selected and the complementary block contribute to the operator. We do not characterize that regime here. Two points about the condition are worth stating plainly. First, it is a condition on the perturbed matrix: it asks that the perturbed complementary singular values stay below the perturbed selected ones, which is not implied by the ordering of the reference spectrum alone. Second, whether the perturbations delivered in our experiments satisfy it is not established by this analysis; we report the delivered magnitudes, and mapping the dose at which the condition fails is left open. Where the condition fails, a complementary-block perturbation is expected to acquire a first-order contribution through precisely the divided differences that vanish under the ordering assumption (Section C), and we make no quantitative claim about the size of that contribution.

B.4 Fixed Reference Tangent and Normal Spaces

At the rank-r matrix $U_{c,r}\Sigma_{c,r}V_{c,r}^\top$, with $\Sigma_{c,r}$ invertible, the tangent space is

$$T_c = \left\{U_{c,r}X^\top + ZV_{c,r}^\top : X \in \mathbb{R}^{n\times r}, Z \in \mathbb{R}^{m\times r}\right\}.$$

Using $P_U = U_{c,r}U_{c,r}^\top$ and $P_V = V_{c,r}V_{c,r}^\top$, its orthogonal projections are

$$\begin{aligned}\Pi_{T_c}(H) &= P_U H + HP_V - P_U H P_V,\\ \Pi_{N_c}(H) &= P_U^\perp H P_V^\perp.\end{aligned} \tag{15}$$

Thus the selected block, 10, and 01 blocks are tangent, and the 00 block is normal (Vandereycken, 2012).

For $G_{00} = P_U^\perp G P_V^\perp$,

$$B_0 B_0^\top G_{00} = 0, \qquad G_{00}A_0^\top A_0 = 0,$$

so $\mathcal{K}_{W_c}(G_{00}) = 0$. Perturbed selected directions can acquire reference-complement components, allowing the perturbed operator to respond to the same G_{00}. This is a possibility, not a guarantee of a nonzero response for every perturbation and gradient.

All comparisons use the fixed reference projections. A difference between two tangent updates need not belong to either branch's tangent space.

B.5 Selected-Block Changes and Subspace Preservation

Consider the reference decomposition

$$W_c = U_S\Sigma_S V_S^\top + U_\perp C_\perp V_\perp^\top,$$

where S contains the selected singular components and Σ_S has positive diagonal entries. The columns of $U_\perp$ and $V_\perp$ complete the corresponding orthonormal bases. For the selected-block intervention, write

$$E_{11} = U_S D V_S^\top, \qquad D = U_S^\top E V_S.$$

Then

$$W_c + E_{11} = U_S(\Sigma_S + D)V_S^\top + U_\perp C_\perp V_\perp^\top. \tag{16}$$

Assume that $\Sigma_S + D$ remains nonsingular and that the selection rule continues to select all r singular components of this block. Let

$$\Sigma_S + D = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top.$$

The perturbed selected singular bases can then be taken as $U_S\widetilde{U}$ and $V_S\widetilde{V}$. Since $\widetilde{U}$ and $\widetilde{V}$ are orthogonal, their projectors satisfy

$$\begin{aligned}P_{11} &= U_S\widetilde{U}\widetilde{U}^\top U_S^\top = U_S U_S^\top = P,\\ Q_{11} &= V_S\widetilde{V}\widetilde{V}^\top V_S^\top = V_S V_S^\top = Q.\end{aligned} \tag{17}$$

Thus the selected subspaces remain unchanged, although their internal singular directions and singular values can change. For the balanced factors

$$B_0 = s^{-1/2}U_S\Sigma_S^{1/2}, \qquad A_0 = s^{-1/2}\Sigma_S^{1/2}V_S^\top,$$

the weighted maps after the intervention are

$$\begin{aligned}L_{11} &= s^{-1}U_S\widetilde{U}\widetilde{\Sigma}\widetilde{U}^\top U_S^\top,\\ R_{11} &= s^{-1}V_S\widetilde{V}\widetilde{\Sigma}\widetilde{V}^\top V_S^\top.\end{aligned} \tag{18}$$

These matrices need not equal $L_0 = s^{-1}U_S\Sigma_S U_S^\top$ and $R_0 = s^{-1}V_S\Sigma_S V_S^\top$. Consequently,

$$Y_{11} = (L_{11} - L_0)G + G(R_{11} - R_0)$$

can be nonzero despite $P_{11} = P$ and $Q_{11} = Q$. Here $\mathcal{K}(G) = LG + GR$ excludes the s^2 prefactor in the effective-weight SGD update.

For example, take $D = \alpha\Sigma_S$ with $\alpha > -1$ small enough to preserve selection. Then

$$L_{11} = (1+\alpha)L_0, \qquad R_{11} = (1+\alpha)R_0, \qquad Y_{11} = \alpha\mathcal{K}_{W_c}(G).$$

This example establishes a response channel without selected-subspace displacement. Its importance for an observed perturbation is determined by the component interventions, rather than by subspace preservation alone.

B.6 Fixed-Context Decomposition and Factor Subspaces

Context-dependent spectral coordinates. CorDA provides a context-dependent instance of the same selected–complement analysis (Yang et al., 2024). Rather than decomposing W_c directly, it constructs the reference spectral matrix

$$Z_c = W_c C^{\mathrm{reg}}, \qquad H = E C^{\mathrm{reg}}, \tag{19}$$

where the regularized context transform C^{reg} is held fixed within each paired comparison. We form the selected–complement partition from the reference singular coordinates of Z_c, decompose the delivered perturbation H in that frame, and map each intervention back to weight space through $(C^{\mathrm{reg}})^{-1}$ before initializing the factors.

The blocks are orthogonal in the decomposition coordinates, but their mapped weight-space perturbations need not remain orthogonal. We therefore measure component energies in the decomposition coordinates and measure selected-subspace displacement from the actual initialized factors. This preserves the distinction between cross-boundary motion and selected-block changes while testing it under a context-dependent spectral construction.

For the context-dependent construction, let $C \in \mathbb{R}^{n\times n}$ denote the fixed, invertible context transform after regularization. Define

$$Z_c = W_c C, \qquad H = EC.$$

The context transform is held fixed across the reference and component interventions. Let $P_Z = U_S U_S^\top$ and $Q_Z = V_S V_S^\top$ be the selected projectors obtained from the reference SVD of Z_c. We form

$$\begin{aligned} H_{11} &= P_Z H Q_Z, & H_{10} &= P_Z H (I - Q_Z), \\ H_{01} &= (I - P_Z) H Q_Z, & H_{00} &= (I - P_Z) H (I - Q_Z), \end{aligned} \tag{20}$$

and map each intervention back to weight space:

$$E_F = H_F C^{-1}, \qquad (W_c + E_F)C = Z_c + H_F. \tag{21}$$

Every intervention is therefore defined using the same reference partition and context transform. The four H blocks are orthogonal under the Frobenius inner product and sum to H. Their mapped counterparts sum to E, but are generally not orthogonal, since

$$\langle E_F, E_J \rangle_F = \mathrm{tr}\left(H_F^\top H_J C^{-1} C^{-\top}\right).$$

Accordingly, block energies in decomposition coordinates and in weight coordinates are distinct quantities.

For a selected decomposition $U_S \Sigma_S V_S^\top$ of Z_c, the decomposition-balanced factors are

$$B_0 = s^{-1/2} U_S \Sigma_S^{1/2}, \qquad A_0 = s^{-1/2} \Sigma_S^{1/2} V_S^\top C^{-1}. \tag{22}$$

They satisfy $sB_0A_0 = U_S \Sigma_S V_S^\top C^{-1}$. The corresponding Gram matrices are

$$\begin{aligned} B_0 B_0^\top &= s^{-1} U_S \Sigma_S U_S^\top, \\ A_0^\top A_0 &= s^{-1} C^{-\top} V_S \Sigma_S V_S^\top C^{-1}. \end{aligned} \tag{23}$$

The right weighted map therefore includes the context transformation explicitly.

We measure the actual factor subspaces through their orthogonal projectors:

$$\begin{aligned} P_B &= B(B^\top B)^{-1} B^\top, \\ Q_A &= A^\top (AA^\top)^{-1} A. \end{aligned} \tag{24}$$

These expressions assume full factor rank; orthonormal bases provide an equivalent evaluation. For intervention F, the displacement measures are

$$d_L(F) = \|P_{B_F} - P_{B_0}\|_F, \qquad d_R(F) = \|Q_{A_F} - Q_{A_0}\|_F.$$

In particular, Q_{A_0} projects onto $\mathrm{span}(C^{-\top} V_S)$, which need not equal $\mathrm{span}(V_S)$.

Under an H_{11} intervention that preserves selection and rank, the selected bases in decomposition coordinates remain within $\mathrm{span}(U_S)$ and $\mathrm{span}(V_S)$. Because C is fixed and invertible, the actual factor spaces also remain

$$\mathrm{col}(B_{11}) = \mathrm{span}(U_S), \qquad \mathrm{col}(A_{11}^\top) = \mathrm{span}(C^{-\top} V_S).$$

Their weighted Gram matrices can nevertheless change through the internal selected-block decomposition. Thus selected-block subspace preservation extends to the mapped factor spaces, while the weighted response can remain nonzero.

Factor scaling is part of this response definition. For example, replacing B, A by $\gamma B, \gamma^{-1} A$ preserves their product and subspaces, but gives

$$\mathcal{K}_\gamma(G) = \gamma^2 B B^\top G + \gamma^{-2} G A^\top A.$$

We therefore specify the factor construction for each comparison and distinguish the balanced construction from native-scaling controls. The selected-block result identifies a structural response channel; the experiments measure its magnitude and its interaction with the cross blocks.

C Reference-State Derivative

We derive the derivative of the weighted spectral maps defining $\mathcal{K}_W$. This calculation concerns the initialized update operator, rather than the truncated matrix $U_r \Sigma_r V_r^\top$ itself. Throughout, s, r, and the probe gradient G are fixed.

C.1 Spectral-Function Representation

Define

$$Q_L(W) = W W^\top, \qquad Q_R(W) = W^\top W.$$

Assume $\sigma_r(W_c) > 0$ and $\sigma_r(W_c) > \sigma_{r+1}(W_c)$. Choose a threshold strictly between $\sigma_{r+1}^2(W_c)$ and $\sigma_r^2(W_c)$. Locally, let f equal $\sqrt{x}$ on neighborhoods of the selected eigenvalues and zero on neighborhoods of the discarded eigenvalues. Then

$$L(W) = s^{-1} f(Q_L(W)), \qquad R(W) = s^{-1} f(Q_R(W)). \tag{25}$$

The separated clusters permit this local representation without differentiating individual singular vectors.

C.2 Divided-Difference Derivative

At W_c, take complete eigendecompositions

$$Q_L(W_c) = \bar{U} \Lambda_L \bar{U}^\top, \qquad Q_R(W_c) = \bar{V} \Lambda_R \bar{V}^\top.$$

For either eigenvalue list, define

$$C_{ij} = \begin{cases} \dfrac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}, & \lambda_i \neq \lambda_j, \\ \dfrac{1}{2\sqrt{\lambda_i}}, & \lambda_i = \lambda_j \text{ in the selected cluster}, \\ 0, & \lambda_i = \lambda_j \text{ in the discarded cluster}. \end{cases} \tag{26}$$

The standard spectral-function derivative, combined with

$$DQ_L[E] = E W_c^\top + W_c E^\top, \qquad DQ_R[E] = E^\top W_c + W_c^\top E,$$

gives

$$\begin{aligned} DL_{W_c}[E] &= s^{-1} \bar{U} \left[C_L \odot \bar{U}^\top (E W_c^\top + W_c E^\top) \bar{U} \right] \bar{U}^\top, \\ DR_{W_c}[E] &= s^{-1} \bar{V} \left[C_R \odot \bar{V}^\top (E^\top W_c + W_c^\top E) \bar{V} \right] \bar{V}^\top, \end{aligned} \tag{27}$$

where $\odot$ denotes the Hadamard product. Every quantity is evaluated at the reference state.

For selected indices i, j, the coefficient is $1/(\sigma_i + \sigma_j)$. For selected i and discarded j, it is $\sigma_i/(\sigma_i^2 - \sigma_j^2)$. Discarded–discarded coefficients vanish. Repeated eigenvalues within either cluster use the limits in Equation equation 26; a repeated eigenvalue across the truncation boundary violates the gap assumption. The thresholds that define the two clusters are chosen strictly between $\sigma_{r+1}^2(W_c)$ and $\sigma_r^2(W_c)$, so the construction requires $\sigma_r(W_c) > \sigma_{r+1}(W_c)$, and it requires no separation between repeated eigenvalues inside a cluster, where the limit value applies. Both conditions are properties of the reference spectrum; they are assumed here rather than tested, and the dose at which the boundary condition fails is not mapped (Section B).

C.3 Full and Cross-Only Responses

For a fixed gradient, define the linear map

$$\mathcal{D}_G[E] = DL_{W_c}[E]G + GDR_{W_c}[E].$$

A pure complementary perturbation gives $\mathcal{D}_G[E_{00}] = 0$. Indeed, its Gram perturbations have only discarded–discarded blocks in the reference eigenbases, where the divided differences vanish. Therefore,

$$\begin{aligned} \widehat{Y} &= \mathcal{D}_G[E_{\text{cross}}] + \mathcal{D}_G[E_{11}], \\ \widehat{Y}_{\text{cross}} &= \mathcal{D}_G[E_{\text{cross}}]. \end{aligned} \tag{28}$$

Cross terms involve the gap-dependent coefficients, but selected–selected terms remain part of the complete first-order response.

Locally,

$$Y(E;G) = \mathcal{D}_G[E] + \mathcal{R}(E;G), \qquad \|\mathcal{R}(E;G)\|_F = O(\|E\|_F^2) \tag{29}$$

for fixed G. The constants depend on the reference spectrum, s, and the gradient scale. Thus

$$Y - \widehat{Y}_{\text{cross}} = \mathcal{D}_G[E_{11}] + \mathcal{R}(E;G). \tag{30}$$

Cross-only prediction error contains both the omitted selected–selected first-order term and the full linearization residual. These terms can reinforce or cancel, so their norms are not additive.

C.4 Finite Interventions and Prediction

The finite cross intervention and its linear prediction are different objects:

$$Y_{\text{cross}} = \big(\mathcal{K}_{W_c + E_{\text{cross}}} - \mathcal{K}_{W_c}\big)(G), \qquad \widehat{Y}_{\text{cross}} = \mathcal{D}_G[E_{\text{cross}}].$$

The former recomputes the perturbed initialization; the latter uses only the reference derivative. Locally,

$$Y_{\text{cross}} = \widehat{Y}_{\text{cross}} + O(\|E_{\text{cross}}\|_F^2).$$

Likewise, first-order additivity implies

$$Y - Y_{\text{cross}} - Y_{\text{within}} = O(\|E\|_F^2)$$

in a sufficiently small neighborhood of W_c.

These local expansions motivate the analysis, but do not determine the relative contribution of each term at the observed magnitudes. The experiments separately evaluate finite cross reproduction, block interaction, and first-order prediction accuracy.

D Experimental Protocol and Coverage

D.1 Upstream Construction and Layer Selection

Experiments use text-classification adaptation states of DeBERTa and RoBERTa. Task adaptation updates both the encoder and classification head. The subsequent paired upstream construction produces a reference weight W_c and a perturbed weight W_δ, with $E = W_\delta - W_c$.

Table 5: Shared configuration for the paired upstream construction. Parameter names identify the corresponding configuration fields.

Parameter	Value
Input sample count (`n_samples`)	64
Validation chunks (`n_val_chunks`)	3
Random-sign draws (`n_random_draws`)	2
Inner learning rate (`inner_lr`)	10^{-3}
Inner steps (`inner_steps`)	2
Input budget (`epsilon`)	0.1
Maximum sequence length (`max_length`)	128
`n_theta`	128
Data split	validation

Input perturbations are applied to word embeddings at the positions of the inserted trigger `banana split`. Both branches retain the same trigger and preprocessing; the reference branch sets the prescribed embedding perturbation to zero. Thus, W_c is the weight after the reference upstream update, rather than the untouched pretrained checkpoint. The evaluated perturbation families are random sign and the layer-targeted spectral construction. Each layer and seed contributes two random-sign draws and one spectral draw.

Table 5 records the shared upstream configuration. The value $\epsilon = 0.1$ denotes the configured input-perturbation budget, not the norm of the resulting weight perturbation E.

Perturbation families and the trigger. Both families perturb the word embeddings at the positions of the inserted trigger tokens, under the configured input budget of Table 5. The random-sign family applies independently drawn signs to the perturbed embedding entries and is drawn twice per layer and seed. The spectral family is layer-targeted: it maximizes the squared Frobenius projection of the induced weight perturbation onto the complementary singular subspace, and is solved as a fixed-budget projected gradient ascent on the trigger-token embedding entries. Its projection basis is taken from the pretrained (step-zero) weight rather than from W_c, so that the target of the construction and the basis of the attribution are different objects; a construction that maximized projections in the attribution basis would make the cross-dominance measurement circular. The trigger is a fluent two-token phrase inserted at a fixed position; both branches retain it, and only the reference branch sets the embedding perturbation to zero, so that the branches differ in the perturbation alone. The norm convention of the budget, the ascent settings, and the trigger construction are recorded in the released configuration.

Core layers are DeBERTa L5v/L6v and RoBERTa R2q/R1q, R11v/R9v, and R10v/R6v. Layer indices are zero-based, and q/v denote query/value projections. RoBERTa core selection followed a spectral census and preceded the corresponding mechanism measurements, using threshold and matching rules.

Prediction layers are DeBERTa L1q/L1v/L9q/L9v and RoBERTa R0q/R1v/R8v/R9q. Selection used module, depth, and documented prior-use criteria. Their spectra had been included in the census, and a reduced pilot was inspected after candidate selection. These layers were held out from the core component-intervention analysis; Table 6 summarizes the coverage.

D.2 Gradient Probes and Aggregation

Seeds 42–51 are the statistical units. For each fixed condition, repeated perturbation draws and gradient probes are first averaged within seed; group summaries are then computed over the ten seed aggregates. A group is the experiment-specific combination of layer, perturbation family, gradient condition, and, where applicable, common initial target and checkpoint. Groups may share seeds and model checkpoints.

Table 6: Experiment-specific coverage. Core denotes eight layers. The six-layer trajectory subset contains L5v/L6v and R2q/R1q/R10v/R6v. The controlled-gradient and task-gradient comparisons use distinct probe protocols.

Evaluation	Layers	Rank
Controlled-gradient comparisons	Core	8
Task-gradient comparisons	Core	8
Controlled rank extension	L5v/L6v	4, 8, 16, 32
Component interventions	Core	8
Matched random controls	Core	8
Short-horizon adaptation	Six-layer subset	8
Conditional prediction	Eight held-out layers	8
Natural-dose example	R10v/R6v	8

The natural component panel contains 80 groups: eight layers, two perturbation families, and five gradient conditions. Component interventions use rank eight, $s = 1$, and the observed perturbation magnitudes. The gradient conditions comprise four synthetic probes ($G_{\text{iso}}, G_{01}, G_{10}, G_{00}$) and one task-gradient condition. Each synthetic condition uses four gradient draws. Task-gradient probes use four batches of 16 unperturbed examples with true labels, disjoint from the upstream-construction examples.

The original component-summary fields `ci_lo` and `ci_hi` are empirical seed quantiles, not confidence intervals for the mean. For sorted seed values $x_{(0)}, \ldots, x_{(n-1)}$, their indices are $\lfloor 0.025n \rfloor$ and $\lfloor 0.975n \rfloor$. With ten seeds, these endpoints equal the observed minimum and maximum. We therefore treat them as descriptive seed ranges.

D.3 Matched Random Controls

For each observed perturbation, we construct five random controls. An independent Gaussian matrix is projected into the four reference blocks, and each block is rescaled to match the corresponding Frobenius norm of the observed perturbation. This preserves all four block energies and hence the total perturbation norm, while randomizing directions within the blocks.

Response ratios are formed before aggregation:

$$q_{s,d,g,b} = \frac{\|Y^{\text{random}}_{s,d,g,b}\|_F}{\|Y^{\text{input}}_{s,d,g}\|_F},$$

where s, d, g, and b index seeds, input-perturbation draws, gradient probes, and random controls. Ratios are averaged within seed and then across seeds. A mean random-to-input ratio is not inverted to obtain a mean input-to-random ratio.

For this ratio analysis, 95% confidence intervals are percentile bootstrap intervals for the mean of the seed aggregates. We use 2,000 resamples with replacement and `numpy.random.RandomState(2026)`, taking the 2.5th and 97.5th percentiles of the resampled means. These are individual group intervals, not simultaneous confidence intervals across groups. With ten seed aggregates the bootstrap operates on a small sample, so these intervals summarize seed variation rather than providing coverage calibration.

A ratio is excluded when its input-induced denominator fails the response threshold. A below-threshold random response is retained when the input-induced denominator is valid. Directional comparisons require both responses to be identifiable. The natural input-induced panel has no exclusions. Four random-control responses fall below threshold; all concern RoBERTa R1q, seed 42, a random-sign draw, and G_{00}, rather than task gradients.

Table 7: Natural-magnitude component interventions. The signed projections p_{cross} and p_{within} are dimensionless. Cross reproduction error is expressed as a percentage. Near-zero lower endpoints for p_{within} are rounded.

Model	Gradients	p_{cross}	p_{within}	Cross error (%)
DeBERTa	Synthetic	0.996890–1.000165	$\approx$ 0–0.003113	0.0107–5.5231
DeBERTa	Task	0.993116–0.994785	0.005295–0.006871	0.6579–0.9645
RoBERTa	Synthetic	0.991781–1.000226	$\approx$ 0–0.008220	0.0130–8.6429
RoBERTa	Task	0.985137–0.999944	0.000194–0.014870	0.1693–3.4583

Table 8: Maximum relative interaction residuals in the natural component panel. Values are dimensionless. Seed maxima refer to seed aggregates.

Model	Maximum group mean	Maximum seed aggregate
DeBERTa	0.000891	0.001968
RoBERTa	0.005733	0.019250

E Component Attribution Details

E.1 Retention, Removal, and Perturbation Energy

Table 7 separates task-gradient results from the synthetic-gradient panel. Entries summarize layer–family–gradient group means, rather than individual-record extrema.

All 80 natural-panel groups have positive $p_{\text{cross}} - p_{\text{within}}$ contrasts in all ten seeds. Across these groups, cross components account for 2.64%–3.51% of perturbation energy in DeBERTa and 1.19%–6.73% in RoBERTa. Their contribution to the response is therefore disproportionate to their energy.

Signed projection and response norm measure different properties. Writing

$$Y_{\text{within}} = p_{\text{within}} Y + Y_{\perp}, \qquad \langle Y_{\perp}, Y \rangle_F = 0,$$

gives

$$\frac{\|Y_{\text{within}}\|_F^2}{\|Y\|_F^2} = p_{\text{within}}^2 + \frac{\|Y_{\perp}\|_F^2}{\|Y\|_F^2}.$$

Thus, the removal result directly establishes attenuation along the full-response direction.

The reproduction error is not bounded by one. With $\beta_F = \|Y_F\|_F^2 / \|Y\|_F^2$, the identity $\epsilon_F^2 = \beta_F + 1 - 2p_F$ holds per record, so ϵ_F grows towards $\sqrt{2}$ as $p_F \to 0$ and exceeds one whenever the component orthogonal to Y dominates Y_F. We therefore verify $\epsilon_F \geq 1 - p_F$, equivalently $\beta_F \geq p_F^2$, at every record and at every group mean, and we read values above 100% as near-orthogonality rather than as a numerical failure.

E.2 Finite Interaction

The interaction residual is

$$I_{\text{cw}} = Y - Y_{\text{cross}} - Y_{\text{within}}, \qquad \iota_{\text{cw}} = \frac{\|I_{\text{cw}}\|_F}{\|Y\|_F}.$$

It measures the departure from additivity of the two finite component interventions. Table 8 reports the maximum residuals.

The largest RoBERTa interaction occurs for R6v under random-sign perturbations and G_{10}; the largest seed aggregate is from seed 47. These residuals indicate that cross reproduction does not rely on a large nonadditive correction between cross and within interventions.

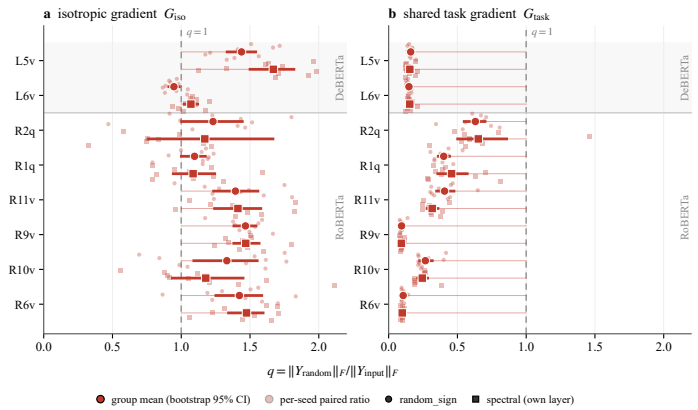

Figure 4: Response ratios for block-energy-matched random controls relative to input-induced perturbations. Panel (a) uses isotropic gradients and panel (b) uses shared task gradients. Matching the four block energies produces broadly comparable response magnitudes under isotropic gradients, but consistently smaller responses under the evaluated task gradients. Ratios are computed for individual paired measurements, averaged within seed, and then summarized across seeds. Points show seed-level paired ratios; larger markers and horizontal intervals show group means and bootstrap 95% confidence intervals.

The same quantities provide a bound on the total within response. Since $Y_{\text{within}} = Y - Y_{\text{cross}} - I_{\text{cw}}$, the triangle inequality gives

$$\rho_{\text{within}} := \frac{\|Y_{\text{within}}\|_F}{\|Y\|_F} \leq \epsilon_{\text{cross}} + \iota_{\text{cw}}. \tag{31}$$

Thus, small cross reproduction error and small interaction jointly constrain the within-response norm, beyond what its signed projection alone establishes. The inequality is preserved under averaging over the same valid records with the same weights. It gives a derived upper bound, rather than a direct measurement of ρ_{within}.

E.3 Block-Energy Matching and Task Responses

Block-energy-matched random controls also exhibit cross-dominated attribution. Under task gradients, their group-mean cross reproduction errors range from 3.2455% to 4.2136% in DeBERTa and 0.8657% to 7.1880% in RoBERTa. This shared pattern is consistent with a geometric contribution to component attribution.

Response magnitude nevertheless differs. Mean random-to-input response ratios range from 0.947 to 1.670 under isotropic gradients, and from 0.093 to 0.652 under task gradients. Each of the 16 task groups has a bootstrap confidence interval with an upper endpoint below one.

In 15 task groups, all ten seed-level ratios are below one. The exception is R2q–spectral, with nine of ten: seed 45 has a ratio of 1.457 and is retained. Matching block energies therefore does not eliminate task-response differences, although it reproduces the broad attribution pattern.

E.4 Block-Energy-Matched Random Controls

Shared attribution does not fix response magnitude. random controls matched in all four block energies exhibit similar cross-component attribution. However, the paired response ratio $q = \|Y_{\text{random}}\|_F / \|Y_{\text{input}}\|_F$ has group means of 0.947–1.670 under isotropic gradients and 0.093–0.652 under task gradients. All sixteen task-gradient groups have individual 95% confidence intervals below one. Block-energy matching therefore reproduces the broad attribution pattern without explaining the larger input-induced task responses. This distinguishes which components control a response from how strongly the delivered perturbation acts through them.Seed-level paired ratios and confidence intervals are shown in Appendix Figure 4.

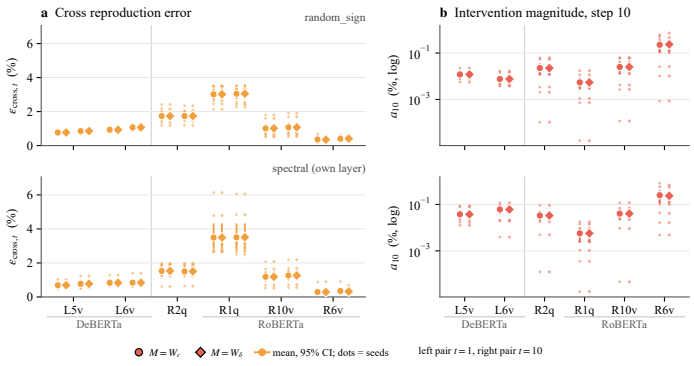

Figure 5: Matched-forward short-horizon validation: cross retention approximately reproduces the full-intervention trajectory separation at steps one and ten, while the separation remains small relative to the reference update.

Table 9: Group-mean cross reproduction errors. Ranges span perturbation families and common initial targets $M \in \{W_c, W_\delta\}$, among effects satisfying Equation 33.

Layer	Step 1	Step 10
L5v	0.69%–0.77%	0.77%–0.86%
L6v	0.84%–0.93%	0.84%–1.07%
R2q	1.53%–1.73%	1.50%–1.74%
R1q	3.02%–3.49%	3.05%–3.51%
R10v	1.01%–1.19%	1.07%–1.27%
R6v	0.29%–0.36%	0.32%–0.41%

F Short-Horizon Results and Precision

For short-horizon validation, we initialize each frame with a frozen residual

$$W_{\text{res}}^{(F)} = M - sB_F A_F, \tag{32}$$

giving all frames the same initial effective weight M.

Short-horizon adaptation uses simultaneous factor SGD with learning rate 0.01 and $s = 1$, without momentum, weight decay, or clipping. Only the target factors are trained, and dropout is disabled. For each common target $M \in \{W_c, W_\delta\}$, the four intervention frames have identical initial effective weights. Non-target parameters and batch schedules are matched. Initial gradients therefore agree, while subsequent gradients are evaluated along each frame's own trajectory.

F.1 Trajectory Reproduction and Effect Scale

For each common initial target M, let

$$Z_t^{(a)} = W_{\text{eff},t}^{(a)} - W_{\text{eff},t}^{(\text{ctrl})}.$$

Cross reproduction compares $Z_t^{(\text{cross})}$ with $Z_t^{(\text{full})}$. Table 9 reports the group-mean reproduction errors.

At step ten, group means of

$$a_{10} = \frac{\|Z_{10}^{(\text{full})}\|_F}{\|W_{\text{eff},10}^{(\text{ctrl})} - M\|_F}$$

Table 10: Step-ten within/full trajectory norm ratios, $100\|Z_{10}^{(\mathrm{within})}\|_F/\|Z_{10}^{(\mathrm{full})}\|_F$. Entries give median (mean) percentages in the layer–target summaries.

Layer	$M = W_c$	$M = W_\delta$
L5v	0.836 (0.830)	0.837 (0.832)
L6v	0.986 (0.991)	0.986 (0.991)
R2q	1.637 (1.622)	1.637 (1.622)
R1q	3.169 (3.166)	3.169 (3.166)
R10v	1.021 (1.189)	1.021 (1.189)
R6v	0.299 (0.302)	0.266 (0.287)

range from 0.0076% to 0.061% in DeBERTa and 0.0054% to 0.25% in RoBERTa. Cross retention therefore approximately reproduces a trajectory separation that remains small relative to the reference update.

Table 10 additionally reports the within-intervention trajectory norm relative to the full-intervention trajectory norm. This is a finite-step trajectory measurement, distinct from the shared-gradient operator response in Appendix E.

F.2 Thresholding and Numerical Precision

Relative cross reproduction errors are summarized only when

$$\|Z_t^{(\mathrm{full})}\|_F > 10^{-6}\|W_{\mathrm{eff},t}^{(\mathrm{ctrl})} - M\|_F. \tag{33}$$

This threshold concerns effect size relative to the reference update, rather than floating-point accuracy.

Across $t \in \{1, 5, 10\}$, all 24 DeBERTa group–checkpoint cells contain ten valid seed aggregates. Of the 48 RoBERTa cells, 28 contain ten and 20 contain nine. R1q cells contain nine, R6v cells contain ten, and R2q/R10v cells contain nine or ten.

There are 38 below-threshold cross-reproduction records, all from seed 42: 14 at step one, 12 at step five, and 12 at step ten. They are omitted only from the corresponding relative-error summaries. The maximum individual-record step-ten reproduction error is approximately 6.05%.

The reported trajectories use the final float64 runs. Checks at initialization give zero differences in effective weights, gradients, and losses across frames for each common target. Earlier float32 late-step query-layer errors are not used to characterize mechanism reproduction.

F.3 Functional Measurements

On the fixed 32-example evaluation batch, no predicted-class disagreements are observed. At step ten, the largest batch logit-difference ℓ_2 norms across intervention frames are 2.904×10^{-5} for DeBERTa and 7.562×10^{-3} for RoBERTa. The corresponding largest absolute loss differences are 2.384×10^{-7} and 1.080×10^{-4}. These measurements describe the same fixed batch; repeated evaluations are not independent test samples.

G Prediction Errors and Dose Diagnostics

G.1 Held-Out Conditional Prediction

Each model contributes four held-out layers, ten seeds, three perturbation draws per layer and seed, and four task-gradient batches, giving 480 records. The draws comprise two random-sign perturbations and one spectral perturbation. Families are summarized separately, and the main prediction analysis has no exclusions.

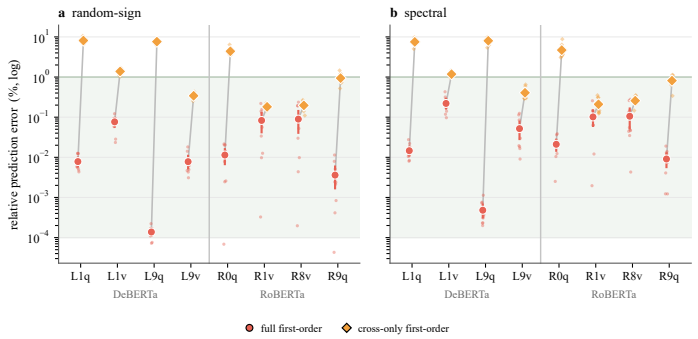

Figure 6: Conditional prediction on held-out layers: the full reference-state first-order model predicts the finite response with errors well below one percent, while cross-only predictions leave larger residuals. Larger markers show group means with 95% confidence intervals; the vertical axis is logarithmic.

Table 11: Held-out prediction errors and omitted first-order contributions. Entries are layer–family group means, expressed as percentages, with ten seed aggregates per group. Full and Cross-only denote relative prediction errors; Omitted denotes $100\|\widehat{Y} - \widehat{Y}_{\text{cross}}\|_F/\|Y\|_F$. RS denotes random sign; SP denotes spectral.

Layer	Family	Full (%)	Cross-only (%)	Omitted (%)
L1q	RS	0.007833	8.1328	8.1335
L1q	SP	0.014630	7.5233	7.5225
L1v	RS	0.076580	1.3815	1.3751
L1v	SP	0.220200	1.1804	1.1598
L9q	RS	0.0001378	7.6215	7.6215
L9q	SP	0.0004807	8.0183	8.0183
L9v	RS	0.007796	0.34169	0.3408
L9v	SP	0.051920	0.40568	0.4050
R0q	RS	0.011400	4.3864	4.3884
R0q	SP	0.021090	4.6969	4.7000
R1v	RS	0.082900	0.18117	0.1418
R1v	SP	0.101600	0.20956	0.1718
R8v	RS	0.089570	0.19693	0.1683
R8v	SP	0.105600	0.25774	0.2108
R9q	RS	0.003602	0.93983	0.9397
R9q	SP	0.009087	0.81363	0.8135

The predictor uses the reference state, observed weight perturbation, and specified gradient. Predictions are saved before the corresponding finite-response evaluation. Perturbed singular decompositions, perturbed factors, and finite-response targets are not predictor inputs. Task gradients are obtained in float32, while the spectral calculations use float64.

The maximum individual-record full-model errors are approximately 0.4424% for DeBERTa and 0.4375% for RoBERTa. These maxima are distinct from the group means in Table 11.

The omitted first-order contribution is $\widehat{Y} - \widehat{Y}_{\text{cross}} = D_G[E_{11}]$, since $D_G[E_{00}] = 0$. The cross-only residual satisfies

$$Y - \widehat{Y}_{\text{cross}} = D_G[E_{11}] + (Y - \widehat{Y}).$$

Table 11 reports the relative norms of the omitted term and the full-model residual, alongside the cross-only error. These norms need not add, because the residual decomposition is a vector identity. Their relative magnitudes vary across conditions; for example, the full-model error in L1v–spectral is approximately 19% of the cross-only error. Dominant-response attribution therefore does not make the remaining prediction terms negligible.

Table 12: Median seed-level PiSSA/RandFP sensitivity ratios on DeBERTa, with $r = 8$ and $\|E\|_F = 10^{-3}$.

Layer	Family	G_{iso}	G_{00}
L5v	Spectral	436.9	623.4
L5v	Random sign	376.9	543.4
L6v	Spectral	253.1	352.2
L6v	Random sign	276.2	390.0

G.2 Dose Diagnostics

A supplementary seed-42 diagnostic evaluates $E/2$ and $E/4$ without regenerating inputs. Each prediction record carries a `dose` field (1, $1/2$, or $1/4$) in place of a separate file per scale, and the dose-ratio summaries below are computed from that field. Ratios are computed only for records valid at both scales.

For DeBERTa, the median relative-error ratio $e(1/2)/e(1/4)$ is 2.0000 across 48 pairs, with a range of 1.9470–2.0353. For RoBERTa, the median is 1.9998 across 36 pairs, with a range of 0.3960–2.0907.

The median scaling is consistent with a quadratic absolute remainder relative to an approximately linear response. The wider RoBERTa range also reflects the sensitivity of relative errors when responses are small. This diagnostic characterizes the tested scales without establishing a uniform validity radius.

H Initialization and Perturbation Delivery

Why measure the delivered perturbation? On four held-out layers per model, mean seed-level Spearman correlations with $\|Y\|_F/\|G\|_F$ range from 0.04 to 0.62 for the static gap-based indicator, compared with 0.82–0.98 for the delivered magnitude $\|E\|_F$. These auxiliary rankings motivate observing the perturbation that reaches the weights. They do not identify which of its components controls the response, which we test next.

H.1 Parameterization and Rank Comparisons

Parameterization comparisons use the same fixed reference tangent space:

$$S_{K,T_c}(E,G) = \frac{\|\Pi_{T_c}\Delta\mathcal{K}_E(G)\|_F}{\|E\|_F\|G\|_F}.$$

For each seed, the PiSSA/RandFP ratio is formed after averaging the sensitivity of each method over its repeated observations:

$$F_s = \frac{\text{mean}_{d,g}\, S^{\text{PiSSA}}_{K,T_c}}{\text{mean}_{d,g,b}\, S^{\text{RandFP}}_{K,T_c}}.$$

Table 12 reports the median of these seed-level ratios.

The direction-related contrasts retain their signs over the tested DeBERTa ranks $4, 8, 16, 32$. This rank extension concerns the controlled comparisons; the component, trajectory, and prediction analyses use rank eight.

H.2 Static Rankings and Natural Magnitudes

Table 13 reports the ranking agreement.

The natural-dose comparison uses

$$S_{K,\text{nat}} = \frac{\|Y\|_F}{\|E\|_F\|G\|_F}.$$

Table 13: Mean seed-level Spearman correlations with $\|Y\|_F/\|G\|_F$ across four prediction layers per model. Observed $\|E\|_F$ requires the upstream construction.

Model	Family	Static gap indicator	$\|E\|_F$
DeBERTa	Random sign	0.56	0.92
DeBERTa	Spectral	0.62	0.98
RoBERTa	Random sign	0.24	0.92
RoBERTa	Spectral	0.04	0.82

Table 14: Natural-dose comparison at rank eight. Entries are means across ten seed aggregates. Sensitivity is evaluated at the observed perturbation.

Layer	$\|E\|_F$	$S_{K,\mathrm{nat}}$	$\|Y\|_F$
R10v	8.140×10^{-4}	0.02939	2.6175×10^{-5}
R6v	8.811×10^{-3}	0.01052	8.0347×10^{-5}

Under random-sign perturbations and unit-norm isotropic gradients, R10v has higher normalized sensitivity, whereas R6v has a larger delivered perturbation and a larger absolute response (Table 14).

The ranking analysis and natural-dose comparison use different layer sets. The former measures ordering agreement; it does not predict response vectors. The natural-dose columns are averaged separately, so their means need not satisfy the per-record multiplicative identity.

I Subspace Motion and Weighted Responses Across Spectral Constructions

This appendix reports the component measurements supporting Section 4.3. The comparison uses rank eight, random-sign perturbations, shared isotropic gradients, and seeds 42–51. Principal and minor selection are evaluated on matched layer sets. The context-dependent constructions use mode-specific layer sets and separately specified factor scalings.

I.1 Selected-Block Responses under Direct Spectral Selection

For each intervention, we measure

$$a_{11} = \frac{\|Y_{11}\|_F}{\|Y\|_F}, \qquad e_{\mathrm{cross}} = \frac{\|Y_{\mathrm{cross}} - Y\|_F}{\|Y\|_F}.$$

The first quantity measures the response to the isolated selected block; the second measures the discrepancy left by cross retention. Table 15 reports their layer-wise medians after within-seed aggregation.

The selected-block displacement ratios are

$$r_L^{11} = \frac{d_L(11)}{d_L(\mathrm{full})}, \qquad r_R^{11} = \frac{d_R(11)}{d_R(\mathrm{full})}.$$

Under minor selection, their largest layer-wise medians are 1.10×10^{-8} and 1.15×10^{-8}, respectively. The corresponding selected-block responses remain identifiable and reach 21.25%–78.74% of the full response. This contrast establishes the empirical importance of internal selected-block changes in these settings.

Principal-selection displacement medians are also small, with maxima of 4.12×10^{-8} on the left and 6.41×10^{-8} on the right. The numerical response summaries are reported together with their identifiability counts. Below-floor values describe the recorded calculation but are not used as evidence of an identifiable selected-block effect.

Table 15: Direct spectral constructions under random-sign perturbations and isotropic gradients. Response ratios and errors are percentages. The principal-selection a_{11} column retains the ten-seed numerical summaries, including below-floor measurements as descriptive values. n_{11} counts seed aggregates for which all constituent selected-block responses pass the identifiability check; it is not the number of values in that descriptive median. All minor-selection groups have ten identifiable full-response and selected-block seed aggregates. Principal-selection cross errors use identifiable full responses: nine seeds for R9q and ten otherwise.

	Principal selection			Minor selection	
Layer	a_{11} (%)	e_{cross} (%)	n_{11}	a_{11} (%)	e_{cross} (%)
L1q	7.83	7.83	10	21.25	21.25
L1v	3.48	3.48	10	50.35	50.35
L3v	3.05	3.05	10	53.51	53.51
L4q	8.53	8.53	10	64.95	64.95
L4v	3.05	3.06	10	54.42	54.42
L8q	8.64	8.64	9	55.73	55.73
L9q	9.64	9.64	1	55.03	55.03
L9v	2.04	2.04	10	53.61	53.61
L11q	5.05	5.05	0	46.21	46.21
L11v	4.34	4.34	10	66.50	66.50
R0q	3.37	3.37	9	42.08	42.08
R0v	3.43	3.43	9	30.02	31.39
R1v	2.18	2.22	9	78.74	78.75
R2v	1.10	1.25	9	78.06	77.75
R4q	6.23	6.24	9	66.27	66.27
R5v	1.96	2.10	9	71.85	71.65
R7q	6.52	6.52	9	66.36	66.34
R8v	1.48	1.61	9	60.98	60.99
R9q	5.21	5.21	9	66.69	66.69

I.2 Actual Factor Subspaces under Context-Dependent Decomposition

The context-dependent measurements use the weight-space factor projectors defined in Appendix B.6. They therefore measure the spaces of B and $A^\top$ after mapping back from decomposition coordinates. The context transform and reference factor scaling are fixed within each intervention comparison.

Table 16 jointly reports selected-block displacement, selected-block response, and cross reproduction for the balanced constructions.

There are 160 model–mode–layer–seed cells in the balanced panel. The selected-block response is fully identifiable in all 80 DeBERTa cells and 79 of the 80 RoBERTa cells. The exception is R0q–IPM at seed 42. Its full response remains identifiable, so its cross reproduction error is retained. The subspace measurements are evaluated separately from the response-identifiability condition.

I.3 Finite Interactions between Component Responses

Finite interactions further determine how the isolated responses combine. For CorDA-KPM at RoBERTa R2v, for example, the layer-wise medians are 37.49% for a_{11} and 47.25% for e_{cross}, while the independently measured residual

$$\frac{\|Y - Y_{\text{cross}} - Y_{11} - Y_{00}\|_F}{\|Y\|_F}$$

is 15.09%. The cross-only discrepancy cannot therefore be assigned entirely to the selected-block response. Selected-block responses establish that subspace motion is insufficient, while finite interactions determine how isolated component responses combine. These separately summarized medians are not used as terms in a numerical vector identity.

Table 16: Context-dependent balanced constructions. Entries are medians across seed aggregates. r_L^{11} and r_R^{11} are dimensionless; response quantities are percentages. Each row has ten seeds for subspace displacement and cross reproduction. The selected-block response uses ten seeds except R0q–IPM, where seed 42 is below the selected-response floor and $n_{11} = 9$.

Mode	Layer	r_L^{11}	r_R^{11}	a_{11} (%)	e_{cross} (%)	n_{11}
IPM	L1v	4.68×10^{-12}	3.02×10^{-9}	10.90	10.90	10
IPM	L8q	4.78×10^{-9}	1.03×10^{-5}	15.58	15.58	10
IPM	L9q	1.59×10^{-8}	4.03×10^{-5}	17.98	17.98	10
IPM	L11v	9.89×10^{-10}	2.75×10^{-6}	32.62	32.62	10
KPM	L1v	1.72×10^{-10}	2.00×10^{-10}	66.26	66.26	10
KPM	L3v	6.51×10^{-10}	1.13×10^{-9}	63.33	63.33	10
KPM	L9q	3.86×10^{-5}	4.33×10^{-5}	52.21	52.21	10
KPM	L11v	3.28×10^{-6}	5.24×10^{-6}	64.61	64.61	10
IPM	R0q	5.30×10^{-11}	5.57×10^{-8}	7.85	7.80	9
IPM	R2v	1.57×10^{-13}	9.56×10^{-11}	10.19	10.42	10
IPM	R5v	2.84×10^{-13}	5.34×10^{-11}	19.93	22.35	10
IPM	R7q	2.47×10^{-11}	8.96×10^{-9}	15.40	15.40	10
KPM	R2v	1.81×10^{-11}	4.48×10^{-11}	37.49	47.25	10
KPM	R4q	3.53×10^{-9}	3.25×10^{-9}	68.31	68.30	10
KPM	R7q	6.01×10^{-9}	5.20×10^{-9}	58.68	58.66	10
KPM	R8v	2.67×10^{-10}	5.30×10^{-10}	78.73	78.73	10

Table 17: Finite three-part interaction residuals in the balanced context-dependent panel. Entries are medians across ten seed aggregates, reported as percentages of the full response norm. Small values are retained rather than rounded to zero.

Mode	Layer	ϵ_{add} (%)
IPM	L1v	0.0312
IPM	L8q	0.000271
IPM	L9q	0.000526
IPM	L11v	0.00334
KPM	L1v	0.293
KPM	L3v	0.0868
KPM	L9q	0.00883
KPM	L11v	0.00588
IPM	R0q	0.00297
IPM	R2v	2.06
IPM	R5v	7.41
IPM	R7q	0.0382
KPM	R2v	15.09
KPM	R4q	0.677
KPM	R7q	0.424
KPM	R8v	1.63

For the three-part decomposition, define

$$J = Y - Y_{\mathrm{cross}} - Y_{11} - Y_{00}, \qquad \epsilon_{\mathrm{add}} = \frac{\|J\|_F}{\|Y\|_F}. \tag{34}$$

This residual is distinct from the cross–within residual $I_{\mathrm{cw}} = Y - Y_{\mathrm{cross}} - Y_{\mathrm{within}}$. Their exact relationship is

$$J = I_{\mathrm{cw}} + (Y_{\mathrm{within}} - Y_{11} - Y_{00}).$$

It is also distinct from decomposing cross into its two separate finite interventions. Table 17 reports these residuals.

The exact response decomposition is

$$Y - Y_{\mathrm{cross}} = Y_{11} + Y_{00} + J.$$

Table 18: Native-scaling measurements on common layers. Entries are layer-wise seed medians in percent. Each response is normalized by its own construction's full response; the table does not compare absolute response magnitudes across modes.

	IPM		KPM	
Layer	a_{11}	e_{cross}	a_{11}	e_{cross}
L1v	10.71	10.71	104.46	104.42
L9q	18.17	18.17	47.96	47.96
L11v	32.76	32.76	78.28	78.28
R2v	10.26	10.26	18.30	25.00
R7q	15.50	15.50	37.75	37.86

Thus an identifiable selected-block response establishes an additional response channel, while the independently measured residual determines whether isolated block responses approximately add. The reported medians summarize each quantity separately and are not substituted into this vector identity.

I.4 Factor-Scaling Controls

This is an exact *difference* for this construction: the finite-step product terms cancel because $B_\beta A_\beta$ is shared. Measured first-step differences agree with Eq. equation 6 to approximately 10^{-15}–10^{-13} relative error. This algebraic agreement checks the intervention; the empirical questions are its size under task gradients and whether it persists when each arm recomputes its own gradients.

Changing the selected singular values. Our complementary SR-D intervention constructs $W_{\text{spec}} = W_0 + U_r D V_r^\top$ and compensates each balanced factorization with a branch-specific frozen residual, so both arms again start from W_0 with the same selected subspaces. The prespecified nonuniform D has $\|D\|_F/\|W_0\|_F = 10^{-3}$. In the five layers that pass the selection-order gate, its first-step difference is 0.279%–0.395% of the reference step, and its ten-step trajectory separation is 0.274%–0.390% of the reference update. All ten R6v nonuniform units fail the prespecified gate because the selected components reorder internally; they are excluded *before* training, rather than selected based on outcomes. The step-10 fixed-evaluation loss differences are only of order 10^{-7}–10^{-6} and produce no classification disagreements.

The control patterns verify the meaning of this intervention. A multiplicative scaling of all selected singular values changes the initial update map by a single scalar, with median scalar-alignment residuals of order 10^{-14} in both models. An additive constant is not a pure scaling control: its median residual is 1.70×10^{-4} in DeBERTa and 7.34×10^{-4} in RoBERTa. The nonuniform intervention increases these values to 3.44×10^{-3} and 2.50×10^{-3}, respectively. Thus changes to selected spectral values enter early training without moving the selected subspaces, although their realized effects are small at this dose.

The balanced and native constructions use different factor scalings. The native control fixes the additional gauge scalar to one. Table 18 reports the common layers of the two modes, using random-sign perturbations and isotropic gradients. Each entry summarizes ten seeds.

Selected-block responses remain appreciable under native scaling, although their relative magnitudes change. Values above 100%, as in L1v–KPM, mean that an isolated response or reproduction error exceeds the full-response norm. They do not identify the source of cancellation or imply near-orthogonality. These controls support examining factor scaling as part of the update map.