NORMFOLD: FUNCTION-IDENTIFIABLE SPECTRAL ADAPTATION AT LAYERNORM INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral adapters derive trainable coordinates from a stored pretrained matrix, but storage coordinates need not be uniquely determined by the function. We expose this ambiguity at a LayerNorm followed by an affine map. The complete local invariant is (G, q), where $G = (I - \mathbf{1}\mathbf{1}^\top/m)\,\mathrm{diag}(\gamma)W$ and $q = \beta W + b$. We prove that two parameter tuples with the same (G, q) can induce spectral tangent spaces of different dimensions, so scale folding and offset preservation alone do not make spectral adaptation well defined on the pretrained function. NORMFOLD instead decomposes the unique minimum-Frobenius representative G, preserves q, and ties exactly repeated singular blocks. Its adapted family and matched optimizer trajectory are therefore determined by the local function in exact arithmetic. A controlled 9,120-fit study spans two CLIP checkpoints, three datasets, two shot counts, five support seeds, seven nonzero rewrite strengths, and five rewrite directions. Across all 2,100 matched nonzero task-direction conditions, NORMFOLD has lower final-logit discrepancy than the closest offset-preserving control. Its mean discrepancy remains between 1.03×10^{-7} and 1.05×10^{-7}, while that control ranges from 8.39×10^{-6} to 2.85×10^{-3}. NORMFOLD changes 9 of 10.88 million repeated paired predictions, compared with 27,145 for the control. Swin-T, seen-to-unseen transfer, fresh-image replay, merged export, and finite-precision analyses support the same function-identifiability account. The claim concerns whether the learning rule is representation independent, not whether NORMFOLD improves task accuracy over every spectral baseline.

1 INTRODUCTION

A pretrained weight matrix serves two roles during adaptation. It implements part of a function, and it supplies coordinates in which that function will be changed. Spectral parameter-efficient fine-tuning makes the second role explicit by decomposing a stored tensor and optimizing coefficients in its singular-vector basis (Han et al., 2023; Meng et al., 2024; Lingam et al., 2024; Zhang & Pilanci, 2024; Koleilat et al., 2025; Baek et al., 2026). This construction is only intrinsic when the decomposed tensor is uniquely determined by the pretrained function. Otherwise, two checkpoints that compute the same predictor can expose different trainable directions.

LayerNorm creates a concrete instance of this problem. Its standardized activations have zero channel mean. Consequently, a row-constant component of the following folded weight is invisible to the local forward map, yet that component can rotate the singular vectors selected by a spectral adapter. The resulting failure is not merely numerical. We give an exact two-dimensional construction in which two function-equivalent representatives induce projected spectral tangent spaces of different dimensions. Thus, an arbitrary storage choice can alter which first-order function updates are available.

We formalize the desired property as *function identifiability*: once the pretrained local map is fixed, an adapter-construction rule should define the same trainable function family at the same adapter coordinates. This criterion separates three operations that are often conflated. Folding the LayerNorm scale removes a scale-transfer ambiguity. Preserving the complete affine offset fixes the constant response. Projecting onto the centered input subspace removes the remaining unobservable matrix component. Only the third operation makes the spectral operator a function of the local map.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

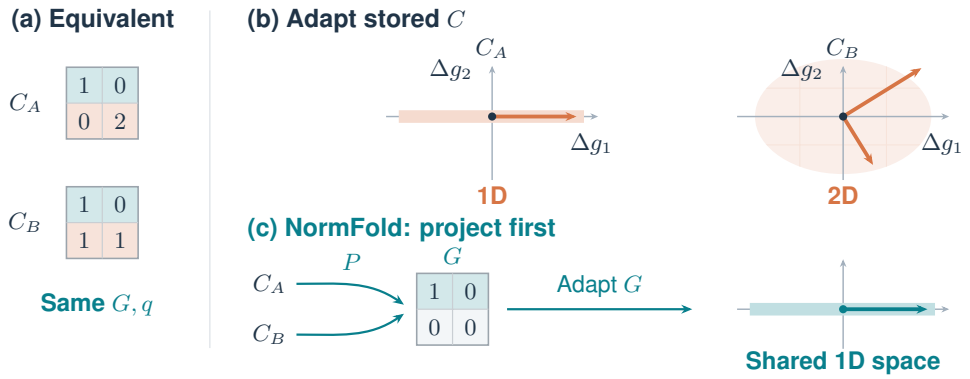

Figure 1: **Equivalent local functions can expose different spectral updates.** (a) The matrices in Equation 6 have the same observable operator G in an e-aligned basis, where $P = \mathrm{diag}(1, 0)$. (b) Spectral scaling of the stored representatives yields one-dimensional and two-dimensional first-order function-update spaces. Axes show the first row of ΔG, and shading indicates a span. (c) Projecting before decomposition produces the same one-dimensional function-update space for both representatives.

NORMFOLD implements this principle. It forms the folded operator, projects out the LayerNorm-null direction, decomposes the resulting canonical representative, preserves the complete offset, and assigns one multiplier to every exactly repeated nonzero singular block. We prove equality of the adapted family and of matched optimization trajectories in exact arithmetic. The learned increment can be merged back into the existing affine tensors, so the method adds no inference module.

The empirical design directly tests the theorem rather than relying on accuracy alone. We generate independently stored checkpoints that preserve the same initial interface function, then hold the support set, objective, optimizer schedule, and realized randomness fixed. The primary 60-task audit is expanded to seven nonzero rewrite strengths and five directions, producing 9,120 fits. The experiment asks whether learning amplifies a storage-only change. We then test the same criterion on a hierarchical architecture, a seen-to-unseen split, fresh full-image inference, merged-weight export, and controlled finite-precision variants.

Contributions. First, we characterize the complete local invariant of a LayerNorm-affine interface and identify its unique minimum-Frobenius representative. Second, we prove that spectral adaptation of a stored representative can change the dimension of the available function-update space, and we derive a function-identifiable alternative with an optimizer-level certificate. Third, we provide a causal experimental audit in which equivalent starting functions are trained independently. The resulting evidence distinguishes representative consistency from predictive accuracy and isolates the numerical boundary between exact function-space guarantees and floating-point execution.

2 RELATED WORK

Spectral and parameter-efficient adaptation. PEFT methods use bottleneck modules, low-rank updates, quantized adapters, random bases, or weight decompositions (Houlsby et al., 2019; Hu et al., 2021; Dettmers et al., 2023; Kopiczko et al., 2024; Liu et al., 2024). Spectral approaches instead derive coordinates from pretrained singular factors. SVDiff tunes singular values, while PiSSA, SVFT, Spectral Adapter, LaMDA, DeLoRA, PiCa, PSOFT, SumRA, CLIP-SVD, and IMSE adapt singular subspaces or related factors (Han et al., 2023; Meng et al., 2024; Lingam et al., 2024; Zhang & Pilanci, 2024; Azizi et al., 2024; Bini et al., 2025; Hwang et al., 2026; Wu et al., 2026; Kwok et al., 2026; Koleilat et al., 2025; Baek et al., 2026). LoFT and LoRA-RITE further show that parameterization and optimizer coordinates affect adaptation dynamics (Tastan et al., 2026; Yen et al., 2025). Our question precedes rank, sparsity, and optimizer choices: is the operator being decomposed itself determined by the pretrained function?

Reparameterization, symmetry, and normalization geometry. Deep networks contain scale, permutation, and more general parameter symmetries that affect optimization and comparisons in weight space (Badrinarayanan et al., 2015; Neyshabur et al., 2015; Dinh et al., 2017; Ainsworth et al., 2022; Entezari et al., 2022; Theus et al., 2026). LayerNorm standardizes across channels (Ba et al., 2016). SliceGPT exploits centering-aware transformations for compression (Ashkboos et al., 2024), algebraic dead-direction analysis identifies LayerNorm null directions (Shirodkar & Narayanan, 2026), and AffineQuant studies equivalent affine transformations (Ma et al., 2024). SITH folds LayerNorm into CLIP projections and removes the all-ones direction for data-free interpretation and weight editing (Gentile et al., 2026). We focus on a different consequence: a null direction can alter the learning family selected by a spectral adapter. Our result supplies the complete affine invariant, an exact dimension-changing obstruction, and a construction whose coordinates descend to the quotient defined by the local function.

Vision-language adaptation. CLIP aligns image and text representations (Radford et al., 2021). CoOp, CoCoOp, MaPLe, CLIP-Adapter, and Tip-Adapter adapt prompts, features, or cached representations (Zhou et al., 2022b;a; Khattak et al., 2023; Gao et al., 2024; Zhang et al., 2022). We use few-shot vision-language classification as a controlled environment for a different evaluation axis. The target is not a new accuracy state of the art. It is whether equivalent parameter tuples lead to the same learned predictor under a fixed adaptation rule.

3 METHOD

3.1 THE QUOTIENT VISIBLE TO THE LOCAL FUNCTION

Use row-oriented inputs $z \in \mathbb{R}^{1\times m}$ with $m \geq 2$. Let $\hat{z}$ denote LayerNorm standardization with fixed $\epsilon \geq 0$. Write $\Gamma = \mathrm{diag}(\gamma)$, $\beta \in \mathbb{R}^{1\times m}$, $W \in \mathbb{R}^{m\times d}$, and $b \in \mathbb{R}^{1\times d}$. The following affine consumer computes

$$y = (\hat{z}\Gamma + \beta)W + b = \hat{z}C + q, \qquad C = \Gamma W, \quad q = \beta W + b. \tag{1}$$

For a bias-free consumer, take $b = 0$. Define

$$e = \mathbf{1}/\sqrt{m}, \qquad P = I - ee^\top, \qquad G = PC. \tag{2}$$

Because $\hat{z}e = 0$, including when $\epsilon > 0$, the observable map is $y = \hat{z}G + q$. The input coordinates and standardization operation remain fixed throughout this local analysis.

Theorem 1 (Complete local invariant)**.** *Two affine parameter tuples of the same shapes agree on every attainable standardized input if and only if*

$$PC = PC' \quad \textit{and} \quad q = q'. \tag{3}$$

Thus, (G, q) completely identifies the local function. Invertibility of Γ is not required.

Proof. Sufficiency follows from $\hat{z}P = \hat{z}$. For necessity, equality at attainable inputs $\hat{z}$ and $-\hat{z}$ gives $q = q'$ and $\hat{z}(C - C') = 0$. Every centered direction has an attainable nonzero multiple, so these inputs span $e^\perp$ and $P(C - C') = 0$. The same spanning argument applies when $\epsilon = 0$, although constant inputs are excluded. Appendix A gives the full construction. □

The matrix representatives of one local function are exactly

$$\mathcal{E}(G) = \{C : PC = G\} = \{G + ea^\top : a \in \mathbb{R}^d\}. \tag{4}$$

The offset must also match because changing q cannot be compensated by centering a weight.

Proposition 1 (Canonical minimum-norm representative)**.** *G uniquely minimizes the Frobenius norm over $\mathcal{E}(G)$. For a centered increment ΔG, it also uniquely minimizes the folded update norm among all ΔC satisfying $P\Delta C = \Delta G$.*

Proof. Since $e^\top G = 0$, $\left\|G + ea^\top\right\|_F^2 = \|G\|_F^2 + \|a\|_2^2$. Equality with $\|G\|_F^2$ holds only for $a = 0$. The same orthogonal decomposition applies to updates. □

The minimum is defined in folded C-space. Stored W-space is rescaled by Γ.

Definition 1 (Local function identifiability)**.** *An adapter-construction rule is locally function-identifiable when any two parameter tuples that share the same local map produce the same adapted map for every common adapter coordinate, after coordinates are indexed by function-defined blocks rather than arbitrary bases inside repeated subspaces.*

3.2 WHY FOLDING AND OFFSET PRESERVATION ARE NOT ENOUGH

Let $M = U\,\mathrm{diag}(s)V^\top$ have simple nonzero singular values. Log-spectral multipliers define $M_\delta = U\,\mathrm{diag}(s \odot e^\delta)V^\top$. At $\delta = 0$, the projected first-order function directions span

$$\mathcal{T}_P(M) = \mathrm{span}\{s_i(Pu_i)v_i^\top : s_i > 0\}. \tag{5}$$

A function-identifiable family cannot change this tangent-space dimension when only the stored representative changes.

Proposition 2 (Representation-dependent spectral tangent space)**.** *There exist equivalent* 2×2 *matrices* C_A, C_B *with simple spectra such that* $\dim \mathcal{T}_P(C_A) = 1$ *and* $\dim \mathcal{T}_P(C_B) = 2$*.*

Proof. In an orthonormal input basis whose second axis is e, take

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad C_A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad C_B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{6}$$

Both matrices project to G. For C_A, P removes one singular component and leaves the first-row direction $[1, 0]$. For C_B, the eigenvalues of $C_B^\top C_B = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ are $(3 \pm \sqrt{5})/2$. Its two orthogonal right singular vectors have nonzero first coordinates. Since $s_i(u_i)_1 = (v_i)_1$, the projected components have first rows $(v_i)_1 v_i^\top$ and are linearly independent. □

The construction is exact and permits the same offset q. Applying P after decomposing C therefore differs from decomposing PC: the former selects directions using an unobservable component. Decomposing W additionally exposes the family to scale-transfer choices.

3.3 NORMFOLD: A SPECTRAL FAMILY ON THE LOCAL FUNCTION

NORMFOLD constructs spectral coordinates from G before decomposition. Let $G = U\,\mathrm{diag}(s)V^\top$ with descending singular values. Initialize log-multipliers δ to zero and define

$$G_\delta = U\,\mathrm{diag}(s \odot \exp(\delta))V^\top, \qquad y_\delta = \hat{z}G_\delta + q. \tag{7}$$

Exactly repeated nonzero singular values share one block multiplier δ_B, and zero singular values remain zero. Every nonzero left singular vector lies in $e^\perp$, so G_δ remains centered.

Proposition 3 (Adapted-family and trajectory certificate)**.** *Suppose two tuples share* (G, q) *and use the same exact singular-block assignment. Their* NORMFOLD *maps coincide for every common block-multiplier vector. If they also share the standardized support inputs, frozen downstream map, objective, initialization, optimizer state, update schedule, and realized randomness, then their multiplier gradients and optimization trajectories coincide in exact arithmetic.*

Proof. For a repeated block with value s_B, alternative SVD bases have the form $(U_B R, V_B R)$ for an orthogonal matrix R. A shared multiplier contributes $s_B e^{\delta_B} U_B R R^\top V_B^\top = s_B e^{\delta_B} U_B V_B^\top$. Sign choices cancel in the same way. The common forward map yields identical block derivatives, and equality of optimizer states propagates by induction. □

The proposition concerns exact multiplicity blocks. The implementation uses a declared finite-precision rank cutoff and grouping policy, whose sensitivity is evaluated in Appendix E. A common SVD seed cannot replace block tying because it does not remove basis freedom inside a repeated subspace. Appendix A.3 gives the block derivative and coordinate correspondence.

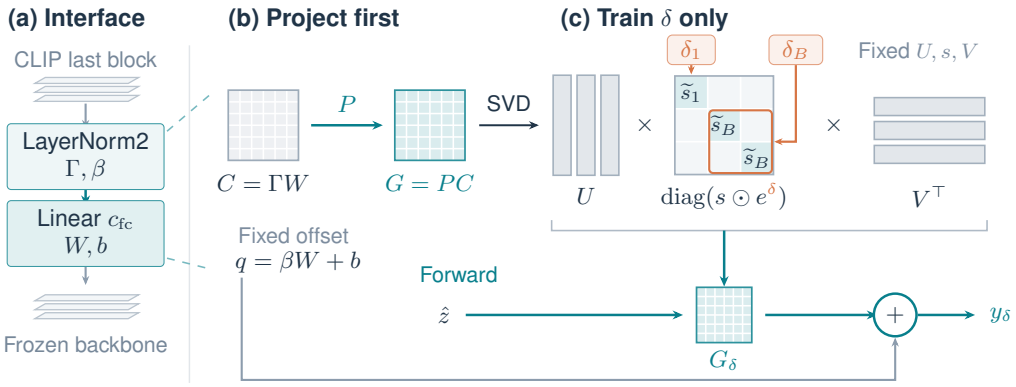

Figure 2: NORMFOLD **at a LayerNorm interface.** (a) The selected final-block CLIP interface. (b) Canonicalization maps $C = \Gamma W$ to $G = PC$ while preserving $q = \beta W + b$. (c) Only log-multipliers δ are trained. The factors U, s, V remain fixed, each repeated nonzero block B shares δ_B, and zero singular values stay zero.

Table 1: **Controlled spectral parameterizations.** Every method uses Equation 8 and the same baseline y_0. The last column is the fixed offset in $x_M M_\delta + o$. ScaleFold-SVA assumes invertible Γ.

Method	Decomposed M	Increment input x_M	Fixed o
Raw-SVA	W	$\hat{z}\Gamma + \beta$	b
ScaleFold-SVA	$C = \Gamma W$	$\hat{z} + \beta\Gamma^{-1}$	b
(C, q)-SVA	$C = \Gamma W$	$\hat{z}$	q
NORMFOLD	$G = PC$	$\hat{z}$	q

3.4 CONTROLLED PARAMETERIZATIONS AND DEPLOYMENT

We learn an increment rather than reconstructing an entire weight. For a decomposed matrix M, use

$$\Delta M_\delta = U \,\text{diag}(s \odot \text{expm1}(\delta))\, V^\top, \qquad y_\delta = y_0 + x_M \Delta M_\delta. \tag{8}$$

This produces exactly zero increment at initialization and preserves inactive baseline directions.

Raw-SVA retains stored coordinates. ScaleFold-SVA removes scale-transfer dependence but lets the folded offset participate in the update. (C, q)-SVA additionally fixes the complete offset. Its only operator-selection difference from NORMFOLD is whether centering occurs before or after the SVD.

For a bias-bearing consumer with invertible Γ, export uses

$$W_\delta = W + \Gamma^{-1}\Delta G_\delta, \qquad b_\delta = b - \beta\Gamma^{-1}\Delta G_\delta, \qquad \Delta G_\delta = G_\delta - G. \tag{9}$$

Then $(\hat{z}\Gamma + \beta)W_\delta + b_\delta = \hat{z}G_\delta + q$. Export modifies existing tensors and leaves the inference graph unchanged. If G has rank r, NORMFOLD trains at most $r \le \min(m - 1, d)$ scalars, with fewer parameters when blocks are tied. Centering costs $O(md)$ and the SVD is computed once. Appendix A covers zero scales, bias-free realizations, and shared consumers.

4 EXPERIMENTS

The experiments test one causal question: can an algebraically neutral change of the stored representative alter the learned predictor? Every A/B comparison starts from the same local function, prepares factors independently, trains both branches, and reports function discrepancy alongside task utility.

4.1 PROTOCOL AND EVALUATION CRITERIA

The primary study uses OpenAI CLIP ViT-B/16 and ViT-B/32 on CIFAR-100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), and Oxford-IIIT Pets (Parkhi et al., 2012). Both 4-shot and 16-

Table 2: **Representative consistency in the fixed-strength 60-task audit.** The rewrite uses $\rho = 0.25$ and direction seed 1729. Every base task receives equal weight.

Method	Tasks	Mean E	Max E	Mean D	Mean dis. (%)
Raw-SVA	60	3.02×10^{-3}	1.12×10^{-2}	1.01×10^{-1}	0.459
ScaleFold-SVA	60	3.30×10^{-3}	1.10×10^{-2}	1.15×10^{-1}	0.481
(C, q)-SVA	60	2.24×10^{-3}	5.34×10^{-3}	9.10×10^{-2}	0.386
NORMFOLD	60	1.04×10^{-7}	1.16×10^{-7}	5.25×10^{-6}	0.000

shot support sets are sampled with five seeds, giving 60 base tasks. Each fit adapts the final vision block's LayerNorm2 to MLP c_{fc} interface. All upstream weights, the nonlinear tail, LayerNorm affine parameters, text features, and the logit scale are frozen.

Independently loaded branch B rewrites natural branch A as

$$W' = W + \Gamma^{-1}\mathbf{1}a^\top, \qquad b' = b - (\beta\Gamma^{-1}\mathbf{1})a^\top, \tag{10}$$

which preserves (G, q) in exact arithmetic. The fixed-strength audit uses a float64 Gaussian direction with seed 1729 and $\rho = \left\|\mathbf{1}a^\top\right\|_F / \|C\|_F = 0.25$. The full sweep uses $\rho \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, 0.10, 0.25, 0.50\}$ and five nonzero direction seeds. The zero-strength condition is evaluated once. There are 240 natural A fits, 240 deterministic A repeats, and 8,640 independently prepared rewritten B fits, for 9,120 fits in total. This organization yields 8,640 A/B pairs and 240 A/A-repeat pairs.

Float32 Adam trains each adapter for 200 steps with batch size 32, learning rate 10^{-3}, and support cross-entropy plus $0.01\,\mathrm{mean}(\delta_B^2)$. Batches are matched across A and B. CPU float64 preparation retains $s_i > 10^{-7}s_1$ and transitively groups adjacent active values when $(s_i - s_{i+1})/s_1 \le 10^{-6}$. AMP and TF32 are disabled. Every branch prepares a fresh SVD and optimizer. Only the frozen CLS residual before the selected LayerNorm is cached, while the learned tail includes the nonlinear MLP, residual, final normalization, and classifier.

For complete final query-logit matrices, define

$$\begin{aligned} E &= \frac{\|L_{T,A} - L_{T,B}\|_F}{\max(\|L_{T,A}\|_F, \|L_{T,B}\|_F, 10^{-12})}, \\ D &= \frac{\|L_{T,A} - L_{T,B}\|_F}{\max(\|L_{T,A} - L_{0,A}\|_F, \|L_{T,B} - L_{0,B}\|_F, 10^{-12})}. \end{aligned} \tag{11}$$

E measures relative predictor discrepancy, while D normalizes that discrepancy by the learned change. We also report top-1 disagreement and natural-branch accuracy. The sweep intervals use 2,000 crossed-bootstrap replicates over support and direction labels, as specified in Appendix C. Query images are never treated as independent training replicates.

4.2 RQ1: DOES A FUNCTION-PRESERVING REWRITE CHANGE LEARNING?

The controls isolate the theoretical obstruction. Raw-SVA depends on the stored weight and its scale assignment. ScaleFold-SVA removes the scale-transfer ambiguity. (C, q)-SVA also fixes the complete offset. Nevertheless, all three controls remain sensitive because their singular bases are selected before the LayerNorm-null component is removed. In contrast, NORMFOLD stays at the scale of the initial float32 mismatch (Table 2).

All 480 A/B fits in this audit have nonzero first gradients and updates, so agreement is not caused by an inactive adapter. Initial A/B decisions are identical and initial E lies between 8.98×10^{-8} and 1.11×10^{-7}. The eight natural-checkpoint repeats are bitwise identical. After learning, NORMFOLD changes none of 310,980 repeated paired clean decisions. Each of its twelve model-dataset-shot cells improves over the frozen predictor by 0.39 to 1.12 percentage points, while its cell-mean accuracy differs from (C, q)-SVA by only -0.04 to $+0.03$ points. Thus, accuracy remains useful for confirming that adaptation occurred, but it does not diagnose whether the learning rule is well defined on the starting function.

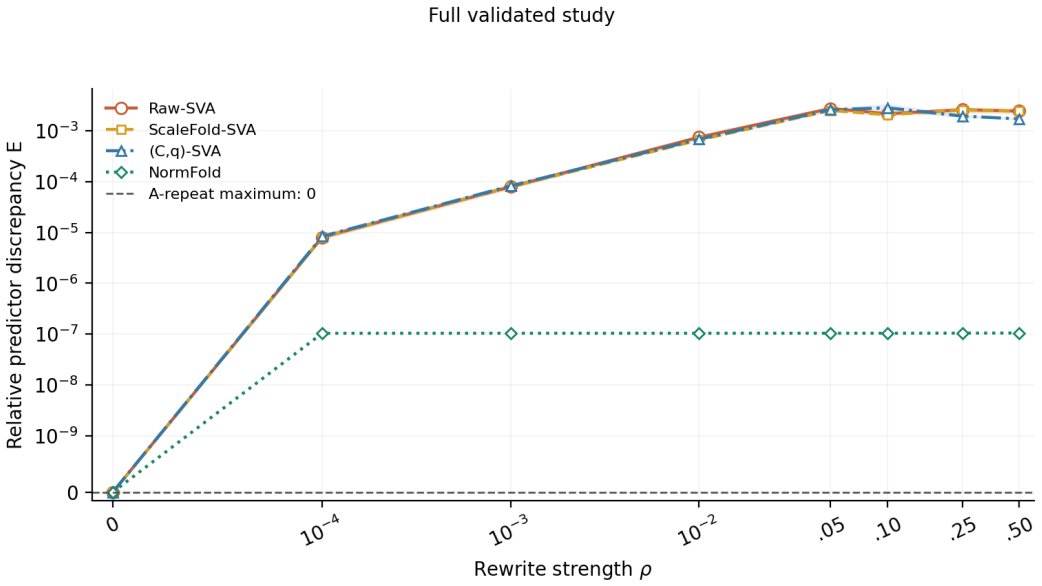

Figure 3: **Relative final-logit discrepancy across the complete rewrite-strength study.** Curves summarize all 60 base tasks and five nonzero directions, with pointwise 95% crossed-bootstrap intervals. The controls respond strongly but not monotonically to rewrite magnitude. NORMFOLD remains close to the initial float32 storage-error scale.

Table 3: **Selected points from the complete strength sweep.** A changed-pair count indicates at least one top-1 disagreement and is out of 300 pairs at each nonzero strength.

ρ	Mean E_{Cq}	Mean E_{NF}	Cq changed pairs	NF changed pairs
10^{-4}	8.39×10^{-6}	1.03×10^{-7}	34	1
10^{-3}	8.23×10^{-5}	1.04×10^{-7}	142	2
10^{-2}	6.85×10^{-4}	1.04×10^{-7}	289	1
0.10	2.85×10^{-3}	1.04×10^{-7}	300	0
0.50	1.72×10^{-3}	1.05×10^{-7}	299	3

4.3 RQ2: DOES THE SEPARATION PERSIST ACROSS REWRITE MAGNITUDE AND DIRECTION?

Across the seven nonzero strengths, NORMFOLD mean E remains between 1.0346×10^{-7} and 1.0474×10^{-7}. The closest control, (C, q)-SVA, reaches 8.3868×10^{-6} at $\rho = 10^{-4}$, peaks at 2.8450×10^{-3} at $\rho = 0.10$, and remains at 1.7204×10^{-3} at $\rho = 0.50$. NORMFOLD has lower E and lower D in every one of the 2,100 matched nonzero task-direction conditions. The ordering therefore does not depend on one dataset, support draw, rewrite direction, or stress-test magnitude.

The nonmonotone control curves show that the experiment is not a simple linear perturbation test. Changing the invisible component changes the learned coordinates, and the resulting optimizer path can vary with direction and magnitude. Across 10,884,300 repeated paired predictions per method, (C, q)-SVA changes 27,145 predictions in 1,664 task-direction pairs. NORMFOLD changes nine predictions in nine pairs, all associated with one base task and vanishing top-two margins. The 240 A/A-repeat pairs and 60 zero-strength A/B pairs are bitwise identical, whereas nonzero-strength NORMFOLD pairs retain positive discrepancies near 10^{-7}. The empirical claim is therefore not bitwise invariance. It is that learning does not amplify the storage-only change beyond the initial numerical scale. Appendix C reports every strength, direction, utility summary, and rare decision change.

Table 4: **Transfer of representative consistency.** Counts aggregate repeated query decisions over the declared task grids.

Setting	Tasks	Max E_{Cq}	Max E_{NF}	Cq diff.	NF diff.
Swin-T, CIFAR-100/Pets	12	3.28×10^{-3}	8.97×10^{-8}	181	0
CUB seen-to-unseen	12	4.48×10^{-3}	1.10×10^{-7}	405	0

Table 5: **Fresh-image replay and merged-weight export.** The fixed seed-0 subset uses both CLIP models, all three datasets, and both shot counts. Every test image is read again.

Compared outputs	Scope	Max E	Different decisions
Full vs. cached tail	48 adapters	1.38×10^{-7}	0
Full A vs. B, (C, q)-SVA	12 pairs	4.82×10^{-3}	193
Full A vs. B, NORMFOLD	12 pairs	1.52×10^{-7}	0
Merged vs. wrapper, NORMFOLD	12 adapters	1.55×10^{-7}	0

4.4 RQ3: DOES THE ACCOUNT TRANSFER BEYOND THE PRIMARY AUDIT?

Swin-T (Liu et al., 2021) changes the architecture and retains all spatial states at the final `norm2 -> mlp[0]` interface. Across 12 tasks, NORMFOLD preserves all 82,014 paired decisions, while (C, q)-SVA changes 181. CUB-200-2011 (Wah et al., 2011) uses the proposed 150-seen and 50-unseen split (Xian et al., 2017) and trains only on seen labels. Across another 12 tasks, NORMFOLD preserves all 56,772 paired decisions. Its seen/unseen harmonic accuracy improves over the frozen model by 0.29 to 0.53 points and stays within 0.021 points of (C, q)-SVA. These studies expand architectural and label scope without converting representative consistency into an accuracy ranking. Appendix D gives the complete protocols and utility tables.

Fresh-image replay reproduces the same gap outside the training cache, and ordinary merged tensors reproduce every wrapper decision in the export subset (Table 5). Numerical diagnostics separate the exact certificate from floating-point execution. Float32 preparation activates one additional numerical direction and raises the tested NORMFOLD maximum E from 1.00×10^{-7} to 7.95×10^{-7} without changing predictions. Coarser grouping preserves A/B agreement but reduces adaptation capacity, confirming that tying exact multiplicities and grouping nearby distinct values serve different roles. Appendix E reports these diagnostics.

Discrete predictions also depend on margins. Let m_i be branch A's top-two logit margin and $\eta_i = \|\ell_A(x_i) - \ell_B(x_i)\|_\infty$. Then

$$m_i > 2\eta_i \quad \Longrightarrow \quad \operatorname*{argmax}_k \ell_{A,k}(x_i) = \operatorname*{argmax}_k \ell_{B,k}(x_i). \tag{12}$$

On the complete CIFAR-100-C grid of 15 corruption types and five severities (Hendrycks & Dietterich, 2019), NORMFOLD changes 3 of 4.5 million repeated paired decisions, compared with 24,414 for (C, q)-SVA. The sufficient condition certifies 4,499,949 NORMFOLD decisions, and every observed flip lies among the remaining near ties. Appendix F preserves the aggregate corruption results and the three counterexamples.

4.5 INTERPRETATION AND SCOPE

Accuracy and function identifiability answer different questions. The A and B branches begin with the same classifier and retain nearly identical task accuracy, yet stored-matrix coordinates can learn measurably different functions. Accuracy can therefore conceal a change in the learning rule. Representative consistency matters when checkpoints are folded, converted, exported, quantized, or otherwise rewritten without an intended change in behavior.

NormFold is an operator-selection principle. NORMFOLD does not prescribe the downstream objective, rank budget, sparsity pattern, or optimizer of a complete spectral PEFT method. It replaces an arbitrary representative by the canonical function-visible operator before those choices are applied. The counterexample proves that preserving scale and offset after decomposition cannot

Table 6: **Decision-margin audit on CIFAR-100-C.** Counts aggregate repeated paired evaluations. An uncertified decision need not change.

Method	$m_i > 2\eta_i$	Uncertified	Different decisions
(C, q)-SVA	3,949,107	550,893	24,414
NORMFOLD	4,499,949	51	3

generally recover this property. The sweep shows that the difference appears even at the smallest tested nonzero rewrite and persists across every matched task and direction.

What the controlled design rules out. The initial-predictor check excludes a functional change caused by the rewrite itself. Nonzero first gradients and updates exclude an inactive-adapter explanation. Independent factor preparation excludes accidental reuse of A's spectral basis, while bitwise A/A repeats verify that the fixed execution path can reproduce itself. Fresh-image replay and merged export then show that the result is not created by the cached prefix or by a spectral wrapper. Taken together, these controls leave representative-dependent operator selection as the explanation that is shared by the theorem and the observed learning trajectories.

A broader quotient-space principle. The same construction applies whenever architecture supplies a known input subspace S and its projector P_S. The function-visible operator is then $P_S C$, and spectral constraints can be applied after this projection. This observation does not justify estimating an arbitrary subspace from a small support set, because estimation error would become part of the adapter definition. It instead identifies a reusable design rule for exact architectural invariances: canonicalize with the known quotient first, then apply the chosen rank, sparsity, regularization, and optimization scheme.

Limitations. The certificate is local to a LayerNorm-affine interface with fixed standardized-input coordinates. Appendix A.5 extends it to multiple affine consumers, but it is not a theorem for arbitrary global network symmetries. The controlled rewrite establishes a causal failure mode and its strength sensitivity; it does not estimate how often naturally produced checkpoints contain large null components. The empirical study isolates operator selection rather than reproducing every end-to-end spectral PEFT system. Finally, exact family equality does not imply bitwise equality after finite-precision SVD, matrix multiplication, or argmax at a vanishing decision margin. The reported numerical diagnostics retain these boundaries instead of treating them as contradictions of the exact statement.

5 CONCLUSION

Spectral adaptation is well posed only when its coordinates are defined by the pretrained function rather than an incidental storage representative. At a LayerNorm-affine interface, $(P\Gamma W, \beta W + b)$ is the complete local invariant. Function-equivalent representatives can nevertheless induce spectral tangent spaces of different dimensions when the SVD is applied before the invisible component is removed. NORMFOLD projects first, preserves the complete offset, and ties repeated singular blocks, which yields a function-identifiable family and a matched-trajectory certificate in exact arithmetic.

The controlled audit closes the corresponding empirical loop. Equivalent starting functions, independent factor preparation, active learning, rewrite-strength and direction sweeps, architectural transfer, fresh-image replay, merged export, and finite-precision boundary cases all support the same mechanism. Across every tested nonzero task-direction condition, NORMFOLD is more representative-consistent than the closest offset-preserving control while maintaining comparable utility. The practical implication is to choose the spectral operator on the quotient defined by the function before applying method-specific spectral constraints.

This perspective also suggests a sharper evaluation protocol for future PEFT methods. In addition to reporting accuracy and parameter count, a method can be tested on matched function-preserving rewrites. Failure under such a test reveals dependence on storage coordinates; success shows that the stated learning rule survives an algebraically neutral checkpoint transformation.

Reproducibility Statement

The main text specifies the assumptions, proofs, adapter construction, controls, rewrite grid, optimizer, and evaluation metrics. Appendix A gives the complete derivations and deployment conditions. Appendix B documents the primary datasets, support sampling, execution controls, complete cell-level outcomes, and full-image replay. Appendix C reports the exact rewrite-strength and direction study. Appendices D, E, and F provide transfer, numerical, and corruption analyses.

AI Use Statement

The authors used generative AI tools for language polishing, structural editing, and LaTeX formatting assistance. These tools were not used to design the method, conduct experiments, generate numerical results, or draw scientific conclusions. The authors verified all manuscript content and take full responsibility for it.

References

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pp. 11682–11701, 2024.

Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Lamda: Large model fine-tuning via spectrally decomposed low-dimensional adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9635–9646, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Vijay Badrinarayanan, Bamdev Mishra, and Roberto Cipolla. Understanding symmetries in deep networks. *arXiv preprint arXiv:1511.01029*, 2015.

Sunghyun Baek, Jaemyung Yu, Seunghee Koh, Minsu Kim, Hyeonseong Jeon, and Junmo Kim. Imse: Intrinsic mixture of spectral experts fine-tuning for test-time adaptation. *arXiv preprint arXiv:2603.07926*, 2026.

Massimo Bini, Leander Girrbach, and Zeynep Akata. Decoupling angles and strength in low-rank adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 20216–20233, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Francesco Gentile, Nicola Dall'Asen, Francesco Tonini, Massimiliano Mancini, Lorenzo Vaquero, and Elisa Ricci. From weights to concepts: Data-free interpretability of clip via singular vector decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2895–2906, 2026.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7289–7300. IEEE, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Junseo Hwang, Wonguk Cho, and Taesup Kim. Pica: Parameter-efficient fine-tuning with column space projection. In *International Conference on Learning Representations*, volume 2026, pp. 114267–114287, 2026.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122. IEEE, 2023.

Taha Koleilat, Hassan Rivaz, and Yiming Xiao. Clip-svd: Efficient and interpretable vision-language adaptation via singular values. *arXiv preprint arXiv:2509.03740*, 2025.

Dawid Kopiczko, Tijmen Blankevoort, and Yuki Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 6815–6835, 2024.

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

Chin Yuen Kwok, Yongsen Zheng, Jia Qi Yip, Kwok Yan Lam, and Ensiong Chng. Sumra: Parameter efficient fine-tuning with singular value decomposition and summed orthogonal basis. In *International Conference on Learning Representations*, volume 2026, pp. 91424–91436, 2026.

Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham K Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. Svft: Parameter-efficient fine-tuning with singular vectors. *Advances in neural information processing systems*, 37:41425–41446, 2024.

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. *arXiv preprint arXiv:2402.09353*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. Affinequant: Affine transformation quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 50932–50951, 2024.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 28, 2015.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Tejas Pradeep Shirodkar and PJ Narayanan. Algebraic dead directions in layernorm transformers: A forward-pass-only diagnostic at llm scale. *arXiv preprint arXiv:2606.19491*, 2026.

Nurbek Tastan, Stefanos Laskaridis, Martin Takáč, Karthik Nandakumar, and Samuel Horváth. Loft: Low-rank adaptation that behaves like full fine-tuning. In *International Conference on Learning Representations*, volume 2026, pp. 71294–71324, 2026.

Alexander Theus, Alessandro Cabodi, Sotiris Anagnostidis, Antonio Orvieto, Sidak Pal Singh, and Valentina Boeva. Generalized linear mode connectivity for transformers. *Advances in Neural Information Processing Systems*, 38:136197–136225, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Efficient orthogonal fine-tuning with principal subspace adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 25798–25826, 2026.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning—the good, the bad and the ugly. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3077–3086. IEEE, 2017.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Venkata Sai Surya Subramanyam Duvvuri, Inderjit Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. Lora done rite: Robust invariant transformation equilibration for lora optimization. In *International Conference on Learning Representations*, volume 2025, pp. 75815–75834, 2025.

Fangzhao Zhang and Mert Pilanci. Spectral adapter: Fine-tuning in spectral space. *arXiv preprint arXiv:2405.13952*, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

A Supporting Derivations and Extensions

A.1 Attainable standardized states and the counterexample basis

For $\bar{z} = m^{-1} z\mathbf{1}$, LayerNorm standardization is

$$\hat{z} = \frac{z - \bar{z}\mathbf{1}^\top}{\sqrt{m^{-1}\left\|z - \bar{z}\mathbf{1}^\top\right\|_2^2 + \epsilon}}. \tag{13}$$

For $\epsilon = 0$, constant inputs are excluded; for $\epsilon > 0$, the denominator is positive for every input. If $u \in e^\perp$ is nonzero, taking $z = tu$, $t \neq 0$, gives a nonzero multiple of u, and changing the sign of t reverses it. The attainable set is therefore sign-symmetric and spans $e^\perp$, which is the property used in Theorem 1. Its nonzero elements have norm $\sqrt{m}$ when $\epsilon = 0$ and norms below $\sqrt{m}$ when $\epsilon > 0$.

Equation 6 is written in an orthonormal input basis aligned with e. To express it in ordinary two-channel coordinates, choose an orthogonal matrix Q with second column $e = (1,1)^\top/\sqrt{2}$. Then $Q^\top PQ = \mathrm{diag}(1,0)$, and the stored matrices are $W_A = QC_A$ and $W_B = QC_B$, with $\Gamma = I$, $\beta = 0$, and common $b = q$. Left multiplication by Q preserves singular values, while $PQ = Q\,\mathrm{diag}(1,0)$ preserves the projected tangent-space dimensions. Thus the diagonal projector in Equation 6 is simply the channel-centering projector expressed in the e-aligned basis.

A.2 Scale transfer, zero scales, and bias-free realization

An invertible diagonal transfer

$$(\Gamma, \beta, W, b) \longmapsto (\Gamma D^{-1}, \beta D^{-1}, DW, b) \tag{14}$$

preserves C and q, so it also preserves (G, q). The null-direction rewrite in Equation 10 moves within the larger local equivalence class while leaving Γ, β fixed. Bias compensation is necessary because a weight-only rewrite can otherwise change q.

For a consumer without a bias tensor, a joint canonical realization is

$$\Gamma_\star = I, \qquad \beta_\star = e^\top, \qquad W_\star = G + eq, \qquad b_\star = 0. \tag{15}$$

Its output is $(\hat{z} + e^\top)(G + eq) = \hat{z}G + q$, using $\hat{z}e = 0$ and $e^\top G = 0$. Replacing G by G_δ gives the adapted map. This joint realization preserves the function while changing the stored tensors at zero multipliers. If the LayerNorm feeds several linear consumers, each affected consumer is rewritten with its own (G_j, q_j) to preserve the complete network output.

The complete-invariant theorem allows zero scales, whereas the inverse-scale deployment in Equation 9 requires invertible Γ. With a zero scale, the canonical joint realization remains available, although a fixed-Γ update may not realize every centered increment. Very small scales can amplify stored-space updates even when the folded increment is well conditioned; the minimum-norm statement therefore applies specifically to folded space.

A.3 Function-defined gradients and numerical blocks

Gradients in function-defined coordinates. For a minibatch of standardized states $\widehat{Z}$ and an exact nonzero singular block B, the derivative of the adapted interface is

$$\frac{\partial Y_\delta}{\partial \delta_B} = s_B e^{\delta_B} (\widehat{Z} U_B) V_B^\top. \tag{16}$$

This derivative is unchanged by a paired rotation of U_B and V_B. Backpropagating through the same frozen tail therefore gives the same block gradient, even when that tail is nonlinear. The regularizer used in the experiments adds $2\lambda\delta_B/K$ for K trainable blocks, which is also independent of the stored representative. This makes the trajectory argument concrete: it concerns the actual coordinates seen by the optimizer, not merely the existence of some functionally equivalent optimum. By contrast, the control derivative contains singular factors selected from W or C, and the counterexample shows that even its available local span can change.

The coordinate correspondence is indexed by distinct singular values and their exact multiplicities, not by an arbitrary ordering inside a repeated block. An implementation must preserve that correspondence when comparing gradients or initializing optimizer moments. We therefore evaluate representative dependence using complete output functions and decision counts, which remain comparable even when the control coordinates differ.

For an exactly repeated nonzero singular value, independent multipliers would choose an arbitrary basis within the repeated block. A common multiplier depends only on its block product, as shown in Proposition 3. A zero singular value remains zero under every multiplier. By contrast, grouping merely close but distinct values reduces capacity and is a numerical design choice. The learned rule is specified in Section 4.1; matched sensitivity results appear in Appendix E.

A.4 General subspace scope

Let the attainable row inputs be a sign-symmetric set spanning a linear subspace $S \subseteq \mathbb{R}^m$, and let P_S be its orthogonal projector. For $y = xC + q$, two pairs represent the same map on those inputs if and only if $P_S C = P_S C'$ and $q = q'$. Necessity follows by adding and subtracting equality at x and $-x$, then using spanning; sufficiency follows from $xP_S = x$. The matrix class is

$$\{C' : P_S C' = P_S C\} = \{P_S C + (I - P_S)A : A \in \mathbb{R}^{m \times d}\}. \tag{17}$$

Orthogonality gives $\|C'\|_F^2 = \|P_S C\|_F^2 + \|(I - P_S)A\|_F^2$, so $P_S C$ is the unique minimum-norm representative. Block-tied spectral adaptation of this projected operator inherits the same family-level certificate. LayerNorm supplies the architectural subspace $e^\perp$ exactly. Extending the result to a subspace estimated from finite support data, upstream coordinate rotations, or consumers beyond the modeled affine interface requires a separate analysis.

A.5 Shared LayerNorm with multiple linear consumers

Corollary 1 (Joint local invariant)**.** *Suppose J affine consumers receive the same fixed standardized input, $y_j = \hat{z}\Gamma W_j + q_j$, with $q_j = \beta W_j + b_j$. The tuple of their outputs is identified by*

$$G_{\text{joint}} = [P\Gamma W_1 \ \cdots \ P\Gamma W_J], \qquad q_{\text{joint}} = [q_1 \ \cdots \ q_J]. \tag{18}$$

Among representatives C_j satisfying $PC_j = G_j$, the unique minimizer of $\sum_j \|C_j\|_F^2$ is $C_j = G_j$ for every j.

Proof. Concatenate the outputs and apply Theorem 1. Writing $C_j = G_j + ea_j^\top$ gives $\sum_j \|C_j\|_F^2 = \sum_j \|G_j\|_F^2 + \sum_j \|a_j\|_2^2$. □

Preparing a separate NormFold spectral family for each G_j preserves the joint output tuple at matched multipliers. Any common frozen downstream function of that tuple consequently has identical outputs and gradients in exact arithmetic, including when the downstream function is nonlinear. The matched-optimization conditions of Proposition 3 still apply. A single SVD of the concatenation is another invariant family, but generally is *not* the same family as these separate SVDs.

Two rewrite operations must be distinguished. Equation 10 holds Γ, β fixed and compensates one consumer's bias; unaffected consumers are then unchanged. A scale transfer $(\Gamma, \beta) \mapsto (\Gamma D^{-1}, \beta D^{-1})$ instead requires $W_j \mapsto DW_j$ for *every* affected affine consumer. A nonlinear consumer of the affine-normalized state is not automatically covered. The corollary is a local extension, not a global transformer symmetry theorem or a new multi-branch adaptation architecture.

Executed synthetic implementation check. For each precision and each of ten seeds (0 to 9), we instantiate a width-24 LayerNorm with three linear consumers of output widths 16, 12, and 8, followed by a shared frozen GELU and five-output linear map. Using 96 synthetic input/target pairs, independently prepared adapters for the original and bias-compensated branches undergo 50 matched full-batch Adam steps (learning rate 0.01; squared-error loss plus $0.01\,\mathrm{mean}(\delta^2)$). Preparation uses float64 SVD in both execution precisions. These are mathematical implementation tests, not additional vision datasets or evidence of downstream classification gains. All 20 trials pass their predeclared numerical checks; full multiplier and gradient traces are retained.

Table 7: **Shared-consumer validation.** Symmetric relative errors, with maxima over ten seeds except the negative-control range. The negative control deliberately compensates only one of three consumers after changing the shared LayerNorm affine parameters. It is not a valid equivalent rewrite.

Comparison	Float64 execution	Float32 execution
Initial output tuple, null rewrite	2.65×10^{-16}	1.42×10^{-7}
Learned downstream outputs, step 50	8.54×10^{-15}	2.72×10^{-7}
Scale transfer, all consumers compensated	1.96×10^{-16}	1.07×10^{-7}
Scale transfer, only one compensated	$[0.252, 0.321]$	$[0.252, 0.298]$

B Primary Experimental Protocol and Complete Results

B.1 Original assets, official splits, and support sampling

Both public model files were verified against their published SHA-256 checksums. CIFAR-100 is loaded from its official Python archive after MD5 and SHA-256 validation. DTD and Oxford-IIIT Pets use the original JPEG bytes reconstructed from public mirrors and verified against SHA-256 manifests. The DTD label archive matches the official Oxford release, while the Pets identifiers and annotations are consistent with the official archive and its MD5 checksum. The loader preserves the official annotation order.

Table 8: **Expanded dataset protocol.** Support seeds resample within the listed training pool. Accuracy is per-image top-1 accuracy.

Dataset	Classes	Training pool	Test	Split rule
CIFAR-100	100	50,000	10,000	Official train/test
DTD	47	1,880	1,880	Partition 1; validation excluded
Pets	37	3,680	3,669	Official trainval/test

For each shot count and support seed, classes are traversed in ascending label order. A fresh NumPy `default_rng(seed)` selects the required number of distinct training indices without replacement in each class. An independent generator with the same seed concatenates permutations of support positions and takes the first 200×32 positions as batches. Support IDs, labels, positions, and hashes are saved before training. The same support and batch schedule are used across methods and A/B branches, and the same seed/shot sampling is used for both models. The two shot settings need not be nested draws. Prompts replace label underscores with spaces; Pets breed names are lowercased. The fixed single-prompt protocol deliberately does not tune dataset-specific prompt ensembles. Preprocessing is the checkpoint's deterministic 224-pixel bicubic resize, center crop, RGB conversion, tensor conversion, and CLIP normalization; no random image augmentation is used.

B.2 Execution identity and independent learning

Every formal fit uses Adam with learning rate 10^{-3}, betas $(0.9, 0.999)$, epsilon 10^{-8}, zero weight decay, no scheduler, and `foreach=False`; the batch size is 32 and the schedule has 200 steps. The objective is mean support cross-entropy plus $0.01\,\mathrm{mean}(\delta_B^2)$, with a fixed final checkpoint. The runtime is Python 3.10.20, PyTorch 2.6.0+cu124, torchvision 0.21.0+cu124, and `openai-clip` 1.0.1 on one RTX 3090 Ti (24 GB). Deterministic algorithms are enabled, cuDNN benchmarking, TF32 and AMP are disabled, and the CUDA workspace configuration is fixed before PyTorch initialization. The resolved configuration, imported source files, and their hashes are snapshotted before the first formal fit. Resume rejects changed configuration or source. Failed attempts are retained in separate directories; a successful result selects a completed attempt by identity, never by query accuracy.

Within each model/dataset/branch, only frozen prefix extraction is shared across support tasks. Every fit calls a new SVD and starts zero multipliers and an empty Adam state. Branches independently load the model and extract their prefix, including A' with the same image union and batch layout as A. The first method-specific fit in each branch is tested on eight real support images for zero/nonzero

Table 9: Mean relative discrepancy E and adaptation-normalized discrepancy D over all paired support draws. Each cell contains five paired support draws.

Model	Dataset	Shot	$Cq\ E$	NF E	$Cq\ D$	NF D
B/16	CIFAR-100	4	2.61×10^{-3}	9.98×10^{-8}	8.61×10^{-2}	3.38×10^{-6}
B/16	CIFAR-100	16	2.50×10^{-3}	1.00×10^{-7}	8.17×10^{-2}	3.35×10^{-6}
B/16	DTD	4	1.73×10^{-3}	9.19×10^{-8}	1.95×10^{-1}	1.03×10^{-5}
B/16	DTD	16	1.35×10^{-3}	9.10×10^{-8}	1.74×10^{-1}	1.16×10^{-5}
B/16	Pets	4	4.58×10^{-3}	1.05×10^{-7}	7.55×10^{-2}	1.81×10^{-6}
B/16	Pets	16	4.42×10^{-3}	1.04×10^{-7}	7.69×10^{-2}	1.89×10^{-6}
B/32	CIFAR-100	4	1.45×10^{-3}	1.11×10^{-7}	7.44×10^{-2}	5.87×10^{-6}
B/32	CIFAR-100	16	1.14×10^{-3}	1.10×10^{-7}	5.35×10^{-2}	5.25×10^{-6}
B/32	DTD	4	1.47×10^{-3}	1.04×10^{-7}	1.12×10^{-1}	8.27×10^{-6}
B/32	DTD	16	1.24×10^{-3}	1.03×10^{-7}	8.55×10^{-2}	7.28×10^{-6}
B/32	Pets	4	2.53×10^{-3}	1.14×10^{-7}	4.49×10^{-2}	2.06×10^{-6}
B/32	Pets	16	1.88×10^{-3}	1.14×10^{-7}	3.25×10^{-2}	1.99×10^{-6}

Table 10: Paired utility differences on the natural checkpoint. Values are means and 95% percentile intervals over support seeds (20,000 resamples; fixed seed 20260910). Five-seed intervals do not establish equivalence or noninferiority.

Model	Dataset	Shot	NF−base	95% CI	NF−Cq	95% CI
B/16	CIFAR-100	4	+0.51	[0.45, 0.56]	+0.03	[0.02, 0.05]
B/16	CIFAR-100	16	+0.51	[0.47, 0.55]	+0.02	[0.01, 0.03]
B/16	DTD	4	+1.12	[0.98, 1.26]	-0.02	[-0.05, 0.02]
B/16	DTD	16	+1.05	[1.00, 1.12]	-0.04	[-0.05, -0.02]
B/16	Pets	4	+0.46	[0.30, 0.61]	+0.00	[-0.04, 0.04]
B/16	Pets	16	+0.56	[0.52, 0.59]	-0.03	[-0.08, 0.01]
B/32	CIFAR-100	4	+0.46	[0.41, 0.52]	-0.03	[-0.03, -0.02]
B/32	CIFAR-100	16	+0.52	[0.49, 0.54]	-0.01	[-0.02, 0.00]
B/32	DTD	4	+0.39	[0.30, 0.48]	-0.04	[-0.10, 0.01]
B/32	DTD	16	+0.43	[0.24, 0.60]	-0.02	[-0.04, 0.00]
B/32	Pets	4	+0.99	[0.83, 1.10]	-0.01	[-0.03, 0.03]
B/32	Pets	16	+1.00	[0.99, 1.02]	-0.03	[-0.05, -0.02]

outputs, loss, multiplier gradient, and agreement with the explicit affine formula. Every fit must have a nonzero first gradient and parameter update. The fixed numerical gate tolerance is 10^{-5}; this is an implementation check, not a theorem about all finite-precision training runs. Checkpoints and optimizer states are saved at steps 0, 1, 10, 50, 100, and 200; every step records the loss, gradients, multipliers, and a fixed support probe. Timing logs reflect the actual workload, including concurrent execution of part of the numerical diagnostic; they are not isolated comparative throughput measurements.

B.3 Complete outcome and uncertainty tables

The following tables contain every model, dataset, shot, method, and support-seed condition in the fixed-strength audit. Paired bootstrap intervals condition on the pretrained model, test set, prompt, and training schedule. The equal-weight aggregate summarizes the declared task grid.

Table 11: All methods on all predeclared cells. Accuracies are mean ± sample SD; disagreement is mean percentage.

Model	Dataset	Shot	Method	A acc.	B acc.	Dis.	E	D
B/16	CIFAR-100	4	(C, q)-SVA	68.84±0.06	68.85±0.08	0.256	2.61×10^{-3}	8.61×10^{-2}
B/16	CIFAR-100	4	NormFold	68.87±0.07	68.87±0.07	0.000	9.98×10^{-8}	3.38×10^{-6}
B/16	CIFAR-100	4	Raw-SVA	68.65±0.04	68.69±0.04	0.394	1.46×10^{-3}	5.58×10^{-2}
B/16	CIFAR-100	4	ScaleFold-SVA	68.67±0.06	68.70±0.06	0.336	3.13×10^{-3}	1.38×10^{-1}

Table 11: All methods on all predeclared cells (continued).

Model	Dataset	Shot	Method	A acc.	B acc.	Dis.	E	D
B/16	CIFAR-100	16	(C, q)-SVA	68.85±0.04	68.85±0.05	0.286	2.50×10^{-3}	8.17×10^{-2}
B/16	CIFAR-100	16	NORMFOLD	68.87±0.05	68.87±0.05	0.000	1.00×10^{-7}	3.35×10^{-6}
B/16	CIFAR-100	16	Raw-SVA	68.65±0.05	68.71±0.04	0.290	1.32×10^{-3}	4.98×10^{-2}
B/16	CIFAR-100	16	ScaleFold-SVA	68.68±0.03	68.69±0.04	0.288	3.07×10^{-3}	1.37×10^{-1}
B/16	DTD	4	(C, q)-SVA	44.38±0.14	44.34±0.14	0.819	1.73×10^{-3}	1.95×10^{-1}
B/16	DTD	4	NORMFOLD	44.36±0.18	44.36±0.18	0.000	9.19×10^{-8}	1.03×10^{-5}
B/16	DTD	4	Raw-SVA	44.49±0.13	44.46±0.12	0.904	2.16×10^{-3}	2.27×10^{-1}
B/16	DTD	4	ScaleFold-SVA	44.46±0.24	44.22±0.14	0.979	2.27×10^{-3}	2.27×10^{-1}
B/16	DTD	16	(C, q)-SVA	44.34±0.06	44.28±0.08	0.426	1.35×10^{-3}	1.74×10^{-1}
B/16	DTD	16	NORMFOLD	44.30±0.08	44.30±0.08	0.000	9.10×10^{-8}	1.16×10^{-5}
B/16	DTD	16	Raw-SVA	44.38±0.14	44.43±0.16	0.777	1.85×10^{-3}	2.29×10^{-1}
B/16	DTD	16	ScaleFold-SVA	44.33±0.14	44.20±0.07	0.915	2.06×10^{-3}	2.28×10^{-1}
B/16	Pets	4	(C, q)-SVA	88.72±0.15	88.69±0.19	0.251	4.58×10^{-3}	7.55×10^{-2}
B/16	Pets	4	NORMFOLD	88.72±0.20	88.72±0.20	0.000	1.05×10^{-7}	1.81×10^{-6}
B/16	Pets	4	Raw-SVA	88.63±0.11	88.74±0.15	0.567	9.50×10^{-3}	1.32×10^{-1}
B/16	Pets	4	ScaleFold-SVA	88.72±0.18	88.66±0.20	0.507	9.97×10^{-3}	1.36×10^{-1}
B/16	Pets	16	(C, q)-SVA	88.84±0.08	88.87±0.05	0.245	4.42×10^{-3}	7.69×10^{-2}
B/16	Pets	16	NORMFOLD	88.81±0.05	88.81±0.05	0.000	1.04×10^{-7}	1.89×10^{-6}
B/16	Pets	16	Raw-SVA	88.77±0.13	88.83±0.08	0.501	9.67×10^{-3}	1.38×10^{-1}
B/16	Pets	16	ScaleFold-SVA	88.79±0.07	88.80±0.12	0.567	9.74×10^{-3}	1.40×10^{-1}
B/32	CIFAR-100	4	(C, q)-SVA	64.97±0.07	64.96±0.05	0.284	1.45×10^{-3}	7.44×10^{-2}
B/32	CIFAR-100	4	NORMFOLD	64.94±0.07	64.94±0.07	0.000	1.11×10^{-7}	5.87×10^{-6}
B/32	CIFAR-100	4	Raw-SVA	65.06±0.07	65.08±0.07	0.342	1.35×10^{-3}	5.77×10^{-2}
B/32	CIFAR-100	4	ScaleFold-SVA	65.01±0.05	65.00±0.03	0.252	1.33×10^{-3}	6.61×10^{-2}
B/32	CIFAR-100	16	(C, q)-SVA	65.01±0.04	65.01±0.02	0.276	1.14×10^{-3}	5.35×10^{-2}
B/32	CIFAR-100	16	NORMFOLD	65.00±0.04	65.00±0.04	0.000	1.10×10^{-7}	5.25×10^{-6}
B/32	CIFAR-100	16	Raw-SVA	65.08±0.03	65.09±0.01	0.318	1.41×10^{-3}	5.82×10^{-2}
B/32	CIFAR-100	16	ScaleFold-SVA	64.98±0.03	64.97±0.03	0.296	1.03×10^{-3}	4.92×10^{-2}
B/32	DTD	4	(C, q)-SVA	43.73±0.12	43.61±0.25	0.681	1.47×10^{-3}	1.12×10^{-1}
B/32	DTD	4	NORMFOLD	43.69±0.12	43.69±0.12	0.000	1.04×10^{-7}	8.27×10^{-6}
B/32	DTD	4	Raw-SVA	43.89±0.17	43.85±0.23	0.660	1.38×10^{-3}	1.02×10^{-1}
B/32	DTD	4	ScaleFold-SVA	43.71±0.23	43.69±0.22	0.691	1.45×10^{-3}	1.00×10^{-1}
B/32	DTD	16	(C, q)-SVA	43.74±0.20	43.65±0.15	0.745	1.24×10^{-3}	8.55×10^{-2}
B/32	DTD	16	NORMFOLD	43.72±0.23	43.72±0.23	0.000	1.03×10^{-7}	7.28×10^{-6}
B/32	DTD	16	Raw-SVA	43.85±0.10	43.85±0.12	0.426	1.09×10^{-3}	8.00×10^{-2}
B/32	DTD	16	ScaleFold-SVA	43.69±0.24	43.61±0.14	0.617	1.15×10^{-3}	8.24×10^{-2}
B/32	Pets	4	(C, q)-SVA	86.06±0.17	86.07±0.14	0.218	2.53×10^{-3}	4.49×10^{-2}
B/32	Pets	4	NORMFOLD	86.06±0.18	86.06±0.18	0.000	1.14×10^{-7}	2.06×10^{-6}
B/32	Pets	4	Raw-SVA	85.97±0.14	85.94±0.15	0.245	2.77×10^{-3}	4.73×10^{-2}
B/32	Pets	4	ScaleFold-SVA	85.96±0.19	86.00±0.14	0.185	2.57×10^{-3}	4.50×10^{-2}
B/32	Pets	16	(C, q)-SVA	86.10±0.02	86.09±0.04	0.147	1.88×10^{-3}	3.25×10^{-2}
B/32	Pets	16	NORMFOLD	86.07±0.02	86.07±0.02	0.000	1.14×10^{-7}	1.99×10^{-6}
B/32	Pets	16	Raw-SVA	85.95±0.09	85.99±0.08	0.087	2.27×10^{-3}	3.76×10^{-2}
B/32	Pets	16	ScaleFold-SVA	86.00±0.04	86.05±0.05	0.142	1.88×10^{-3}	3.22×10^{-2}

B.4 FRESH-IMAGE REPLAY AND MERGED-WEIGHT EXPORT

The fixed replay subset contains support seed 0, both shots, both models, all three datasets, (C, q)-SVA and NORMFOLD, and A/B: 48 final adapters. The replay independently reloads the public checkpoint, reproduces the rewrite and text classifier, and prepares fresh factors. It verifies the factor hashes and block assignments against training metadata before restoring the final multipliers. It reads every test image again and never reads the training prefix cache. A live all-token prefix may be shared by the four adapters evaluated within the same model/dataset/branch; each adapter still evaluates its full tokenwise MLP and downstream path. An eight-image check compares this shared live path with a separate call to the complete encoder implementation.

For the 12 NORMFOLD 4-shot adapters in that subset, export also evaluates Equation 9. It forms the increment in float64, merges it into the existing W, b, casts those tensors to float32, and uses an ordinary LayerNorm, linear, and MLP path without the spectral wrapper. The original stored tensors are checked to be unchanged at zero update. Nonzero export error and decisions are compared with the learned wrapper on the full test set, retaining any rounding-induced discrepancy. These checks cover the declared subset, not every adapter or arbitrary upstream interfaces.

All 48 replays satisfy the full-versus-cached gate, with maximum relative logit error 1.3794×10^{-7} and zero prediction discrepancies. All 12 merged exports pass, with maximum wrapper-versus-merged error 1.5490×10^{-7} and zero prediction discrepancies. Per-adapter and 24 replay A/B-pair records are retained. Preprocessing identity is checked by a regression-tested serialization that removes process-specific memory addresses while preserving function names and transform parameters.

C COMPLETE REWRITE-STRENGTH AND DIRECTION STUDY

C.1 FULL PREDECLARED DESIGN

The main strength study varies the magnitude and direction of the same bias-compensated LayerNorm-null rewrite used in Equation 10. We define

$$\rho = \frac{\left\|\mathbf{1}a^\top\right\|_F}{\|C\|_F} \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, 0.10, 0.25, 0.50\}. \tag{19}$$

Each nonzero strength uses five float64 Gaussian direction seeds, 1729, 2718, 3141, 5772, and 7919. Zero strength is evaluated once because its direction is immaterial. The study crosses two CLIP checkpoints, three datasets, two shot counts, five support seeds, four methods, and independently prepared A/B branches. Together with 240 natural-checkpoint reference fits and 240 deterministic A-repeat fits, the design contains 9,120 fits, 8,640 A/B pairs, and 240 A/A-repeat pairs.

The optimizer, support draws, batches, preparation precision, rank cutoff, block rule, prompt, and final-step evaluation match Section 4.1. Model, dataset, and shot strata are fixed. Pointwise 95% intervals use 2,000 crossed-bootstrap replicates: support labels are resampled within each dataset and paired across models, shots, methods, and strengths, while the five direction labels are resampled independently and crossed with those support draws. Query images are not treated as independent bootstrap replicates.

The 480 conditions at $\rho = 0.25$ and direction 1729 reproduce the original formal records bitwise in final dense logits and all recorded multiplier steps. The final validation retains the earlier probe diagnostics and uses the first actual 32-example training minibatch as its primary same-input implementation check. All rule revisions and prior diagnostic flags were retained in the experiment records rather than discarded.

C.2 CONTINUOUS PREDICTOR DISCREPANCY AND EXACT SUMMARIES

Figure 3 in the main text shows the complete mean-E curves. For NORMFOLD, the mean nonzero-strength discrepancy lies between 1.0346×10^{-7} and 1.0474×10^{-7}. The closest control, (C, q)-SVA, rises from 8.3868×10^{-6} at $\rho = 10^{-4}$ to an aggregate maximum of 2.8450×10^{-3} at $\rho = 0.10$, then decreases to 1.7204×10^{-3} at $\rho = 0.50$. Raw-SVA and ScaleFold-SVA show the same broad separation but not a globally monotone response. In all 2,100 matched nonzero task-direction conditions, NORMFOLD has lower E and lower D than the corresponding (C, q)-SVA pair.

C.3 DISCRETE DECISIONS, UTILITY, AND DIRECTION EFFECTS

Across the seven nonzero strengths, each method has 10,884,300 repeated paired query evaluations. (C, q)-SVA changes 27,145 predictions across 1,664 of 2,100 task-direction pairs. NORMFOLD changes nine predictions across nine pairs. Each NORMFOLD pair contains one changed prediction among 3,669 Pets queries, and all nine pairs share the B/16, Pets, 16-shot, support-seed-2 base task. These repeated evaluations are descriptive counts rather than independent trials. They establish that

Table 12: **Exact full-sweep summaries for the closest control and NormFold.** Changed-pair counts are out of 300 at each nonzero strength and out of 60 at zero. E and D are means over task-direction pairs, not pooled-query ratios.

ρ	E_{Cq}	E_{NF}	D_{Cq}	D_{NF}	Cq pairs	NF pairs
0	0	0	0	0	0	0
10^{-4}	8.39×10^{-6}	1.03×10^{-7}	3.90×10^{-4}	5.1×10^{-6}	34	1
10^{-3}	8.23×10^{-5}	1.04×10^{-7}	3.81×10^{-3}	5.1×10^{-6}	142	2
10^{-2}	6.85×10^{-4}	1.04×10^{-7}	3.16×10^{-2}	5.1×10^{-6}	289	1
0.05	2.61×10^{-3}	1.04×10^{-7}	1.15×10^{-1}	5.2×10^{-6}	300	1
0.10	2.85×10^{-3}	1.04×10^{-7}	1.27×10^{-1}	5.2×10^{-6}	300	0
0.25	1.96×10^{-3}	1.04×10^{-7}	8.65×10^{-2}	5.2×10^{-6}	300	1
0.50	1.72×10^{-3}	1.05×10^{-7}	7.64×10^{-2}	5.3×10^{-6}	299	3

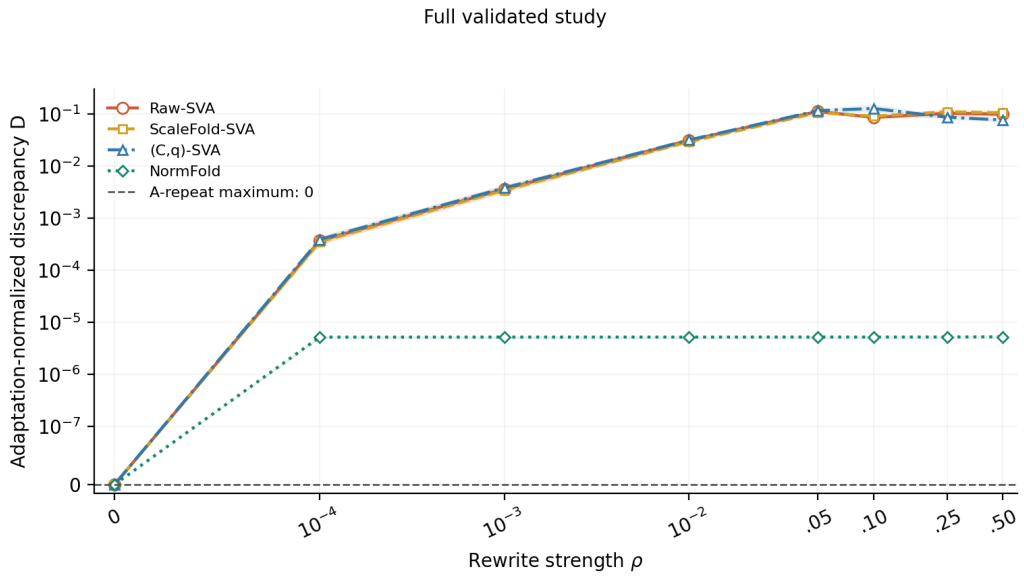

Figure 4: **Discrepancy normalized by the learned update magnitude.** D uses each fit's own initial logits. No main-study or repeat pair activates the 10^{-12} denominator clamp.

finite-precision decision identity is not universal, while the continuous discrepancy remains close to the initial storage-error scale.

Within-method A/B accuracy shifts remain small: the largest absolute method-strength mean is 0.0275 percentage points. Thus the sweep tests representation consistency rather than a between-method accuracy advantage. Direction-specific results in Figure 6 retain all five declared directions. The nearly overlapping NormFold curves are not jittered, while the controls show substantially larger direction sensitivity.

D Cross-Architecture and Unseen-Class Transfer

These experiments apply the same representative-consistency criterion across a hierarchical architecture and a seen-to-unseen label split. Every predeclared seed and method is included, with protocols fixed before execution and complete utility results reported below.

D.1 A distinct hierarchical architecture

We use torchvision's public ImageNet-1K Swin-T checkpoint, with the final `norm2 -> mlp[0]` interface. Unlike the CLIP CLS-tail protocol, all $7 \times 7 \times 768$ real spatial states are retained. Gra-

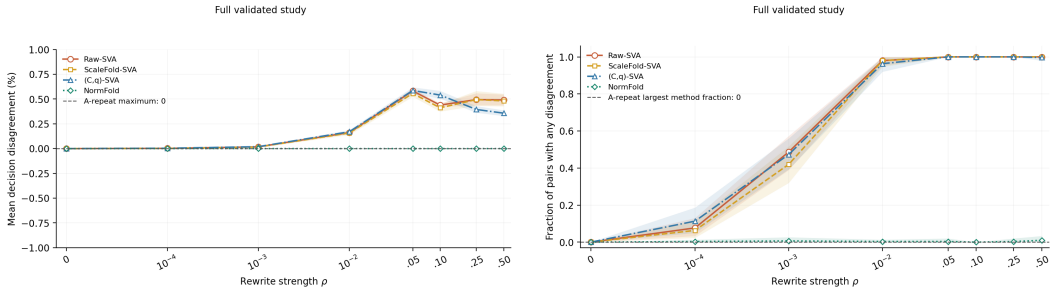

Figure 5: **Discrete prediction sensitivity.** Left: mean query-level disagreement. Right: fraction of task-direction pairs containing at least one changed prediction. Nonzero points contain 300 pairs per method; zero-strength points contain 60.

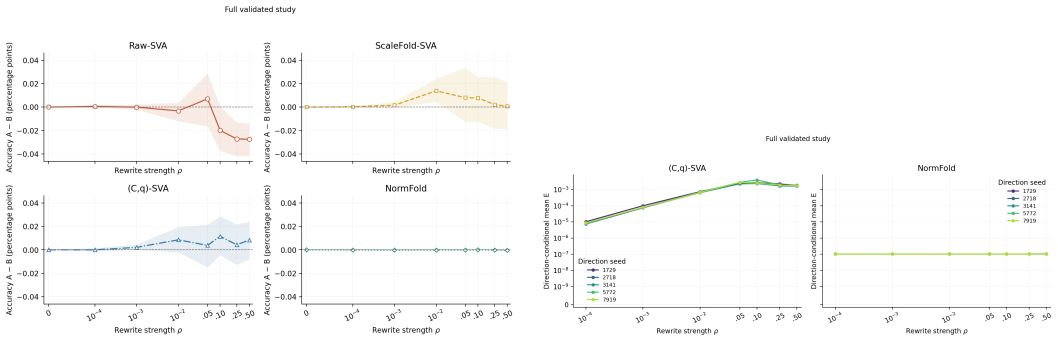

Figure 6: **Utility and direction sensitivity.** Left: signed within-method accuracy difference, natural A minus rewritten B. Right: mean E for all five nonzero rewrite directions. Intervals containing zero are descriptive and do not establish equivalence.

dients propagate through the frozen MLP, residual, final normalization and spatial average; only the spectral multipliers are updated. Fixed cosine prototypes are computed only from each task's original-branch support set, normalized, and shared across methods and A/B; temperature 30 is fixed before query encoding. No query or ImageNet classifier weight forms a prototype. This is supervised-pretrained few-shot transfer, not CLIP zero-shot inference or an isolated backbone-only comparison.

CIFAR-100 and Pets, 4/16 shots, seeds 0 to 2, four methods and two branches give 96 fits. The 200-step optimizer, rewrite, preparation, active cutoff and transitive relative-gap rule match the main protocol. Branches independently load weights and extract full spatial states. Every fit has zero/nonzero full-image versus cached-tail and formula gates, and a final learned-adapter check on eight support images. Query predictions use the complete spatial cached tail; this is not a full-query image replay.

Table 13: **Swin-T utility, all cells.** Mean ± sample SD over three support draws; differences use the natural A branch. No superiority or non-inferiority test is implied.

Dataset	Shots	NormFold accuracy (%)	Difference from (C, q) (pp)	Difference from initial (pp)
CIFAR-100	4	51.2533 ± 0.4972	$+0.0133 \pm 0.0252$	$+0.3633 \pm 0.0709$
CIFAR-100	16	61.7967 ± 0.5459	$+0.0033 \pm 0.0252$	$+0.3667 \pm 0.1537$
Pets	4	87.4171 ± 1.3726	-0.0182 ± 0.0157	$+0.1454 \pm 0.1731$
Pets	16	90.3698 ± 0.4539	-0.0091 ± 0.0157	$+0.1999 \pm 0.1597$

The natural NormFold minus (C, q) differences across all 12 support tasks are three positive, five negative and four exactly zero, ranging from -0.0273 to $+0.0400$ points. Stability transfers to this architecture; accuracy superiority does not follow. All 768 Raw-SVA singular values remain active

Table 14: **Swin-T representative sensitivity.** Each method has 12 task pairs and 82,014 repeated query decisions.

Method	Max E	Max D	Different decisions
Raw-SVA	9.18×10^{-3}	0.3611	628
ScaleFold-SVA	4.10×10^{-3}	0.1860	206
(C, q)-SVA	3.28×10^{-3}	0.2052	181
NORMFOLD	8.97×10^{-8}	8.24×10^{-6}	0

in both branches, but the relative grouping rule produces 768 groups in A and 302 in B: the rewrite changes the leading singular scale from approximately 19.90 to 4823.89. The 302 value denotes groups, not a truncation to 302 active directions. NORMFOLD has 767 groups in both branches. The LayerNorm's smallest scale is 3.42×10^{-4}; we retain this conditioning rather than changing the cutoff after observing results.

D.2 SEEN-SUPERVISED TRANSFER TO UNSEEN CLASSES

We use CUB-200-2011 (Wah et al., 2011) with the proposed split of Xian et al. (2017): 7,057 trainval images from 150 seen classes, 1,764 seen-query images, and 2,967 query images from 50 unseen classes. Official MATLAB metadata supply filenames and labels only, never stored ResNet features. Their class ordering is mapped explicitly to original image-folder names. The original archive matches its published MD5; all 11,788 images are decoded and checked for exact byte/pixel duplicates. Unseen denotes absence from *adaptation supervision*; CLIP's web pretraining cannot be audited for category or image overlap. This is not a claim of pretraining-unseen or attribute-only GZSL state of the art.

Both CLIP models use four/16 examples per seen class and three support seeds, the four methods, and independently loaded A/B branches: 96 200-step fits. All hyperparameters and the final visual interface match the main protocol. The training cross-entropy includes only the 150 seen columns with explicit global-to-seen label mapping; an implementation test verifies zero gradient on unseen logits. All 200 fixed text prompts are used as candidates for generalized evaluation, without calibrated stacking, query calibration, checkpoint selection, or extra training. Each fit passes the full-image/cached-tail implementation gates and a learned-update check on eight support images. Query predictions use the real cached tail; this is not a complete-query image replay.

Table 15: **CUB proposed-split transfer.** A-branch per-class seen accuracy S, unseen accuracy U, and $H = 2SU/(S + U)$, in percent. Mean $\pm$ sample SD over three support seeds; paired NORMFOLD $-$ (C, q) harmonic-mean differences in percentage points. All four cells are shown.

Model	Shots	S	U	H	ΔH
B/16	4	$54.7452 \pm .2085$	$55.7659 \pm .2769$	$55.2507 \pm .2144$	$+.0013 \pm .0603$
B/16	16	$54.6196 \pm .1346$	$55.7757 \pm .1940$	$55.1913 \pm .0543$	$+.0208 \pm .0983$
B/32	4	$50.8593 \pm .0446$	$51.6263 \pm .3033$	$51.2396 \pm .1355$	$+.0206 \pm .0033$
B/32	16	$50.8456 \pm .0534$	$51.7693 \pm .0706$	$51.3033 \pm .0494$	$+.0094 \pm .0226$

Conventional ZSL evaluation restricts candidates to the 50 unseen classes. In table order, NORMFOLD obtains 70.1218 ± 0.1457, 69.9965 ± 0.0364, 68.9882 ± 0.0519, and 68.9537 ± 0.1188 percent. Its mean H improvements over the initial model are 0.3493, 0.2899, 0.4636, and 0.5273 points. The much smaller differences from (C, q) do not establish a practical accuracy advantage.

The 12 NORMFOLD A/B pairs have maximum $E = 1.10 \times 10^{-7}$, maximum $D = 1.11 \times 10^{-5}$, and zero differing decisions out of 56,772 repeated query decisions. Raw-SVA, ScaleFold-SVA and (C, q)-SVA have 597, 623 and 405 differences, respectively; their maximum E is 0.01069, 0.00965 and 0.00448, with maximum $D = 0.3420$, 0.3367 and 0.2070. These are descriptive matched-representative counts, not independent image trials or an accuracy-based ranking.

Table 16: **Complete learned numerical diagnostic.** Each row includes all three four-shot support seeds. Accuracy is natural-branch A mean ± sample SD (percentage points). Blocks are A/B counts, constant across the three seeds. The paired error E and update-normalized discrepancy D are defined in Equation 11. The E column shows $(\mathrm{maximum}/\mathrm{mean}) \times 10^k$ over seeds; D is a seed mean. Disagreements sum counts over three evaluations of the same 10,000-image test set, giving a denominator of 30,000 per row, not 30,000 independent training trials. Both preparation precisions use float32 training.

Method	Prep.	τ	Blocks A/B	Accuracy A	E max / mean	Disagr.	Mean D
(C, q)	64	10^{-6}	768/768	68.82 ± 0.08	$(2.96/2.69)\times10^{-3}$	73	8.88×10^{-2}
(C, q)	64	10^{-4}	686/634	68.79 ± 0.06	$(2.27/2.06)\times10^{-3}$	85	6.96×10^{-2}
(C, q)	64	10^{-3}	64/47	68.55 ± 0.01	$(1.67/1.48)\times10^{-3}$	92	5.02×10^{-2}
(C, q)	32	10^{-6}	768/768	68.82 ± 0.08	$(2.96/2.69)\times10^{-3}$	73	8.88×10^{-2}
(C, q)	32	10^{-4}	686/634	68.79 ± 0.06	$(2.27/2.06)\times10^{-3}$	85	6.96×10^{-2}
(C, q)	32	10^{-3}	64/47	68.55 ± 0.01	$(1.67/1.48)\times10^{-3}$	92	5.02×10^{-2}
NormFold	64	10^{-6}	767/767	68.86 ± 0.10	$(1.00/0.998)\times10^{-7}$	0	3.39×10^{-6}
NormFold	64	10^{-4}	687/687	68.81 ± 0.07	$(1.01/1.00)\times10^{-7}$	0	3.41×10^{-6}
NormFold	64	10^{-3}	66/66	68.57 ± 0.02	$(1.01/1.00)\times10^{-7}$	0	3.46×10^{-6}
NormFold	32	10^{-6}	768/768	68.86 ± 0.10	$(7.95/7.37)\times10^{-7}$	0	2.51×10^{-5}
NormFold	32	10^{-4}	688/688	68.81 ± 0.07	$(4.90/4.61)\times10^{-7}$	0	1.56×10^{-5}
NormFold	32	10^{-3}	67/67	68.57 ± 0.02	$(1.59/1.54)\times10^{-7}$	0	5.29×10^{-6}

E Preparation Precision, Rank Selection, and Spectral Grouping

E.1 Matched precision and grouping experiment

This diagnostic separates preparation precision from model execution and learning. On CLIP ViT-B/16 and CIFAR-100, we use four-shot support seeds 0, 1, and 2, both A/B representations, (C, q)-SVA and NormFold, CPU preparation in float64 or float32, and $\tau \in \{10^{-6}, 10^{-4}, 10^{-3}\}$. Scale folding, centering, and SVD use the declared preparation dtype; model execution, fixed factors, and Adam updates remain float32. The active cutoff is $s_i > 10^{-7}s_1$. Each of the 72 fits independently prepares factors and trains a fresh optimizer for 200 steps with the formal objective, batches, prompt, and final-checkpoint rule, using the same three support draws as the primary experiment.

All 72 fits pass the zero/nonzero full-image versus cached-tail and explicit-formula checks and the nonzero first-learning-step check. The largest full/cache gradient relative error is 4.69×10^{-6}, below the fixed 10^{-5} implementation threshold. The twelve float64/$\tau = 10^{-6}$ fits exactly repeat corresponding formal conditions: independently trained final logits are bitwise identical to their formal counterparts over all 10,000 test images. This is a reproduction calibration, not twelve additional statistical replicates. Table 16 includes every method, precision, threshold, and support seed.

Preparation precision changes numerical rank selection. At $\tau = 10^{-6}$, float64 preparation gives NormFold 767 active directions, whereas float32 gives 768. The smallest-to-largest singular-value ratios in float32 are 1.21×10^{-7} and 1.34×10^{-7} in A and B, respectively, just above the unchanged 10^{-7} active cutoff; the corresponding float64 ratios are 1.78×10^{-16} and 2.41×10^{-16}. Thus finite-precision residuals can activate a direction that is null in the exact centered operator. We retain that outcome rather than changing the cutoff after observing it. Float32 preparation raises NormFold's mean A/B logit error from 9.98×10^{-8} to 7.37×10^{-7} at the main threshold, but gives no A/B prediction disagreements in any tested condition. Matched float32/float64 predictions also agree for both methods on every tested branch, threshold, and seed. These results show that the chosen float32 training remains stable across the tested preparation variants, while numerical rank selection remains part of the practical algorithm specification.

Grouping changes capacity as well as agreement. Coarsening τ from 10^{-6} to 10^{-3} reduces NormFold's float64 block count from 767 to 66 and lowers mean A accuracy from 68.86% to 68.57%, despite zero A/B prediction disagreements at both settings. For (C, q)-SVA, the independently chosen block counts differ between A and B at coarser thresholds (686/634 and 64/47). Its

mean E decreases with coarsening, while total prediction disagreements increase from 73 to 92 per 30,000 paired decisions. The experiment therefore separates two roles of grouping: exact repeated-block tying provides basis independence, whereas grouping distinct values controls adaptation capacity.

This experiment learns every declared active block on actual internal image states using the recorded adjacent-gap rule. It characterizes the chosen implementation over three support draws at one model and dataset; broader rewrite strengths, spectra, interfaces, and numerical environments remain open dimensions for evaluation.

F Standard Corruptions and Finite-Precision Decision Boundaries

F.1 Complete standard-corruption grid

We evaluate the original CIFAR-100-C benchmark following Hendrycks & Dietterich (2019): all 15 core corruption types at five severities, using the same 10,000 CIFAR-100 test identities. Both CLIP models retain the 24 clean-trained four-shot adapters from the primary study: three support seeds, two methods, and A/B. Evaluation uses no retraining, test-time adaptation, condition selection, or corruption-dependent tuning; natural branch A is fixed for accuracy reporting.

Before inference, the 15 image-array files and label file pass published per-file MD5 checks, full-file and severity-slice SHA checks, array-shape checks, and clean-label ordering checks. Files obtained from Zenodo and the authors' public mirror agree on independently retrieved labels and Gaussian-noise arrays; all evaluations use the original benchmark arrays.

All 152 model conditions complete: two models times 75 corruptions plus one clean reference. A raw-output audit recomputes all 912 learned A/B pairs and 152 zero-update pairs. Twenty-eight full-image versus cached-tail gates pass, and all 24 clean adapter replays reproduce their saved query logits bitwise. Prefix reuse is limited to identical upstream parameters, while A/B spectral preparations remain independent.

Table 17: **All 75 standard conditions, equally weighted.** A-branch top-1 accuracy (%), not AlexNet-normalized mCE. Each seed first averages the complete grid; mean $\pm$ sample SD then uses three support seeds. The last column is paired NormFold minus (C, q) in percentage points.

Model	Initial	(C, q)-SVA	NormFold	Difference
B/16	48.3895	$48.8271 \pm .0208$	$48.8343 \pm .0224$	$+.0072 \pm .0027$
B/32	43.7473	$44.3810 \pm .0140$	$44.3812 \pm .0153$	$+.0003 \pm .0016$

NormFold exceeds its zero-update reference by 0.4448 and 0.6339 points, respectively. Its much smaller differences from (C, q) do not establish a practically meaningful accuracy advantage. The aggregate retains all 75 corruption-severity conditions without condition selection. Seventy-five conditions are not 75 independent support seeds.

Table 18: **Standard-corruption representative consistency.** Arithmetic means of condition-level E, D, not ratios after concatenating conditions. Disagreement denominators include repeated evaluations; zero-update rows have no learned update, so D is undefined. Clean conditions are excluded from this table.

Model	Method	Mean E	Mean D	Different decisions
B/16	Initial	9.89×10^{-8}	n/a	$0/750{,}000$
	(C, q)-SVA	2.54×10^{-3}	0.1101	$11{,}202/2{,}250{,}000$
	NormFold	9.66×10^{-8}	4.33×10^{-6}	$1/2{,}250{,}000$
B/32	Initial	1.09×10^{-7}	n/a	$1/750{,}000$
	(C, q)-SVA	1.40×10^{-3}	0.0880	$13{,}212/2{,}250{,}000$
	NormFold	1.08×10^{-7}	6.99×10^{-6}	$2/2{,}250{,}000$

Across both models, NORMFOLD differs in three of 4,500,000 paired corrupted decisions, versus 24,414 for (C, q). Its largest condition-level E is below 1.16×10^{-7}. Even the B/32 zero-update pair has one disagreement, illustrating a floating-point boundary present before adaptation; no method is declared exactly argmax-invariant.

The same *a posteriori* sufficient condition $m_i > 2\eta_i$ in Equation 12 holds for 4,499,949 NORMFOLD decisions; 51 are uncertified, including all three flips. For (C, q), 3,949,107 satisfy it and 550,893 do not; all 24,414 flips lie in the latter set. These counts describe repeated decisions, not independent trials for a significance test. We preserve every flip and all 100 original A/B logits for each NORMFOLD counterexample. All three NORMFOLD flips are between two incorrect labels, so unchanged accuracy alone would have hidden these prediction changes.

Table 19: **All three standard-corruption NORMFOLD flips.** Indices, support seeds and class IDs are zero-based. m is the A top-two margin and η the largest absolute A/B logit difference.

Model	Corruption	Severity	Seed	Index	True; A/B	m	η
B/32	Impulse noise	5	0	4529	50; 40/20	1.91×10^{-6}	3.82×10^{-6}
B/32	Motion blur	2	0	5089	80; 75/74	3.82×10^{-6}	7.63×10^{-6}
B/16	Defocus blur	4	2	8445	19; 20/40	1.91×10^{-6}	7.63×10^{-6}