TRANSFORM COORDINATES SEPARATE VULNERABILITY FROM COMPENSATORY CAPACITY UNDER DEPLOYMENT PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Local sensitivity identifies functional directions vulnerable to deployment perturbations without necessarily revealing which can compensate for the induced error. A three-parameter Iwasawa representation of the linear canonical transform (LCT), together with a real-valued quadrature of its complex kernel, yields a token-mixing matrix whose derivatives with respect to these parameters can be obtained analytically. Analog in-memory computing (AIMC) supplies a controlled post-training perturbation, followed by hardware-aware fine-tuning to measure adaptation. These differentials capture pronounced anisotropy across coordinates, with additional variation remaining unexplained. After mapping, fractional order retains the highest sensitivity yet exhibits the smallest displacement during adaptation, while the relative contributions of scale and shear change with depth. A local response estimate produces candidate-subset rankings that align more closely with restricted recovery than sensitivity-based selection, and comparison with unrestricted fine-tuning shows that the LCT parameters account for only part of the available recovery. Transform coordinates thereby separate vulnerability from compensatory capacity and provide an interpretable basis for prioritizing post-deployment adaptation.

1 INTRODUCTION

Sensitivity can identify directions most affected by a deployment perturbation without revealing which directions can absorb the resulting error. A highly sensitive degree of freedom may move little during fine-tuning, while another with weaker local sensitivity carries more of the correction. Final task performance can conflate vulnerability, perturbation-induced forcing, realized parameter movement, and recoverability under restricted adaptation (Hassibi & Stork, 1992; Dinh et al., 2017; Koh & Liang, 2017; Bae et al., 2022). Distinguishing these roles requires a representation in which changes correspond to identifiable transformations rather than isolated weights.

Weight-space analyses provide local importance and response measures, but individual matrix entries rarely identify which transformation of the computation has changed. One displacement can combine variations in gain, scale, phase, and spatial coupling. Low-dimensional transform parameterizations provide a complementary interface in which sensitivity, local response, and subsequent adaptation can be compared along functional directions without treating vulnerability and compensatory capacity as equivalent (Dao et al., 2019; Şahinuç & Koç, 2022; Koç et al., 2024).

Analog in-memory computing (AIMC) performs matrix-vector products where weights are stored, reducing repeated weight movement and making dense linear operations a natural target for in-memory execution (Gokmen & Vlasov, 2016; Ielmini & Wong, 2018; Sebastian et al., 2020). Mapping a pretrained model to AIMC introduces finite DAC/ADC resolution and hardware nonidealities into those operations, while hardware-aware fine-tuning can recover part of the resulting performance loss (Rasch et al., 2023; Lammie et al., 2024). Dense token mixing generated from a small parameter set connects hardware mapping to analysis in identifiable transformation coordinates. The linear canonical transform (LCT) supplies such a parameterization through an Iwasawa form that separates fractional-order rotation, scale, and shear and admits closed-form derivatives of the selected quadrature matrix (Ozaktas et al., 2001; Alieva & Bastiaans, 2005).

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.

The central contribution is a coordinated analysis that keeps sensitivity, perturbation-induced forcing, realized movement, and restricted recoverability in a common LCT parameterization, allowing vulnerability and compensatory capacity to be compared without conflating them. Analytic differentials of the selected LCT quadrature then separate transform-level anisotropy from the depth dependence imposed by the surrounding trained model. A regularized local response estimate turns that distinction into a restricted adaptation criterion, identifying parameter sets that recover more loss than sensitivity-based selection in most models while parameter-group comparisons show that the LCT subspace contains a minority of recoverable capacity in most models.

2 Related Work

Adaptation after deployment perturbations. Post-training quantization and hardware-aware optimization seek to preserve or recover task performance when deployment changes the implemented computation. AdaRound optimizes rounding decisions through local reconstruction (Nagel et al., 2020), OPTQ uses approximate second-order information for sequential weight quantization (Frantar et al., 2023), and RAOQ adapts quantization to in-memory output constraints (Zhang et al., 2024). AIMC introduces additional mismatch through programming error, read noise, drift, array effects, and finite DAC/ADC resolution (Gokmen & Vlasov, 2016; Ielmini & Wong, 2018; Sebastian et al., 2020). Hardware demonstrations span resistive random-access memory (RRAM) and phase-change memory (PCM) arrays through large neural-network inference (Prezioso et al., 2015; Ambrogio et al., 2018; Chen et al., 2019; Joshi et al., 2020; Yao et al., 2020; Wan et al., 2022; Le Gallo et al., 2023; Chen et al., 2025), while AIHWKit and related hardware-aware training provide software models for mapped computation with hardware nonidealities (Rasch et al., 2021; 2023). Performance recovery alone does not determine whether the directions most affected by deployment are also those used for compensation.

Structured token mixing and transform coordinates. Mixer architectures replace or complement attention with linear, multilayer-perceptron, Fourier-domain, or other structured token interactions (Dao et al., 2019; Liu et al., 2021; Rao et al., 2021; Tolstikhin et al., 2021; Dao et al., 2022; Guibas et al., 2022; Lee-Thorp et al., 2022; Touvron et al., 2023; Wang et al., 2022). Among these structured approaches, fractional-transform models make transform order trainable (Şahinuç & Koç, 2022; Koç et al., 2024). Such parameterizations are usually introduced to define or learn the mixing operation itself. Iwasawa coordinates instead provide three interpretable directions within one LCT family for post-mapping sensitivity and restricted adaptation, without assuming that the LCT parameterization contains all available compensation.

Sensitivity and local response geometry. Classical second-order pruning already distinguishes the cost of removing a weight from compensatory adjustment of the remaining parameters (Hassibi & Stork, 1992). Natural-gradient, curvature, influence, and implicit-differentiation methods provide local geometries for parameter response (Amari, 1998; Schraudolph, 2002; Martens & Grosse, 2015; Koh & Liang, 2017; Lorraine et al., 2020; Martens, 2020). Parameter-space sensitivity can vary under function-preserving reparameterization, empirical-Fisher outer products need not equal the model Fisher or loss Hessian, and local influence predictions can deteriorate when retraining leaves the neighborhood in which they were measured (Dinh et al., 2017; Kunstner et al., 2019; Bae et al., 2022). Selective fine-tuning can release subsets chosen by parameter-importance or Fisher-style criteria (Xu et al., 2021), while parameter importance itself can change during adaptation (Lin et al., 2026). Existing approaches do not jointly separate vulnerability, perturbation-induced forcing, realized movement, and restricted recoverability in identifiable transformation coordinates.

3 Linear Canonical Transform Token Mixing and Hardware Mapping

3.1 Model architecture and analysis interface

Each block applies token mixing generated from one real-valued quadrature of the complex linear canonical transform (LCT) kernel, followed by a two-layer channel multilayer perceptron (MLP). Both stages use independent affine preprocessing and LayerScale within residual paths (Touvron et al., 2021; 2023). Only token mixing carries the three LCT parameters per block. The same trained

model is tracked from its FP32/FP16 mixed-precision checkpoint through hardware mapping and hardware-aware fine-tuning (Fig. 1).

3.2 LCT QUADRATURE AND IWASAWA PARAMETERIZATION

The LCT is specified by a real matrix

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2, \mathbb{R}), \qquad ad - bc = 1,$$

where a, b, c, d are the transform coefficients. Under the Iwasawa parameterization used here, normalized fractional order θ defines the physical rotation angle $\phi = \frac{\pi}{2}\theta$, $s > 0$ is the scale, and k is the shear. The four LCT coefficients are then

$$a = s\cos\phi, \quad b = s\sin\phi, \quad c = ks\cos\phi - \frac{\sin\phi}{s}, \quad d = ks\sin\phi + \frac{\cos\phi}{s}. \tag{1}$$

Fractional order controls the rotation-like component, scale controls dilation, and shear enters the quadratic phase.

Both the continuous LCT kernel and its finite-grid sampling are generally complex valued (Ozaktas et al., 2001; Alieva & Bastiaans, 2005). The model retains one real-valued quadrature by selecting either the real-part or imaginary-part coefficients, producing a real mixing matrix and preserving real-valued multiply-accumulate arithmetic after hardware mapping. Let H denote the number of tokens along one spatial axis. For integer indices $n, m \in \{0, \ldots, H-1\}$, define centered normalized coordinates $x_n = (n - (H-1)/2)/H$ and $x_m = (m - (H-1)/2)/H$. For $b > 0$, define

$$q(x_n, x_m) = \cot\phi\, x_n^2 - \frac{2x_n x_m}{s\sin\phi} + \left(k + \frac{\cot\phi}{s^2}\right) x_m^2, \tag{2}$$

where q is the quadratic phase polynomial. The sampled one-dimensional matrix is

$$K_\psi[n, m] = \frac{1}{H\sqrt{s\sin\phi}} \psi\left(\pi q(x_n, x_m) - \frac{\pi}{4}\right), \qquad \psi \in \{\cos, \sin\}. \tag{3}$$

Here $K_\psi \in \mathbb{R}^{H\times H}$ is the selected real-valued quadrature matrix, with $\psi = \cos$ taking the real part and $\psi = \sin$ taking the imaginary part under the adopted phase convention. Separable two-dimensional token mixing uses $K_\psi \otimes K_\psi$, where $\otimes$ is the Kronecker product. Analysis uses $p = (\theta, \ell, \kappa)$, with $\ell = \log s$ and $\kappa = \operatorname{asinh} k$. Parameter ranges and boundary derivatives are specified in Appendix B.

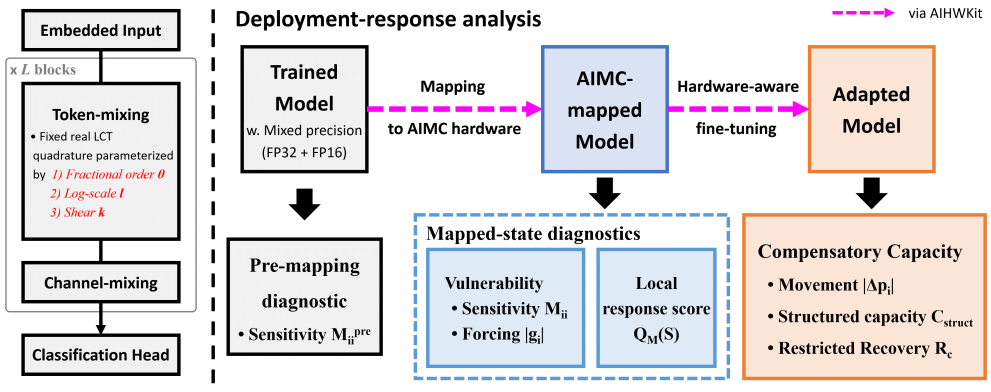

Figure 1: Model architecture and deployment-response analysis. The same L-block LCT mixer is tracked from the trained state through AIMC mapping and hardware-aware fine-tuning using AIHWKit. Pre-mapping sensitivity provides a reference diagnostic. At the mapped state, vulnerability is characterized by sensitivity $\mathcal{M}_{ii}$ and forcing $|g_i|$, while $Q_{\mathcal{M}}(S)$ scores candidate local interventions. After adaptation, compensatory capacity is characterized by movement $|\Delta p_i|$, structured-subspace capacity C_{struct}, and restricted recovery R_c.

3.3 HARDWARE MAPPING AND FINE-TUNING

After FP32/FP16 mixed-precision training, eligible linear operations are mapped with AIHWKit under a fixed RPU configuration (Rasch et al., 2021; 2023). Hardware-aware fine-tuning keeps the mapped forward computation active while updating only designated parameters. Mapped weights use the hardware-aware backward path, whereas the digital LCT parameters regenerate the selected quadrature matrix after each update and use a straight-through derivative for writing it to the mapped module. RPU configuration, conversion, and optimization details are given in Appendix C.

4 SENSITIVITY, RESPONSE, AND RESTRICTED RECOVERY

4.1 SENSITIVITY, FORCING, MOVEMENT, AND RECOVERABILITY

Let $p_0 \in \mathbb{R}^d$ denote all transformed LCT coordinates immediately after hardware mapping, with d the total number of coordinates across blocks, and let p_{FT} denote the coordinates after hardware-aware fine-tuning. For sample t from the calibration set, $z_t \in \mathbb{R}^d$ is the backward field propagated to these coordinates through the straight-through hardware-aware computation using the observed target label. It is not assumed to be the exact gradient of a scalar potential. With $\mathbb{E}_t[\cdot]$ denoting the empirical average over calibration samples, define

$$g = \mathbb{E}_t[z_t], \qquad \mathcal{M} = \mathbb{E}_t[z_t z_t^\top] = gg^\top + \mathrm{Cov}(z), \tag{4}$$

where $\mathrm{Cov}(z) = \mathbb{E}_t[(z_t - g)(z_t - g)^\top]$. For coordinate i, sensitivity is $\mathcal{M}_{ii}$, forcing is $|g_i|$, movement is $|\Delta p_i|$ with $\Delta p = p_{\mathrm{FT}} - p_0$, and recoverability is the loss reduction obtained when adaptation is restricted to a selected subset. The second moment has empirical-Fisher form but is neither the model Fisher nor a generalized Gauss–Newton matrix (Schraudolph, 2002; Kunstner et al., 2019). The straight-through field is generally nonconservative, so its Jacobian is not interpreted as a Hessian (Appendix G.3).

For a subset $S \subseteq \{1, \dots, d\}$ of trainable LCT coordinates, let $\delta \in \mathbb{R}^d$ denote a local coordinate displacement. The notation $\mathrm{supp}(\delta)$ denotes the indices of nonzero components, so $\mathrm{supp}(\delta) \subseteq S$ permits movement only within S. The regularized local response is

$$\min_{\mathrm{supp}(\delta) \subseteq S} g^\top \delta + \tfrac{1}{2}\delta^\top (\mathcal{M} + \mu I)\delta, \qquad \delta_{\mathcal{M},S} = -(\mathcal{M}_{SS} + \mu I)^{-1} g_S, \tag{5}$$

where g_S is the subvector of g indexed by S, $\mathcal{M}_{SS}$ is the corresponding principal submatrix, I is the identity matrix on that subspace, and $\mu > 0$ is a regularization coefficient fixed for each trained model. The resulting response score is

$$Q_{\mathcal{M}}(S) = \tfrac{1}{2} g_S^\top (\mathcal{M}_{SS} + \mu I)^{-1} g_S. \tag{6}$$

The score ranks restricted interventions near p_0 rather than predicting recovered loss or the optimizer trajectory. Its absolute magnitude depends on the coordinate parameterization, while subset ordering remains stable across the alternatives examined in Appendix B.3–B.4.

4.2 QUADRATURE-LEVEL DIFFERENTIAL PREDICTIONS

The selected real-valued LCT quadrature provides a transform-level reference for sensitivity anisotropy independent of the learned model. Write $K_\psi = A\psi(\Phi)$, where $A = (H\sqrt{s \sin\phi})^{-1}$ is the amplitude factor and $\Phi = \pi q - \pi/4$ is the phase. For $u \in \{\phi, s, k\}$,

$$\frac{\partial K_\psi}{\partial u} = \frac{\partial A}{\partial u}\psi(\Phi) + A\psi'(\Phi)\pi\frac{\partial q}{\partial u}. \tag{7}$$

Here u is one physical LCT parameter and ψ' is the derivative of the selected quadrature. Shear changes only the phase through $\partial q/\partial k = x_m^2$, whereas fractional order changes both amplitude and phase through ϕ. The corresponding phase derivative is

$$\frac{\partial q}{\partial \phi} = -\csc^2\phi\Big(x_n^2 + \frac{x_m^2}{s^2}\Big) + \frac{2x_n x_m \cos\phi}{s\sin^2\phi}. \tag{8}$$

Define $D_u = \|\partial K_\psi/\partial u\|_F$, where $\|\cdot\|_F$ is the Frobenius norm. The ratio D_ϕ/D_k follows a $\csc^2\phi$ envelope. In $p = (\theta, \ell, \kappa)$, the chain rule gives $D_\theta = (\pi/2)D_\phi$, $D_\ell = sD_s$, and $D_\kappa = \sqrt{1+k^2}D_k$.

These derivatives predict sensitivity anisotropy only when upstream backward-field factors are comparable. The exact separable two-dimensional derivative is given in Appendix B.5.

Quadrature choice also gives a conditional prediction for the direction of shear adaptation. Let K_{real} denote Eq. (3) with $\psi = \cos$ and K_{imag} denote Eq. (3) with $\psi = \sin$. Because A is independent of k,

$$\frac{\partial \|K_{\mathrm{imag}}\|_F^2}{\partial k} = A^2 \sum_{n,m} \pi x_m^2 \sin 2\Phi = -\frac{\partial \|K_{\mathrm{real}}\|_F^2}{\partial k}. \tag{9}$$

The two quadratures predict opposite preferred shear directions when local training favors increasing the mixing-matrix energy.

4.3 Intervention and recovery measures

Equation (6) is evaluated in diagonal and full-matrix forms to separate coordinate-wise response scale from coupling. Q_{diag} uses only the diagonal of $\mathcal{M}_{SS}$, while Q_{full} retains the complete principal submatrix. Greedy subset search is used, with beam and hill-climbing checks reported in Appendix G.5. Pre-mapping sensitivity, post-mapping sensitivity, both response scores, and stratified random selection are compared at matched subset size.

Two normalized quantities separate the amount of recovery available through the LCT coordinates from the quality of selecting a smaller coordinate set,

$$C_{\mathrm{struct}} = \frac{\mathcal{L}_{\mathrm{conv}} - \mathcal{L}_{\mathrm{struct}}}{\mathcal{L}_{\mathrm{conv}} - \mathcal{L}_{\mathrm{full}}}, \qquad R_c = \frac{\mathcal{L}_{\mathrm{conv}} - \mathcal{L}_c}{\mathcal{L}_{\mathrm{conv}} - \mathcal{L}_{\mathrm{struct}}}. \tag{10}$$

Here $\mathcal{L}_{\mathrm{conv}}$ is the loss immediately after hardware mapping, $\mathcal{L}_{\mathrm{struct}}$ is the loss after hardware-aware fine-tuning with all LCT coordinates trainable, $\mathcal{L}_{\mathrm{full}}$ is the loss after full-model hardware-aware fine-tuning, and $\mathcal{L}_c$ is the loss after fine-tuning the coordinate set chosen by criterion c. C_{struct} measures the fraction of full-model recovery available within the LCT coordinates, while R_c measures how much of that recovery is reached by criterion c.

The CIFAR analysis spans both quadratures. Sensitivity and response are measured immediately after mapping, before hardware-aware fine-tuning, and recovery is evaluated on held-out data. Tiny ImageNet is used as a separate validation.

5 Results and Discussion

5.1 Quadrature geometry

The differential lens in Eqs. (7)–(8) separates transform-level geometry from the effects of the surrounding trained computation. Derivatives of the selected LCT quadrature reproduce the dominant sensitivity ordering between transformation directions, while the exact separable two-dimensional correction adds little beyond the finite-grid one-dimensional calculation. The leading anisotropy is thus already present in the quadrature derivatives rather than being created by hardware mapping (Fig. 2a). By contrast, neither the exact separable derivative nor a data-conditioned derivative removes the block-to-block residual. The same residual is visible before mapping, placing the depth dependence primarily in the trained computation surrounding the token-mixing transform rather than in the selected quadrature or the deployment perturbation.

Equation (7) captures only the transform-side differential. The measured backward field also contains activation and loss-gradient factors supplied by the trained model, so the derivative norm is not expected to reconstruct the full sensitivity map. Agreement in the between-direction ordering identifies a geometric contribution from the selected quadrature, whereas the persistent depth residual identifies where model conditioning enters. Measured sensitivity is thus consistent with quadrature-governed anisotropy modulated by the surrounding computation.

Functional substitution tests the derivative-based interpretation without using backward fields. Replacing fractional order alone reproduces most of the degradation caused by replacing all three LCT parameters, and matching the substituted quadrature-matrix norm does not remove the effect. The dominant fractional-order sensitivity is not reducible to a trivial change in overall mixing magnitude.

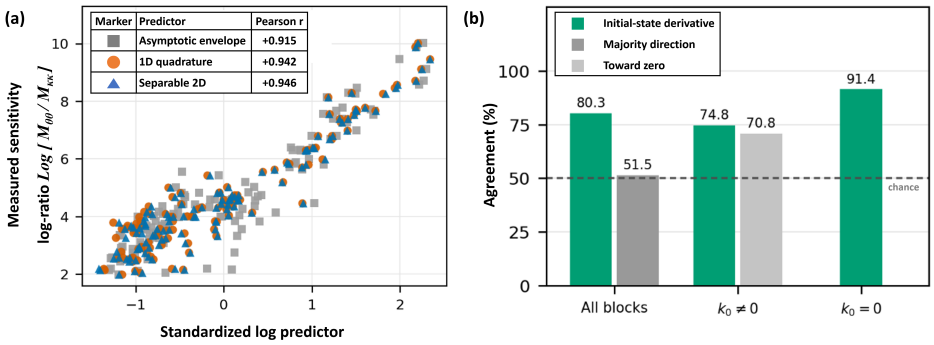

Figure 2: Quadrature-level predictions of sensitivity anisotropy and shear-direction change. (a) Transform-side predictors are compared with the mapped-state sensitivity ratio $\log(\mathcal{M}_{\theta\theta}/\mathcal{M}_{\kappa\kappa})$ over 121 clamp-excluded CIFAR blocks. The three predictors are $\log(\csc^2\phi)$, the one-dimensional quadrature ratio $\log(D_\theta^2/D_\kappa^2)$, and its exact separable two-dimensional counterpart. Each predictor is standardized over the 121 blocks, while the measured sensitivity ratio is not standardized. The inset reports Pearson r. (b) Initial-state shear-direction prediction over CIFAR-10/100 blocks using the sign of $\partial\|K\|_F^2/\partial k$ at (θ_0, s_0, k_0). Agreement is 80.3% over all blocks ($n = 1296$), 74.8% for $k_0 \neq 0$ ($n = 864$), and 91.4% for $k_0 = 0$ ($n = 432$). For nonzero initial shear, movement toward zero gives 70.8% agreement.

The corresponding finite-grid relation also persists on Tiny ImageNet, supporting the quadrature-level explanation without requiring recalibration to the CIFAR token grid.

Equation (9) gives a more specific directional prediction for shear. The observed quadrature dependence is consistent with the predicted sign, but its advantage over a simple movement-toward-zero heuristic is modest when initial shear is nonzero. The directional result is consequently weaker than the sensitivity-anisotropy result and is used as supporting evidence rather than as a separate claim (Fig. 2b).

5.2 Sensitivity and Adaptation

Sensitivity, forcing, movement, and recoverability locate different stages of the post-mapping response. Fractional order dominates the sensitivity $\mathcal{M}_{ii}$ but contributes the least realized movement, and the ordering has already changed at the level of the mean forcing $|g_i|$ before the first fine-tuning update. Equation (4) clarifies why these quantities need not agree. A large diagonal second moment can reflect either a coherent mean field or sample-dependent variation, while only the former supplies a consistent immediate direction for adaptation. Fractional order can remain highly vulnerable without receiving the strongest coherent drive after mapping. Realized movement adds a further constraint because optimization can redistribute correction across other LCT directions and parameters outside the LCT subspace (Fig. 3a).

Depth introduces a second separation between forcing and movement. Total LCT-coordinate displacement decreases with block depth even though displacement normalized by forcing increases, because the forcing magnitude itself falls more strongly. The relative allocation of movement also changes from greater shear contribution near the input side toward greater scale contribution at larger depth, with repeated changes in ordering rather than a single transition. A global coordinate ranking cannot represent this depth-conditioned redistribution (Fig. 3).

The local response in Eq. (5) captures part of the realized displacement but not the full allocation across coordinates. Its weaker account of the scale contribution and the failure of later re-estimation to restore alignment limit its interpretation as a trajectory model. Updating the local geometry without recovering the movement pattern suggests that the mismatch is not merely caused by an outdated expansion point. Redistribution through coupled parameter groups and the nonlinear optimization path remain outside the restricted local description. The sensitivity–movement dissociation also persists after excluding coordinates affected by bounds and under alternative coordinate parameterizations, reducing the likelihood that it is created by a particular parameterization choice.

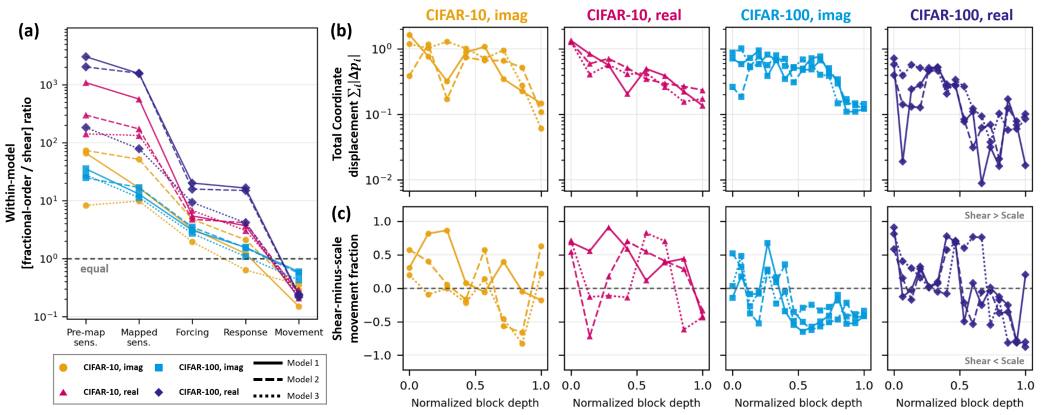

Figure 3: Sensitivity, forcing, and movement across mapping and adaptation. (a) Within-model fractional-order-to-shear ratios for pre-mapping sensitivity, mapped-state sensitivity, forcing, regularized local response, and realized movement. Sensitivity ratios use diagonal entries of the per-example observed-label field second moment $\mathcal{M}$, evaluated before and after mapping, respectively. (b) Total LCT-coordinate displacement $\sum_i |\Delta p_i|$ across normalized block depth. (c) Shear-minus-scale movement fraction across normalized block depth, with positive values indicating greater shear contribution and negative values greater scale contribution. Block-level values are shown without averaging across trained models.

Parameter-group recovery extends the distinction beyond the LCT subspace. Fine-tuning the channel-mixing parameters absorbs most of the loss introduced by hardware mapping, whereas adaptation restricted to the LCT parameters reaches a lower recovery ceiling. The token-mixing transform is still functionally important because substituting its selected quadrature matrix from another independently trained model causes a large performance loss. Functional importance, local vulnerability, and compensatory capacity describe different properties (Fig. 5a). The same qualitative recovery allocation is observed in the two Tiny ImageNet models, with channel-branch recovery above token-branch recovery and the LCT-coordinate subspace providing the lowest recovery ceiling (Fig. 5a; Appendix J).

5.3 Restricted recovery

The response score in Eq. (6) is useful primarily as an ordinal criterion around a mapped state. Its ranking agrees with realized restricted recovery more reliably than its absolute magnitude transfers across trained models. At matched subset size, response-based selection also outperforms pre-mapping sensitivity, post-mapping sensitivity, and random selection in most cases. The gain is consistent with the structure of $Q_{\mathcal{M}}$, which combines the mean forcing with the local response scale instead of using vulnerability alone as a proxy for adaptability (Fig. 4). Composition-matched random controls preserve the selected axis allocation, or both axis and block composition, yet remain below the response criteria in most model–criterion comparisons at the primary subset size (Appendix G.7). The advantage is not explained only by preferential allocation to favorable axes or blocks.

Comparing Q_{diag} and Q_{full} clarifies the role of coordinate coupling. Retaining off-diagonal terms mainly suppresses unusually poor selections rather than uniformly improving typical recovery. Most of the ordering benefit is already present in the diagonal response estimate, while coupling contributes chiefly to robustness when interactions between selected coordinates are unfavorable. The off-diagonal structure is thus useful as a selection correction, not as evidence that the full matrix reproduces the curvature followed by optimization.

The capacity measure C_{struct} and selection measure R_c answer different questions. Total recovery available inside the LCT subspace varies substantially across the CIFAR set, yet the best restricted selection does not follow the same variation. Reliable allocation within a limited subspace can coexist with a low recovery ceiling. Restricted recovery should be interpreted as a problem of choosing useful degrees of freedom inside a declared adaptation space, not as evidence that the chosen space contains most of the available compensation. $Q_{\mathcal{M}}$, C_{struct}, and R_c separate local coordinate ranking,

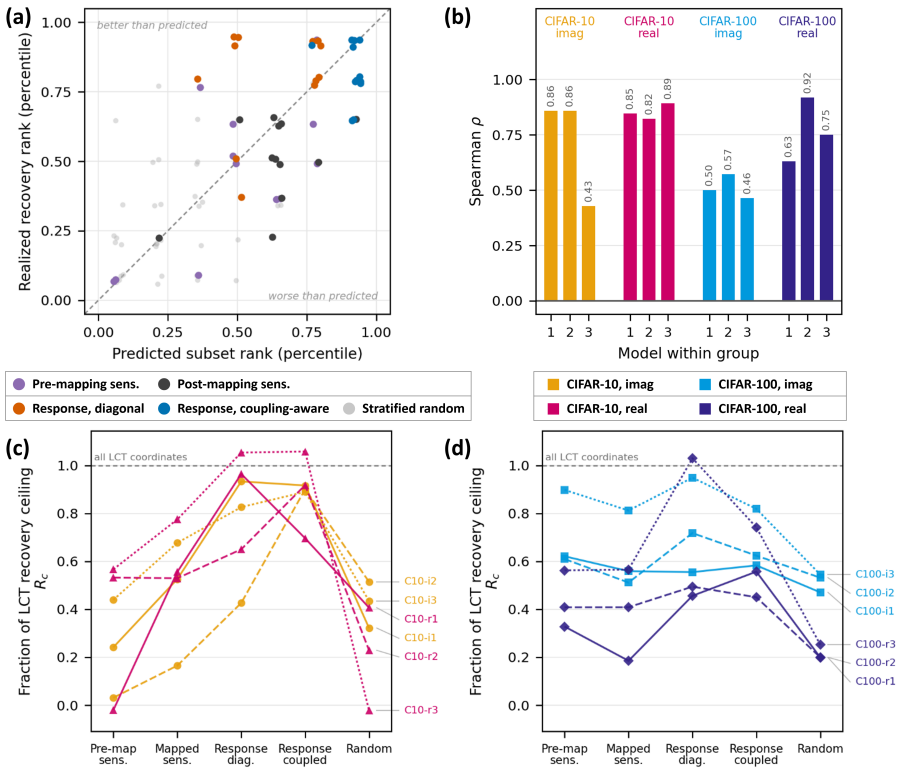

Figure 4: Restricted-recovery ranking and selection. (a) Within-model percentile rank of criterion scores against realized restricted-recovery rank for matched one-third-size LCT-coordinate subsets. Points distinguish pre-mapping sensitivity, mapped-state sensitivity, diagonal and coupling-aware response criteria, and stratified-random controls. The aggregated within-model Spearman association is $\rho = +0.705$, with a model-resampled 95% interval of $[+0.607, +0.798]$. (b) Model-wise Spearman rank association for the same matched-size comparison, grouped by dataset and quadrature. (c,d) Restricted recovery R_c under the five selection criteria for CIFAR-10 and CIFAR-100, expressed as a fraction of the recovery ceiling obtained with all LCT coordinates trainable.

structured-subspace capacity, and use of that capacity. Their different behavior shows that useful local ranking does not require the selected subspace to be globally sufficient for recovery.

The second-moment formulation remains appropriate for that ranking role without being interpreted as optimizer curvature. The straight-through field has a nonsymmetric Jacobian under the examined controls, and symmetrized restricted Jacobians can be indefinite. In contrast, $\mathcal{M}$ is positive semidefinite by construction and $\mathcal{M} + \mu I$ is positive definite for $\mu > 0$, giving a stable regularized ordering criterion even when the underlying field is nonconservative.

Cross-model calibration behaves differently from within-model ranking. The observed-label second moment becomes inflated as mapping loss increases, compressing the predicted recovery scale while preserving useful local ordering. Re-estimating only $\mathcal{M}$ with labels sampled from the mapped model weakens this scale dependence while retaining the observed-label mean forcing g. The label convention affects absolute calibration more strongly than subset ranking, so the selection experiments retain observed targets while the alternative estimate is used only to diagnose calibration error (Fig. 5b).

5.4 Limitations

The LCT coordinates provide a low-dimensional, architecture-specific view of adaptation rather than a complete parameterization of compensatory capacity. Hardware effects are evaluated under a single RPU configuration, so quantitative ordering may change under other configurations without affecting

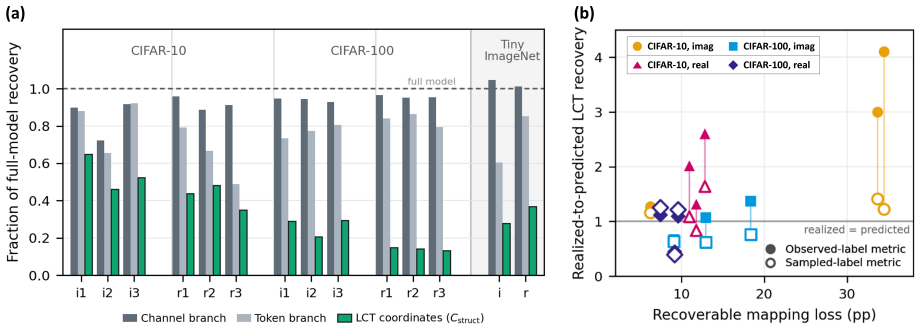

Figure 5: Recovery capacity and cross-model calibration. (a) Recovery obtained by channel-branch, token-branch, and LCT-coordinate adaptation across the twelve CIFAR models, expressed as a fraction of full-model recovery. The LCT-coordinate ratio is C_{struct}, and the dashed line marks full-model recovery. The rightmost group shows the corresponding recovery ratios for the two Tiny ImageNet models. Both reproduce the channel-branch $>$ token-branch $>$ LCT-coordinate ordering and are excluded from the CIFAR model-level summaries. (b) Realized-to-predicted LCT-coordinate recovery versus recoverable mapping loss, with one point per model. Filled markers use the observed-label second moment, whereas open markers re-estimate the second moment with sampled labels while retaining the same observed-label mean forcing g. Colors and marker shapes denote the same dataset–quadrature groups as in Figs. 3 and 4. The horizontal line denotes exact agreement between realized and predicted recovery.

the within-condition comparison between sensitivity and recovery. The response estimate is local to the post-mapping state and supports restricted ranking rather than a global model of fine-tuning. Tiny ImageNet broadens task difficulty and token-grid size but remains descriptive because only one trained model per quadrature is evaluated with a different fine-tuning schedule.

6 Conclusion

Deployment perturbations reveal a mismatch between functional vulnerability and the degrees of freedom that absorb compensation. For LCT-parameterized token mixing, derivatives of the selected real-valued quadrature separate a transform-level source of sensitivity anisotropy from the depth dependence imposed by the surrounding trained model. Fractional order remains highly sensitive after hardware mapping but receives weaker coherent forcing and contributes little realized movement, while the relative participation of scale and shear changes across depth. Sensitivity, forcing, and movement consequently describe different stages of the post-mapping response rather than interchangeable notions of importance.

Restricted adaptation makes this distinction operational. The regularized local response estimate ranks candidate parameter sets more effectively than sensitivity alone, yet its reliable use is local and ordinal rather than an absolute model of recovery or optimization. The amount of recovery available inside the LCT subspace also varies independently of how well useful coordinates can be selected within it. In most evaluated models, the larger share of corrective capacity lies outside the LCT parameters, even though the selected quadrature remains functionally important to the trained computation.

A low-dimensional transform parameterization can expose interpretable perturbation directions, separate geometric sensitivity from model-dependent conditioning, and support targeted adaptation without claiming to span the full recovery space. Similar analyses may be useful whenever deployment changes act on structured transformations whose real-valued implementation admits tractable derivatives.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript language editing and consistency checks, literature retrieval and discovery, and research ideation, including feedback on experimental design and analysis framing. The authors independently reviewed and verified the resulting material and take responsibility for the final manuscript, including its claims, analyses, citations, and presentation.

REFERENCES

Tatiana Alieva and Martin J. Bastiaans. Alternative representation of the linear canonical integral transform. *Optics Letters*, 30(24):3302–3304, 2005. doi: 10.1364/OL.30.003302.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Stefano Ambrogio, Pritish Narayanan, Hsinyu Tsai, Robert M. Shelby, Irem Boybat, Carmelo di Nolfo, Severin Sidler, Massimo Giordano, Martina Bodini, Nathan C. P. Farinha, Benjamin Killeen, Christina Cheng, Yassine Jaoudi, and Geoffrey W. Burr. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature*, 558:60–67, 2018. doi: 10.1038/s41586-018-0180-5.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems*, volume 35, pp. 17953–17967, 2022. doi: 10.52202/068431-1305. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7234e0c36fdbcb23e7bd56b68838999b-Abstract-Conference.html.

An Chen, Stefano Ambrogio, Pritish Narayanan, Atsuya Okazaki, Charles Mackin, Andrea Fasoli, Malte J. Rasch, Alexander Friz, Jose Luquin, Takeo Yasuda, Masatoshi Ishii, Takuto Kanamori, Kohji Hosokawa, Timothy Philicelli, Seiji Munetoh, Vijay Narayanan, Hsinyu Tsai, and Geoffrey W. Burr. Demonstration of transformer-based ALBERT model on a 14nm analog AI inference chip. *Nature Communications*, 16:8661, 2025. doi: 10.1038/s41467-025-63794-4.

Wei-Hao Chen, Chunmeng Dou, Kai-Xiang Li, Wei-Yu Lin, Pin-Yi Li, Jian-Hao Huang, Jing-Hong Wang, Wei-Chen Wei, Cheng-Xin Xue, Yen-Cheng Chiu, Ya-Chin King, Chorng-Jung Lin, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, J. Joshua Yang, Mon-Shu Ho, and Meng-Fan Chang. CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. *Nature Electronics*, 2:420–428, 2019. doi: 10.1038/s41928-019-0288-0.

Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, and Christopher Ré. Learning fast algorithms for linear transforms using butterfly factorizations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1517–1527. PMLR, 2019. URL https://proceedings.mlr.press/v97/dao19a.html.

Tri Dao, Beidi Chen, Nimit S. Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, and Christopher Ré. Monarch: Expressive structured matrices for efficient and accurate training. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4690–4721. PMLR, 2022. URL https://proceedings.mlr.press/v162/dao22a.html.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Tayfun Gokmen and Yurii Vlasov. Acceleration of deep neural network training with resistive cross-point devices: Design considerations. *Frontiers in Neuroscience*, 10:333, 2016. doi: 10.3389/fnins.2016.00333.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EXHG-A3jlM.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pp. 164–171, 1992. URL https://papers.nips.cc/paper_files/paper/1992/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Daniele Ielmini and H.-S. Philip Wong. In-memory computing with resistive switching devices. *Nature Electronics*, 1:333–343, 2018. doi: 10.1038/s41928-018-0092-2.

Vinay Joshi, Manuel Le Gallo, Simon Haefeli, Irem Boybat, S. R. Nandakumar, Christophe Piveteau, Martino Dazzi, Bipin Rajendran, Abu Sebastian, and Evangelos Eleftheriou. Accurate deep neural network inference using computational phase-change memory. *Nature Communications*, 11:2473, 2020. doi: 10.1038/s41467-020-16108-9.

Emirhan Koç, Tuna Alikaşifoğlu, Arda Can Aras, and Aykut Koç. Trainable fractional fourier transform. *IEEE Signal Processing Letters*, 31:751–755, 2024. doi: 10.1109/LSP.2024.3372779.

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 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4158–4169, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/46a558d97954d0692411c861cf78ef79-Abstract.html.

Corey Lammie, Athanasios Vasilopoulos, Julian Büchel, Giacomo Camposampiero, Manuel Le Gallo, Malte J. Rasch, and Abu Sebastian. Improving the accuracy of analog-based in-memory computing accelerators post-training. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2024. doi: 10.1109/ISCAS58744.2024.10558540.

Manuel Le Gallo, Riduan Khaddam-Aljameh, Milos Stanisavljevic, Athanasios Vasilopoulos, Benedikt Kersting, Martino Dazzi, Geethan Karunaratne, Matthias Brändli, Abhairaj Singh, Silvia M. Müller, Julian Büchel, Xavier Timoneda, Vinay Joshi, Malte J. Rasch, Urs Egger, Angelo Garofalo, Anastasios Petropoulos, Theodore Antonakopoulos, Kevin Brew, Samuel Choi, Injo Ok, Timothy Philip, Victor Chan, Claire Silvestre, Ishtiaq Ahsan, Nicole Saulnier, Vijay Narayanan, Pier Andrea Francese, Evangelos Eleftheriou, and Abu Sebastian. A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics*, 6:680–693, 2023. doi: 10.1038/s41928-023-01010-1.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. FNet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.319. URL https://aclanthology.org/2022.naacl-main.319/.

Zekai Lin, Chao Xue, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Lei Jiang, Yu Lu, Bob Simons, Shuang Liang, and Minlong Peng. Parameter importance is not static: Evolving parameter isolation for supervised fine-tuning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32846–32864. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1516. URL https://aclanthology.org/2026.acl-long.1516/.

Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to MLPs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9204–9215, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4cc05b35c2f937c5bd9e7d41d3686fff-Abstract.html.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552. PMLR, 2020. URL https://proceedings.mlr.press/v108/lorraine20a.html.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. doi: 10.1007/s10107-006-0706-8.

Haldun M. Ozaktas, Zeev Zalevsky, and M. Alper Kutay. *The Fractional Fourier Transform: With Applications in Optics and Signal Processing*. Wiley, 2001. ISBN 978-0-471-96346-2.

Mirko Prezioso, Farnood Merrikh-Bayat, Brian D. Hoskins, Geoffrey C. Adam, Konstantin K. Likharev, and Dmitri B. Strukov. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, 521:61–64, 2015. doi: 10.1038/nature14441.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/07e87c2f4fc7f7c96116d8e2a92790f5-Abstract.html.

Malte J. Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. A flexible and fast PyTorch toolkit for simulating training and inference on analog crossbar arrays. In *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 1–4, 2021. doi: 10.1109/AICAS51828.2021.9458494.

Malte J. Rasch, Charles Mackin, Manuel Le Gallo, An Chen, Andrea Fasoli, Frédéric Odermatt, Ning Li, S. R. Nandakumar, Pritish Narayanan, Hsinyu Tsai, Geoffrey W. Burr, Abu Sebastian, and Vijay Narayanan. Hardware-aware training for large-scale and diverse deep learning inference workloads using in-memory computing-based accelerators. *Nature Communications*, 14:5282, 2023. doi: 10.1038/s41467-023-40770-4.

Furkan Şahinuç and Aykut Koç. Fractional fourier transform meets transformer encoder. *IEEE Signal Processing Letters*, 29:2258–2262, 2022. doi: 10.1109/LSP.2022.3217975.

Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002. doi: 10.1162/08997660260028683.

Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. Memory devices and applications for in-memory computing. *Nature Nanotechnology*, 15:529–544, 2020. doi: 10.1038/s41565-020-0655-z.

Trond Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983. doi: 10.1137/0720042.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lučić, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021. doi: 10.1109/ICCV48922.2021.00010.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5314–5321, 2023. doi: 10.1109/TPAMI.2022.3206148.

Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, H.-S. Philip Wong, and Gert Cauwenberghs. A compute-in-memory chip based on resistive random-access memory. *Nature*, 608(7923):504–512, 2022. doi: 10.1038/s41586-022-04992-8.

Ziyu Wang, Wenhao Jiang, Yiming M. Zhu, Li Yuan, Yibing Song, and Wei Liu. DynaMixer: A vision MLP architecture with dynamic mixing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22691–22701. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22i.html.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9514–9528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.749. URL https://aclanthology.org/2021.emnlp-main.749/.

Peng Yao, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J. Joshua Yang, and He Qian. Fully hardware-implemented memristor convolutional neural network. *Nature*, 577 (7792):641–646, 2020. doi: 10.1038/s41586-020-1942-4.

Bonan Zhang, Chia-Yu Chen, and Naveen Verma. Reshape and adapt for output quantization (RAOQ): Quantization-aware training for in-memory computing systems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58739–58762. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24i.html.

APPENDIX

The stored Iwasawa parameters are (θ, s, k), with derived physical angle $\phi = \frac{\pi}{2}\theta$. The complex linear canonical transform (LCT) kernel is reduced to one selected real-valued quadrature matrix for token mixing. Optimization and coordinate-wise analysis use $p = (\theta, \ell, \kappa)$, with $\ell = \log s$ and $\kappa = \operatorname{asinh} k$.

Conventions. Accuracy is reported in percent and accuracy differences in percentage points (pp). Losses, ratios, correlations, and overlaps are unitless. Values written mean±x use SEM. Unless stated otherwise, n denotes independently trained models. Model IDs follow Table A4, with the ordinal ordered by the initialization (θ_0, s_0, k_0) within each dataset–quadrature group. Appendix equations are numbered by section (A.1, B.1, and so on). Appendix tables are numbered A1–A17 and appendix figures A1–A8, independently of the section letters.

Table A1. Symbols and declared coordinate conventions. Quantities are unitless unless a definition supplies its own scale.

symbol	meaning
$\theta \in (0, 1)$	stored normalized fractional order
$\phi = \frac{\pi}{2}\theta$	derived physical rotation angle. Never an independently optimized parameter
$s > 0$, $k \in \mathbb{R}$	stored scale and shear
$p = (\theta, \ell, \kappa)$	coordinate system used for optimization and analysis, $\ell = \log s$, $\kappa = \operatorname{asinh} k$
$j = 1..L$	block index, $3L$ LCT coordinates in total
n, m	integer token-grid indices in $\{0, \ldots, H-1\}$
x_n, x_m	centered normalized token-grid coordinates, $x_n = (n - (H-1)/2)/H$ and $x_m = (m - (H-1)/2)/H$
$K_{\text{full}}[n, m]$	complex linear canonical transform (LCT) kernel
$K_\psi[n, m]$	selected real-valued quadrature matrix obtained from the real or imaginary part of K_{full}
$A(\phi, s)$, $\Phi = \pi q - \pi/4$, q	quadrature amplitude, phase, and phase polynomial
$\mathcal{L}_0$, $\mathcal{L}_\varepsilon$	digital objective and its local deployment-perturbation expansion
$\mathcal{L}_{\text{conv}}$, $\mathcal{L}_{\text{struct}}$, $\mathcal{L}_{\text{full}}$, $\mathcal{L}_c$	held-out losses after conversion, all-LCT adaptation, full-model adaptation, and criterion-c restricted adaptation
$v = \nabla V(p^\star)$	digital-state perturbation gradient in the local deployment expansion
g, $\mathcal{M}$	hardware-conditioned mean field (the forcing), observed-label second moment of the sample-level adaptation field
J_{field}, S_{field}	adaptation-field Jacobian $\partial g/\partial p$ and its symmetrized part. The field has no scalar potential, so the latter is not a Hessian (G.3)
μ	coefficient for identity regularization
$Q_{\mathcal{M}}(S)$	regularized response score of coordinate subset S
C_{struct}, R_c	LCT-subspace recovery capacity and selection quality within that subspace
R_c^{branch}	recovery under adaptation scope c, normalized by full-model recovery

A Architecture and model configurations

The repeated block is

$$\widetilde{X}_j^{\mathrm{tok}} = X_j \odot \gamma_j^{\mathrm{tok}} + \beta_j^{\mathrm{tok}}, \qquad Z_j = K_j \widetilde{X}_j^{\mathrm{tok}} K_j^{\top}. \tag{A.1}$$

$$X'_j = X_j + \mathcal{D}_j^{\mathrm{tok}}\left(\lambda_j^{\mathrm{tok}} \odot Z_j\right), \qquad \widetilde{X}_j^{\mathrm{ch}} = X'_j \odot \gamma_j^{\mathrm{ch}} + \beta_j^{\mathrm{ch}}. \tag{A.2}$$

$$X_{j+1} = X'_j + \mathcal{D}_j^{\mathrm{ch}}\left(\lambda_j^{\mathrm{ch}} \odot \mathrm{MLP}_j(\widetilde{X}_j^{\mathrm{ch}})\right). \tag{A.3}$$

Patch embedding, the classification head, and the final norm complete the model. Token mixing uses $K_j = K_\psi(\theta_j, s_j, k_j) \otimes K_\psi(\theta_j, s_j, k_j)$, where K_ψ is the selected real-valued LCT quadrature matrix. Three Iwasawa parameters determine the mixing matrix in each block, giving 24 LCT parameters at depth 8 and 48 at depth 16.

Table A2. Model configurations. Spatial input and patch sizes are in pixels. All other entries are counts. The Tiny ImageNet row describes the separate larger-grid hardware-aware evaluation and is not part of the primary inferential model set.

setting	classes	input	patch	token grid H	tokens N	depth L	embed	LCT coords
CIFAR-10	10	32	4	8	64	8	256	24
CIFAR-100	100	32	4	8	64	16	256	48
Tiny ImageNet	200	64	4	16	256	16	256	48

The token-branch parameter count is identical for the depth-16 CIFAR-100 and Tiny ImageNet models (12,336), because the affine, bias, and LayerScale vectors depend on embedding dimension rather than token count N. The four-fold increase in N changes the quadrature-matrix size but not the token-branch parameter count, which is relevant to the branch comparison in Appendix H.

Training recipes. CIFAR-100 and Tiny ImageNet share a recipe with depth 16, DropPath 0.4, weight decay 0.05, mixup 0.5, label smoothing 0.1, batch 256, 200 epochs, learning rate 3×10^{-3}, warmup 10, and early stopping 20. CIFAR-10 uses the same training family and also applies mixup 0.5. Because task and digital pretraining configuration differ together, the sensitivity-criterion reversal analyzed in Appendix G is not attributed causally to any single recipe choice.

B Coordinate parameterization, real-valued quadrature, and Jacobians

B.1 Iwasawa parameterization and implementation check.

$$a = s\cos\phi, \quad b = s\sin\phi, \quad c = ks\cos\phi - \tfrac{\sin\phi}{s}, \quad d = ks\sin\phi + \tfrac{\cos\phi}{s}, \qquad \det M = 1\ (< 10^{-12}). \tag{B.1}$$

$$M(\phi, s, k) = \underbrace{\begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix}}_{N\ \text{(shear/chirp)}} \underbrace{\begin{pmatrix} s & 0 \\ 0 & 1/s \end{pmatrix}}_{A\ \text{(dilation)}} \underbrace{\begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}}_{K\ \text{(FrFT order)}}. \tag{B.2}$$

The factorization was verified directly against the implemented float32 quadrature-matrix generator.

Implemented domain. The mixing layer evaluates $\phi = \frac{\pi}{2}\,\mathrm{clamp}(\theta,\, 0.05 + \epsilon,\, 0.95 - \epsilon)$ and $s = \exp \mathrm{clamp}(\ell,\, -\log H,\, \log H)$, so the realized ranges are

$$\theta \in [0.05,\, 0.95], \qquad \phi \in [0.0785,\, 1.492], \qquad s \in [H^{-1},\, H]. \tag{B.3}$$

with $s \in [0.125, 8]$ at $H = 8$ and $[0.0625, 16]$ at $H = 16$. Shear is unclamped, $k = \sinh \kappa$. Both bounds are enforced by the implemented clamp. Its derivative is 1 on the closed interval, **including at the endpoints**, and 0 strictly outside, so the freezing condition is saturation of the *raw* trainable coordinate outside the interval, not equality of the effective coordinate with a bound. The distinction matters because the trajectory records effective values and cannot separate the two cases.

A small subset of fractional-order coordinates is saturated by the implemented clamp. For these coordinates the clamp derivative vanishes, giving $g_i = 0$ and $\mathcal{M}_{ii} = 0$. The log-scale bound is inactive in the analyzed models. Saturation belongs to the parameter-to-quadrature map rather than to a coordinate transformation, and all analyzed states satisfy $b = s \sin \phi > 0$.

B.2 Real-valued quadrature realization and scope. For integer matrix indices $n, m \in \{0, \ldots, H-1\}$, the implementation uses centered normalized coordinates $x_n = (n - (H-1)/2)/H$ and $x_m = (m - (H-1)/2)/H$. The standard LCT kernel sampled on this grid is complex valued, $K_{\mathrm{full}}[n, m] = \frac{1}{\sqrt{ib}} \exp\!\big(i\pi \frac{a x_n^2 - 2 x_n x_m + d x_m^2}{b}\big)$. The implementation selects either $\frac{1}{H} \operatorname{Re} K_{\mathrm{full}}$ or $\frac{1}{H} \operatorname{Im} K_{\mathrm{full}}$, producing a real-valued quadrature matrix that matches the corresponding analytic component to all-close 10^{-6}. This real matrix is used for token mixing. Composition of two realized quadrature matrices is not assumed to reproduce composition of continuous LCT operators, and the finite grid does not preserve exact unitarity.

B.3 Coordinate-transformation Jacobian. With $s = e^{\ell}$ and $k = \sinh \kappa$, the transformation between the raw coordinates (θ, s, k) and the optimization coordinates $p = (\theta, \ell, \kappa)$ is diagonal,

$$T = \mathrm{diag}\Big(1,\; s,\; \sqrt{1 + k^2}\Big), \qquad g_p = T\, g_{\mathrm{raw}}, \qquad \mathcal{M}_p = T\, \mathcal{M}_{\mathrm{raw}}\, T. \tag{B.4}$$

T **remains nonsingular.** It relates two coordinate representations of the same effective parameters, while the clamp belongs to the separate parameter-to-quadrature map. The fractional-order entry is identically 1, and the other two are $s > 0$ and $\sqrt{1 + k^2} \geq 1$. The smallest realized scale factor is 0.479, so T remains invertible throughout. Saturation instead affects the regularized restricted response described below.

For the undamped full subspace, the quadratic is invariant whenever the metric is nonsingular, with $g_p^\top \mathcal{M}_p^{-1} g_p = g_{\mathrm{raw}}^\top \mathcal{M}_{\mathrm{raw}}^{-1} g_{\mathrm{raw}}$. Identity regularization does not transform covariantly, so $g^\top (\mathcal{M} + \mu I)^{-1} g$ remains coordinate dependent. Because T is diagonal, $\mathrm{diag}\, \mathcal{M}$ transforms in the same way as $\mathcal{M}$, which makes Jacobi regularization invariant for any subset on which the restricted matrix is nonsingular.

$$Q_{\mathrm{cov}}(S) = \tfrac{1}{2} g_S^\top (\mathcal{M}_{SS} + \lambda (\mathrm{diag}\, \mathcal{M})_{SS})^{-1} g_S \quad \text{is coordinate independent for any diagonal } T. \tag{B.5}$$

Domain restriction under saturation. If a selected coordinate is saturated, Jacobi regularization remains singular along that zero row, whereas identity regularization remains well posed. Comparisons between the two regularizations therefore use only their common nonsingular domain. On this domain, the raw-to-optimization transformation agrees to numerical precision. Coordinate-wise sensitivity and selected subsets remain representation dependent, so all coordinate-wise statements use $p = (\theta, \ell, \kappa)$.

A separate derivative convention concerns fractional order. Differentiation with respect to θ or $\phi = \frac{\pi}{2}\theta$ differs by exactly $\pi/2$. Eq. (8) uses ϕ, while optimization and coordinate-wise analysis use θ. E.1 reports both explicitly.

B.4 Coordinate-system robustness. The control compares sensitivity ranking, both response scores, selection overlap, and realized-recovery ranking under four parameterizations, namely (θ, s, k),

$(\theta, \log s, k)$, $(\theta, \log s, \operatorname{asinh} k)$, and a functional normalization by each coordinate's quadrature-matrix Jacobian norm D_i. The regularized Q_{full} ordering is empirically stable, while small-subset membership changes modestly. Between raw and optimization coordinates, Spearman agreement for stored-subset Q ranking ranges from $+0.893$ to $+1.000$ with median $+0.998$, and top-r membership Jaccard ranges from 0.778 to 1.000 with median 0.882. This is an empirical robustness result rather than an invariance claim for identity regularization.

The two regularizations are compared on the 154-subset domain where both are well posed.

regularization	rank agreement with the other	Spearman vs realized recovery, common domain
identity, $\mathcal{M} + \mu I$	—	median $+0.834$
Jacobi, $\mathcal{M} + \lambda \operatorname{diag} \mathcal{M}$	median $+0.933$	median $+0.857$

The two regularizations produce nearly identical subset rankings and similar agreement with realized recovery. Selected subsets at the evaluated subset size still differ (Jaccard median 0.333), partly because Jacobi regularization cannot include a saturated coordinate while identity regularization remains well defined. Identity regularization is used because it applies to every subset, with its measured coordinate dependence reported explicitly.

B.5 One-dimensional and separable quadrature Jacobians. Let K denote the selected real-valued quadrature matrix. The derivative predictor uses $J_i = \partial K / \partial p_i$ and $D_i^{\text{1D}} = \|J_i\|_F$. The implemented two-dimensional mixing matrix is $T_K = K \otimes K$, whose derivative is

$$\frac{\partial T_K}{\partial p_i} = J_i \otimes K + K \otimes J_i, \qquad \left\| \frac{\partial T_K}{\partial p_i} \right\|_F^2 = 2\|J_i\|_F^2 \|K\|_F^2 + 2\langle J_i, K \rangle_F^2. \tag{B.6}$$

For activations, $\partial(KXK^\top)/\partial p_i = J_i X K^\top + K X J_i^\top$.

The separable comparison is also bounded before measurement. Writing $\rho_i = \langle J_i, K \rangle_F / (\|J_i\|_F \|K\|_F)$,

$$D_i^{\text{2D}} = \sqrt{2}\, \|K\|_F\, D_i^{\text{1D}} \sqrt{1 + \rho_i^2}. \tag{B.7}$$

and $\|K\|_F$ is common to every axis within a block, so the two predictors can disagree *within a block* by at most $\sqrt{2}$ in D and a factor 2 in D^2. The model-level comparison and the data-conditioned derivative control are reported in E.1b.

C Experimental protocol

C.1 Precision sweep selecting 8-bit ADC and 6-bit DAC

Conversion-only evaluation uses the official test split with repeated stochastic evaluations for all trained models.

Table A3. Median accuracy loss relative to the pre-mapping software model, in pp. Conversion only, no fine-tuning.

setting	CIFAR-10 ($n = 6$)	CIFAR-100 ($n = 6$)	Tiny ImageNet ($n = 2$)
12-bit ADC / 10-bit DAC	0.79	0.38	0.41
10-bit ADC / 8-bit DAC	1.01	0.81	1.30
9-bit ADC / 7-bit DAC	3.50	2.42	**9.74**

setting	CIFAR-10 ($n = 6$)	CIFAR-100 ($n = 6$)	Tiny ImageNet ($n = 2$)
8-bit ADC / 6-bit DAC	**11.85**	**10.51**	**23.48**

The operating point was fixed from the full mapped-model sweep before any recovery analysis. At 10-bit ADC and 8-bit DAC or higher precision, the CIFAR deployment loss is small, leaving limited recoverable degradation for discriminating response criteria. The 8-bit ADC and 6-bit DAC condition was therefore chosen to provide a measurable intervention range rather than to favor a particular method. No claim is made that the same geometry is invariant to converter precision.

Tiny ImageNet does not determine the operating point. The converter condition is selected from the CIFAR models. Tiny ImageNet is evaluated at the same 8-bit ADC and 6-bit DAC setting only to apply the same conversion operation on the larger token grid, and its values are not used in the CIFAR model-level statistics.

C.2 Trained-model selection

The primary analysis uses prespecified CIFAR models spanning both datasets and quadratures. Selection is fixed before hardware analysis, with initial LCT coordinates drawn from the corresponding $3 \times 3 \times 3$ grid over $\theta \in \{0.25, 0.5, 0.75\}$, $s \in \{0.8, 1.0, 2.0\}$, and $k \in \{-1, 0, 1\}$.

Table A4. Prespecified trained models and pre-mapping starting points. Digital accuracy is in percent. Tiny ImageNet is a separate evaluation (J) and does not enter CIFAR model-level statistics.

group	rule	models (θ_0, s_0, k_0)	digital accuracy
CIFAR-10 imag	accuracy-based selection	$(0.25, 0.8, -1)$, $(0.25, 0.8, 0)$, $(0.75, 0.8, 0)$	88.54, 88.20, 87.81
CIFAR-10 real	prespecified diverse initializations	$(0.25, 1.0, 0)$, $(0.75, 1.0, 0)$, $(0.75, 2.0, 1)$	87.39, 87.66, 87.28
CIFAR-100 imag	accuracy-based selection	$(0.25, 0.8, 0)$, $(0.5, 2.0, 1)$, $(0.75, 0.8, 0)$	70.91, 70.53, 70.92
CIFAR-100 real	prespecified diverse initializations	$(0.25, 0.8, 0)$, $(0.25, 0.8, 1)$, $(0.25, 2.0, 1)$	70.61, 70.61, 68.70
Tiny ImageNet **[separate evaluation, J]**	own-grid best per quadrature	imag $(0.25, 2.0, 0)$, real $(0.25, 1.0, 0)$	57.04, 56.11

The imaginary-quadrature groups use an accuracy-based rule, whereas the real-quadrature groups use a prespecified spread across distinct initialization regions. This avoids concentrating every group on a narrow part of the initialization grid without assigning a causal role to the initialization regime itself.

C.3 Conversion, adaptation scopes, and optimization

AIHWKit conversion utilities replace eligible linear operations with hardware-mapped counterparts and import the pretrained digital weights. All reported comparisons use the same fixed RPU configuration. Hardware-aware fine-tuning keeps the mapped forward computation active while gradients pass through the differentiable surrogate and a digital optimizer updates reference parameters. Iwasawa parameters remain digitally represented. Updated (θ, ℓ, κ) values are transformed back to (θ, s, k), regenerate the selected quadrature matrix K_j, and enter the same hardware-conditioned forward computation. No reported condition performs in-memory device updates.

RPU configuration used for hardware mapping.

setting	value
tile and MVM	maximum 256×256, one-pass analog MVM
conversion precision	8-bit ADC, 6-bit DAC
bounds and management	input/output bounds 10/10, iterative bound management, no noise management
nonidealities	IR drop 0.05, input/output noise 0.01/0.05, PCM read noise 0.05
scaling	column-wise weight and output scaling, learned output scaling, $\omega = 1$

Parameters outside the released adaptation scope are frozen in every reported condition. The adaptation scopes release the full model, the token branch, the channel branch, or all LCT coordinates, while restricted selection releases only the chosen LCT subset. CIFAR conditions run 20 epochs and Tiny ImageNet runs 10 because of its higher computational cost; this schedule difference is stated wherever the settings are compared.

C.4 Observed-label field second moment and estimator stability

Estimator definition. The estimator accumulates one backward field for each example and each read-noise draw. With z_n denoting the sample-level field propagated into the LCT coordinates at the observed label,

$$\mathcal{M} = \frac{1}{N} \sum_n z_n z_n^\top = g g^\top + \mathrm{Cov}(z). \tag{C.1}$$

The field z_n is not $\nabla_p C$ for the hardware-conditioned loss because the straight-through path of G.3 routes LCT-coordinate derivatives through a digital proxy. Throughout, $\mathcal{M}$ is the **sample-level observed-label empirical-Fisher-form second moment of the adaptation field**. It is not a mini-batch gradient second moment, the model Fisher, a generalized Gauss–Newton matrix, or physical-device curvature (Schraudolph, 2002; Kunstner et al., 2019).

Averaging order, and the counterfactual that bounds it. The read-noise loop sits inside the sample-level loop and accumulates one outer product per draw, rather than averaging fields across draws before forming the outer product. The hardware-noise covariance is consequently retained in full and is **not** shrunk by $1/R$.

The opposite averaging order is measured as a control. Recomputing the metric with gradients averaged over draws *before* the outer product, on four models at $R = 2$ and $R = 10$, removes 2.0 to 17.0% of the trace, so read noise is a real but minority component of $\mathcal{M}$. It does not change the selected coordinates. The diagonal sensitivity ranking agrees with the deployed estimate at Spearman ≥ 0.999 in all eight comparisons, and the top-8 selected set is identical with Jaccard 1.000. The averaging order does not affect the reported selection results.

Objective match. The estimator differentiates sample-level hard-label cross-entropy. Hardware-aware fine-tuning minimizes cross-entropy on hard labels, with no mixup and no label smoothing. The metric estimator and hardware-aware fine-tuning use the same hard-label cross-entropy objective. Mixup is used only during digital pretraining and is upstream of the measurements in this appendix.

Calibration and held-out evaluation. Coordinate subsets are constructed from a cached calibration set and scored by held-out-validation cross-entropy, so selection quality is evaluated on samples distinct from those used for subset construction. Each repetition uses 128 examples and $R = 2$ read-noise draws, computed separately for the pre-mapping and mapped pre-fine-tuning states, with tensors cached after augmentation so both estimators see identical inputs. Each of the twelve models contributes three calibration repetitions, giving 768 sample-level fields per state.

Stability. On the representative model, the worst-pair agreement between two calibration draws is Spearman 0.949 at the pre-mapping state and 0.987 at the mapped one, the read-noise-seed worst pair is 0.997, and the leave-one-out minima for subsets selected by Q_{diag} and Q_{full} are 0.876 and 0.740. A minimum Spearman agreement of 0.50 and top-set Jaccard overlap of $1/3$ are required per criterion and subset size. Conditions below either prespecified stability threshold are excluded from the corresponding criterion comparison (Figure A1).

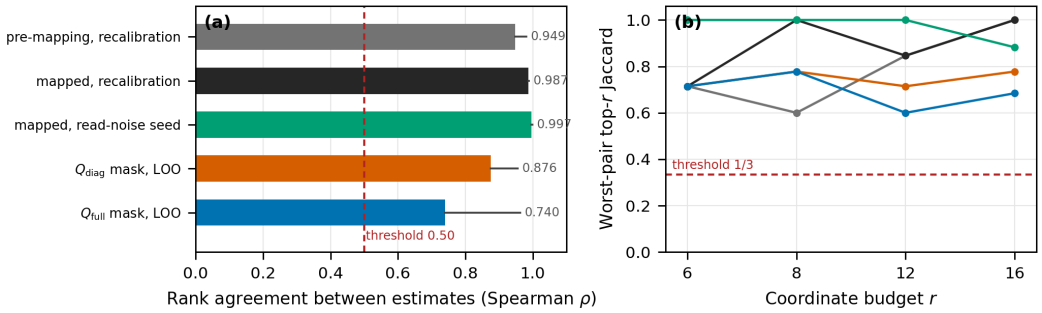

Figure A1: Stability of the empirical-Fisher-form field second moment on a representative model over calibration repetitions and read-noise draws. (a) Leave-one-out rank agreement. (b) Top-r overlap. Horizontal lines mark the prespecified Spearman 0.50 and Jaccard $1/3$ stability thresholds.

Regularization uses μ set from the trace scale of $\mathcal{M}$. The subset-recovery ranking is empirically stable across $\mu \in [0, 1]$ in the declared coordinate system, with Spearman correlation from $+0.73$ to $+0.89$ and $+0.874$ as $\mu \to 0$. This sweep does not establish reparameterization invariance.

C.5 Selection criteria and random controls

Five criteria are evaluated at each subset size. They are pre-mapping sensitivity, post-mapping sensitivity, Q_{diag}, Q_{full} with coupling-aware greedy search over Eq. (6), and stratified random selection. Three independently sampled random subsets are evaluated per condition. G also includes axis-composition-matched and axis-and-block-composition-matched random controls that reproduce the composition of the response-based subset.

D Local response, restricted recoverability, and validity

Section 4.1 states the local response compactly. The derivation and an algebraic property of the subset score are given below.

D.1 Deployment as a perturbation. Write $\mathcal{L}_\varepsilon = \mathcal{L}_0 + \varepsilon V + O(\varepsilon^2)$. Assume that $\nabla_p \mathcal{L}(p, \varepsilon)$ has the required local regularity, $p^\star$ is an interior digital stationary point, and $H_0 = \nabla^2 \mathcal{L}_0(p^\star)$ is nonsingular on the LCT subspace. The implicit-function theorem then defines a local stationary branch with $v = \nabla V(p^\star)$ and

$$p^\star_\varepsilon - p^\star = -\varepsilon H_0^{-1} v + O(\varepsilon^2), \qquad \Delta p_i \approx -\varepsilon\, v_i / H_{0,ii} \text{ under a diagonal approximation.} \quad \text{(D.1)}$$

This is a stationary-branch response, not automatically a local-minimum branch when H_0 is indefinite. The diagonal expression illustrates that forcing and curvature both enter movement. It is not used as the operational predictor. Coordinates on active clamp boundaries require a KKT or active-set analysis and lie outside this unconstrained reference.

D.2 Regularized restricted response. A finite minimum of the unrestricted quadratic requires a positive-definite restricted second-order matrix. Under the chosen identity regularization, that condition holds for only 7 of 214 restricted subsets of the symmetrized field Jacobian (G.3). The mapped-state intervention uses the separately defined regularized second-moment problem

$$\min_{\operatorname{supp}(\delta) \subseteq S} g_S^\top \delta_S + \tfrac{1}{2} \delta_S^\top (\mathcal{M}_{SS} + \mu I) \delta_S, \qquad \delta_{\mathcal{M},S} = -(\mathcal{M}_{SS} + \mu I)^{-1} g_S. \quad \text{(D.2)}$$

with surrogate decrease

$$Q_{\mathcal{M}}(S) = \tfrac{1}{2} g_S^\top (\mathcal{M}_{SS} + \mu I)^{-1} g_S. \quad \text{(D.3)}$$

$Q_{\mathcal{M}}$ is exact for this declared regularized metric problem and is compared ordinally with realized recovery. It is not a Fisher-information quantity, not a leading-order Taylor recoverable loss, and not the update rule the optimizer follows.

The coupling-aware score is monotone under coordinate addition. With $A = \mathcal{M} + \mu I \succ 0$ and $j \notin S$, the Schur complement gives

$$Q_{\mathcal{M}}(S \cup \{j\}) - Q_{\mathcal{M}}(S) = \frac{1}{2} \frac{(g_j - A_{jS} A_{SS}^{-1} g_S)^2}{A_{jj} - A_{jS} A_{SS}^{-1} A_{Sj}} \geq 0. \tag{D.4}$$

The nonnegative increment justifies greedy addition as a score-increasing procedure but does not establish submodularity or global optimality at a prescribed subset size.

E Quadrature differentials and independent confirmation of coordinate ordering

E.1 Analytic quadrature differentials

All quantities in this subsection are obtained by differentiating the selected real-valued quadrature matrix in float64 with autograd, without using trained-model states. The fractional-order derivatives with respect to normalized order θ and physical angle $\phi = \frac{\pi}{2}\theta$ differ by exactly $\pi/2$, as detailed in B.3.

E.1a Analytic sweep in the optimization coordinates ($p = (\theta, \ell, \kappa)$, $s = 1$, $k = 0$). Eq. (8) is written in ϕ, while optimization uses θ. The complete analytic comparison is summarized in Figure A2.

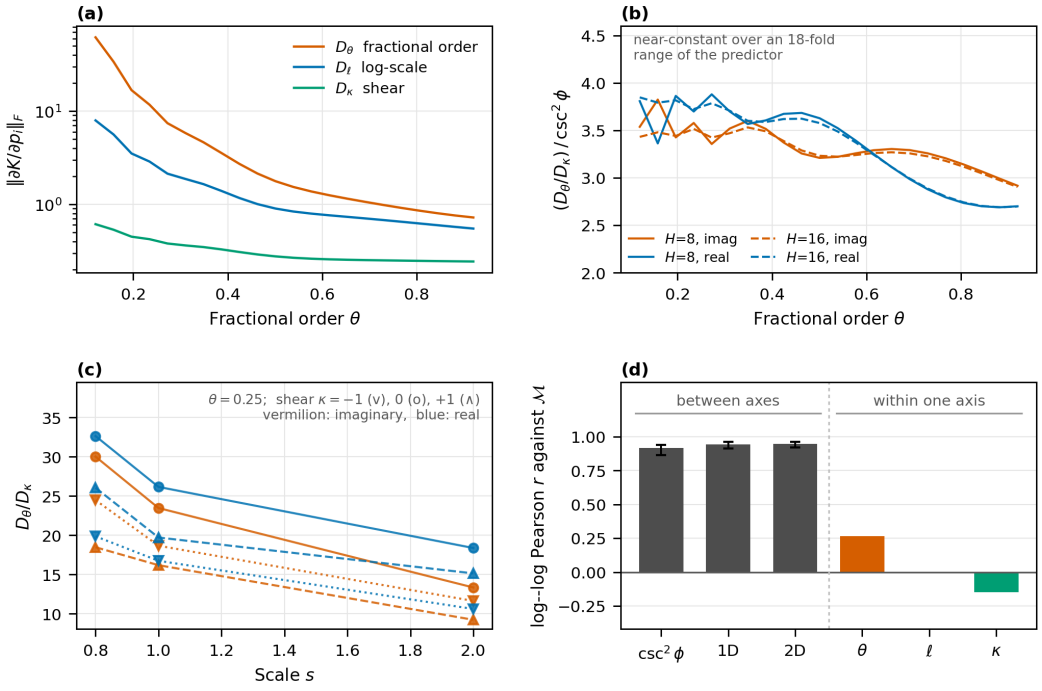

Figure A2: Quadrature differentials on the analytic grid at $H = 8$ and $H = 16$. Panels show axis norms, the D_ϕ/D_k envelope comparison, scale/shear modulation, and between-axis versus within-axis model association.

Table A5. Quadrature-Jacobian sweep. Both quadratures and $H \in \{8, 16\}$. Norms in the declared coordinate system, ratios unitless.

H	quad.	θ	$\csc^2\phi$	D_θ	D_ℓ	D_κ	D_θ/D_κ	residual $/\csc^2\phi$
8	imag	0.15	18.350	39.211	6.159	0.546	71.764	3.911
8	imag	0.25	6.828	9.432	2.486	0.402	23.479	3.438
8	imag	0.50	2.000	1.776	0.903	0.277	6.422	3.211
8	imag	0.90	1.025	0.740	0.561	0.244	3.030	2.956
8	real	0.25	6.828	9.945	2.519	0.380	26.170	3.833
8	real	0.90	1.025	0.670	0.476	0.243	2.760	2.692
16	imag	0.25	6.828	9.792	2.576	0.412	23.752	3.478
16	real	0.90	1.025	0.685	0.481	0.248	2.759	2.691

The residual quotient sits between 2.69 and 3.91 over an operating range across which the predictor itself varies by a factor of 18, supporting the $\csc^2\phi$ scaling as a property of the selected quadrature rather than a training artifact. The quotient is nearly independent of quadrature and grid size, and the three-way ordering $D_\theta > D_\ell > D_\kappa$ holds pointwise at every sampled operating point, in both quadratures and at both grid sizes, the quadrature-level counterpart of the $\theta > \ell > \kappa$ sensitivity ordering measured in the twelve models (F.1), obtained without trained-model information.

Scale and shear modulate the anisotropy without disturbing that scaling. Between $s = 0.8$ and $s = 2.0$ at $\theta = 0.25$ the ratio moves by a factor of 2.2 (imaginary) or 1.8 (real), the direct signature of the x_m^2/s^2 term in $\partial q/\partial\phi$, while shear modulates weakly and antisymmetrically between quadratures, consistent with the same antisymmetry E.2 derives exactly for the energy derivative.

E.1b Model-level comparison. The quadrature-derivative predictors are evaluated at each block's converged coordinates after excluding near-boundary fractional-order states. Table A6 compares each predictor with $\mathcal{M}_{\theta\theta}/\mathcal{M}_{\kappa\kappa}$ on a log scale.

Table A6. Between-axis anisotropy predictors. Intervals resample trained models because blocks are nested within models.

predictor	Pearson r	95% CI
$\csc^2\phi$ asymptotic envelope	$+0.915$	$[+0.865, +0.940]$
1D sampled-quadrature Jacobian	$+0.942$	$[+0.914, +0.961]$
2D separable mixing matrix $K \otimes K$	$+0.946$	$[+0.919, +0.962]$

The finite-grid predictor improves modestly on the asymptotic envelope, while the exact separable derivative adds little beyond the one-dimensional quadrature predictor. The dominant between-axis anisotropy is therefore already captured at the quadrature level.

E.1c Between-axis scope and depth-dependent residual. The quadrature Jacobian predicts the dominant axis substantially better than block-to-block variation within one axis. Normalizing by the quadrature derivative does not remove the depth-dependent residual, and the data-conditioned derivative on cached activations does not close the gap. The remaining variation is therefore attributed primarily to the upstream factor supplied by the surrounding trained model.

E.1d Deployment does not create the limit. The same between-axis agreement, weak within-axis association, and depth dependence are already present at the pre-mapping trained state. Hardware mapping inherits and slightly perturbs this structure rather than creating it.

E.2 Quadrature antisymmetry of the shear-energy derivative

E.2a Exact linearity in shear. With $b = s\sin\phi$ and $d/b = k + \cos\phi/(s^2\sin\phi)$,

$$q = \frac{a}{b}x_n^2 - \frac{2x_n x_m}{b} + \underbrace{k\,x_m^2}_{\text{linear}} + \frac{\cos\phi\, x_m^2}{s\,b}, \qquad \frac{\partial q}{\partial k} = x_m^2, \qquad \boxed{\frac{\partial^2 q}{\partial k^2} = 0}, \qquad \frac{\partial A}{\partial k} = 0. \quad \text{(E.1)}$$

Shear does not touch the amplitude and enters the phase exactly linearly, $\Phi(k) = \Phi_0(\phi, s) + \pi k x_m^2$. All nonlinearity along this axis comes from $\sin$ and $\cos$ alone.

E.2b Antisymmetry. Expanding the selected quadrature-matrix energy,

$$\|K_{\text{imag}}\|^2 = A^2 \sum \tfrac{1-\cos 2\Phi}{2}, \qquad \|K_{\text{real}}\|^2 = A^2 \sum \tfrac{1+\cos 2\Phi}{2},$$
$$\frac{\partial \|K_{\text{imag}}\|^2}{\partial k} = +A^2 \sum_{n,m} \pi x_m^2 \sin 2\Phi = -\frac{\partial \|K_{\text{real}}\|^2}{\partial k}. \tag{E.2}$$

The two derivatives are exactly opposite in sign and equal in magnitude. This is verified numerically at all nine (ϕ, s) combinations of the canonical grid (for example $\theta 0.25/s1.0$. Imaginary -0.2065, real $+0.2065$).

E.2c Premise and its strength. The prediction requires local training pressure toward larger token-mixing energy. This premise holds for a majority of blocks across datasets and quadratures but is not universal. Restricting the direction test to blocks satisfying the premise gives near-perfect agreement, although that conditional result is not treated as independent evidence because the predictor is the corresponding energy derivative on those blocks.

E.2d Controls on the directional prediction. Four alternatives are scored on the same model set.

Table A7. Shear-direction controls. The broader pre-mapping training grids are used only for this directional diagnostic and are separate from the CIFAR inferential model set. The nonzero-shear subset is shown separately because the shrink-toward-zero baseline is undefined at $k_0 = 0$.

predictor	block subset	agreement
hybrid-state $\partial\|K\|^2/\partial k$	all	82.9%
initial-state $\partial\|K\|^2/\partial k$	all	83.3%
majority realized direction (constant)	all	50.8%
hybrid-state derivative	$k_0 \neq 0$	75.1%
initial-state derivative	$k_0 \neq 0$	74.8%
shrink toward zero, $-\operatorname{sign}(k_0)$	$k_0 \neq 0$	70.8%
majority direction	$k_0 \neq 0$	50.5%
initial-state derivative	$k_0 = 0$	93.6%
hybrid-state derivative	$k_0 = 0$	92.2%
initial-state derivative vs. first recorded epoch	all	92.7%
hybrid-state derivative vs. first recorded epoch	all	91.8%

The hybrid construction adds no predictive information. Evaluating the derivative at the fully initial (θ_0, s_0, k_0) scores 83.3% against the hybrid 82.9%, so the prediction requires no converged information and is available before training. **The margin over shrinkage toward zero is small where that baseline exists.** On the 864 blocks initialized away from zero shear, the initial-state predictor scores 74.8% against 70.8% for shrink-toward-zero. The broader-grid advantage is strongest in the 720 blocks initialized at $k_0 = 0$, where the initial-state predictor is correct for 674 blocks, or 93.6%, while the hybrid-state control gives 92.2%. All 288 Tiny ImageNet blocks have $k_0 = 0$ and score 96.9% in this test. To keep the main-text diagnostic separate from the descriptive Tiny ImageNet extension, Fig. 2b uses only the 1296 CIFAR blocks: 80.3% overall, 74.8% for $k_0 \neq 0$, and 91.4% for $k_0 = 0$. **The association is strongest early.** The initial-state predictor agrees with displacement to the first recorded epoch in 92.7% of the broader grid, above its 83.3% endpoint agreement; the corresponding hybrid-state values are 91.8% and 82.9%.

E.3 Functional coordinate substitution as an independent confirmation

Coordinate substitution provides a field-independent intervention. All parameters of a recipient model are frozen except the LCT coordinates, which are replaced by values from separately trained source models before evaluation without retraining. Recipient–source pairs span distinct initial fractional-order settings in both CIFAR tasks and quadratures.

CIFAR-100. Fractional-order-only substitution reproduces most of the degradation caused by replacing all three coordinates, with interactions between coordinates producing some non-additive cases.

Quadrature-matrix distance tracks the resulting degradation, but this association is not interpreted as a global diagonal-geometry result because fractional order affects both quantities.

Quadrature-matrix rescaling rules out a global-gain explanation. Because θ also changes the LCT amplitude factor, substituted matrices were rescaled to the recipient Frobenius norm while the other coordinates were held fixed. The degradation persists under this control, so the substitution effect is not explained by a global quadrature-matrix norm mismatch.

CIFAR-10 shows the same overall pattern, although the imaginary-quadrature comparisons are less uniform when independently trained models converge to similar fractional-order states. The Frobenius-rescaling control leads to the same conclusion.

Scope of coordinate substitution. Because the channel MLP, affine parameters, LayerScale, head, and patch embedding of the recipient model are all frozen, substitution shows that the remaining computation is adapted to its own token-mixing matrix. It does not attribute recovered accuracy exclusively to the channel branch, which is tested separately in H.2.

F Model-level geometry, movement, and its prediction

Movement is defined at the block level, with blocks nested inside independently trained models. Block-level counts are descriptive rather than independent samples. Trained models are the inferential unit, and every interval or permutation test below resamples or permutes models unless stated otherwise.

F.1 Coordinate ordering. Movement is measured from each model's mapped pre-fine-tuning state to the epoch of lowest validation loss. The coordinate ordering is unchanged when epoch 20 is used instead or when blockwise displacement is summarized by the mean, median, or RMS.

Table A8. Group-level coordinate ordering before and after hardware-aware adaptation. Entries summarize the dominant ordinal pattern in the declared optimization coordinates.

group	pre-mapping sensitivity	post-mapping sensitivity	realized movement
CIFAR-10 imag	$\theta > \ell > \kappa$	$\theta > \ell > \kappa$	mostly $\kappa > \ell > \theta$
CIFAR-10 real	$\theta > \ell > \kappa$	$\theta > \ell > \kappa$	$\kappa > \ell > \theta$
CIFAR-100 imag	$\theta > \ell > \kappa$	$\theta > \ell > \kappa$	$\ell > \kappa > \theta$
CIFAR-100 real	$\theta > \ell > \kappa$	$\theta > \ell > \kappa$	$\kappa > \ell > \theta$

Fractional order is consistently the most sensitive axis before and after mapping and the least moved axis after adaptation. Whether scale or shear dominates the realized movement depends on the dataset–quadrature group. Tiny ImageNet is evaluated separately in Appendix J.

F.2 One-step prediction against realized movement. From the stored per-epoch trajectories, with no new runs. Regularization follows the deployed convention $\mu = 0.2 \times \mathrm{mean}(\mathrm{diag}\,\mathcal{M})$, per model. The trajectory comparison is summarized in Figure A3.

Sign structure does not depend on the fine-tuning scope. Fractional order and shear agree in sign under both full-model and all-coordinate adaptation, while scale is the only axis whose sign is unstable. This is a within-axis question about whether the predicted step tracks the direction of a moving coordinate, and it is distinct from the between-axis allocation in F.4.

Alignment is strongest early. The cosine between the full-$\mathcal{M}$ response of Eq. (5) and realized displacement is consistently higher early in adaptation than at the endpoint. Realized displacement is also larger than the local predicted step, and matching the adaptation scope generally improves whole-space alignment.

Relinearizing does not restore the prediction. Re-forming the state during fine-tuning and predicting only the next interval separates a stale-linearization explanation from failure of the local approximation. Median alignment decreases from $+0.382$ early to $+0.060$ later in training, so repeatedly updating the local geometry does not restore trajectory-level prediction.

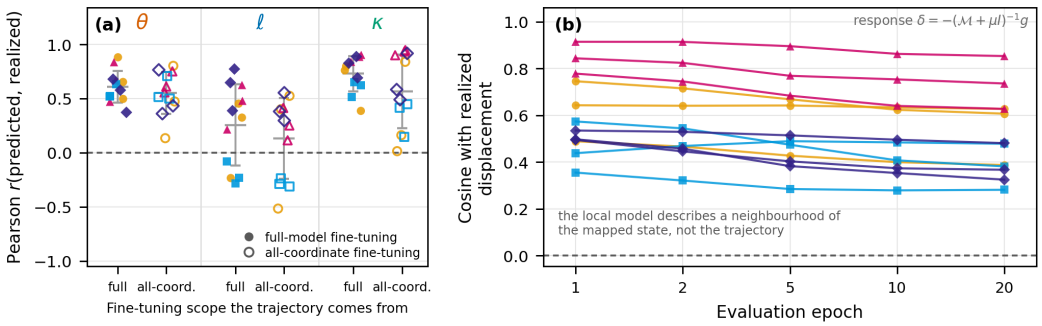

Figure A3: Movement prediction on twelve CIFAR models. Coordinate-wise predicted-versus-realized correlations are separated by adaptation scope, and whole-space alignment is shown against epoch from the mapped-state linearization.

Shrinking displacement partly confounds the decay. Realized displacement becomes smaller later in training, so late-epoch alignment is measured in a weaker-signal regime. Within-epoch checks do not explain away the overall decline, supporting the narrower interpretation that the local model best describes the opening of adaptation.

Scale is the unstable axis. Its movement-prediction alignment differs by quadrature, with an exact model-level permutation test giving $p = 0.013$. The corresponding movement-ordering comparison points in the same direction but remains unresolved.

Per-block form. The diagonal step $\delta p_i^\star = -g_i/(\mathcal{M}_{ii} + \mu)$, evaluated once at the mapped state, is compared with cumulative 20-epoch movement using block-wise Pearson correlation on the full-model trajectory.

F.3 Three controls on the movement inversion.

The sensitivity ordering predates mapping. Fractional order already ranks first in sensitivity at the pre-mapping state and retains that ordering after mapping, so the sensitivity–movement inversion is not produced by conversion.

Saturation inflates it but does not manufacture it. Excluding saturated fractional-order coordinates weakens the movement inversion but does not remove it, while the sensitivity ordering becomes cleaner. The dissociation is therefore not an artifact of pinned coordinates.

F.4 Blockwise depth structure. Relative depth is $b/(L-1)$, so the eight- and sixteen-block stacks share an axis. Figure 3(b,c) plots the profiles. Per-model correlations and counts of models are not used. They describe a shape only if it is monotone and similar everywhere, and it is neither.

Adaptation concentrates in the early blocks without being a boundary artifact. The shallower half of the blocks carries a median 0.69 of a model's total displacement against the 0.50 expected from an even split, and the first block carries 0.17 against 0.11. The forcing magnitude decreases with depth more strongly than displacement per unit forcing increases. Total displacement can consequently fall while normalized response rises. The regularization term in Eq. (5) affects the local response scale, but it is not used as the causal explanation for the observed depth profile.

The composition changes gradually rather than at a single transition depth. Shear contribution generally decreases with depth while scale contribution rises, but their ordering can change repeatedly rather than at one stable crossover. Fractional-order share increases with depth yet remains typically the smallest component.

Eq. (5) does not account for how movement divides between the axes. The predicted top-moving axis agrees only modestly with the realized one, while bottom-axis agreement is stronger. In particular, scale contribution remains systematically underestimated in the deeper blocks, limiting interpretation of the local response as a detailed movement-allocation model.

Axis-allocation prediction is sensitive to the regularization coefficient. Sweeping the regularization ratio two decades around the deployed 0.2, the across-axis agreement changes substantially even

though the recovery ranking remains comparatively stable. At about a hundredth of the deployed value the predicted fractional-order share approaches the realized share, but the scale contribution remains underestimated across the sweep. Taking $\mu \to 0$ also over-corrects the fractional-order ranking. Changing regularization can improve one component of the movement allocation without reproducing the full three-axis pattern.

F.5 Sensitivity to the regularization coefficient. $\mu = 0.2 \times \mathrm{mean}(\mathrm{diag}\,\mathcal{M})$ is fixed by the protocol in C.4. Figure A4 contrasts subset-ranking stability with axis-allocation sensitivity over the regularization sweep. Across the same sweep, the median within-model association between $Q_{\mathcal{M}}(S)$ and realized recovery remains positive and comparatively stable, whereas the predicted axis allocation changes substantially. The main selection claim is less sensitive to μ than any attempt to interpret Eq. (5) as a detailed model of the realized movement split.

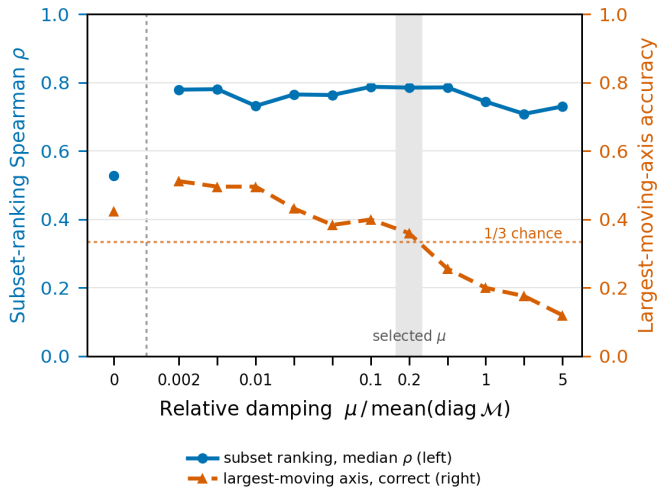

Figure A4: Regularization sensitivity over a factor of 2500. The median within-model subset-ranking correlation remains comparatively stable while the fraction of blocks whose largest-moving axis is predicted changes substantially. The deployed ratio is shaded.

Relative rather than absolute regularization matters. The deployed rule scales μ to each model's metric. Replacing this rule with one fixed small value destabilizes the response along soft coordinates, so analyses of $\delta^\star$ use the declared relative regularization rather than a shared absolute constant.

G Subset Scoring and Selection Controls

G.1 $Q_{\mathcal{M}}(S)$ against realized recovery within each model. For every subset fine-tuned within a model, $Q_{\mathcal{M}}(S) = \frac{1}{2} g_S^\top (\mathcal{M}_{SS} + \mu I)^{-1} g_S$ is computed at the mapped state from stored g and full $\mathcal{M}$ and compared with realized loss reduction.

The reported coefficient uses a **matched one-third subset-size comparison**, with the same fractional coordinate budget in both CIFAR tasks. Deterministic criteria and stratified-random controls are compared at this matched budget, while denser CIFAR-10 sweeps are retained as supporting analyses.

Table A9. Within-model subset prediction at the matched one-third subset size. Spearman coefficients are unitless. Intervals resample trained models.

model set	aggregated within-model rank ρ	95% CI
matched one-third subset size, all CIFAR models	$+0.705$	$[+0.607, +0.798]$
matched one-third subset size, CIFAR-10	$+0.774$	—
matched one-third subset size, CIFAR-100	$+0.637$	—
support: all stored subset sizes	$+0.768$	—

The matched-budget association is positive in both CIFAR tasks. The denser mixed-size sweep gives a somewhat larger aggregate coefficient and is treated as supporting evidence rather than the basis of the reported comparison.

Two normalizations must not be confused. Raw $Q_{\mathcal{M}}(S)$ depends on each model's mean-field magnitude, second-moment geometry, and regularization, and absolute predicted levels have little cross-model relation to realized levels (I.4). The reported association uses ranks normalized within each model and supports an ordinal within-model claim.

G.2 Clustered inference. Evaluated subsets are nested inside trained models, so resampling and paired contrasts use the trained model as the independent unit. At the matched one-third budget, the aggregated within-model rank association is $\rho = +0.705$ with 95% CI $[+0.607, +0.798]$, and the denser full-subset sweep gives a similar result. Paired contrasts favor the response criteria over both stratified random selection and sensitivity-based selection. A denominator-floor control likewise leaves the normalized recovery summary qualitatively unchanged, indicating that low-capacity models are not driving the result.

G.3 Adaptation-field Jacobian and second-moment metric. The adaptation field is differentiated a second time by double backward through the differentiable hardware-training surrogate for all twelve models. The measured object is not a Hessian. Figure A5 summarizes the diagnostic checks. A Hessian of a fixed scalar realization is symmetric, whereas this Jacobian is not and the field has no scalar potential:

$$J_{\text{field}} = \frac{\partial g}{\partial p}, \qquad S_{\text{field}} = \tfrac{1}{2}\left(J_{\text{field}} + J_{\text{field}}^{\top}\right). \tag{G.1}$$

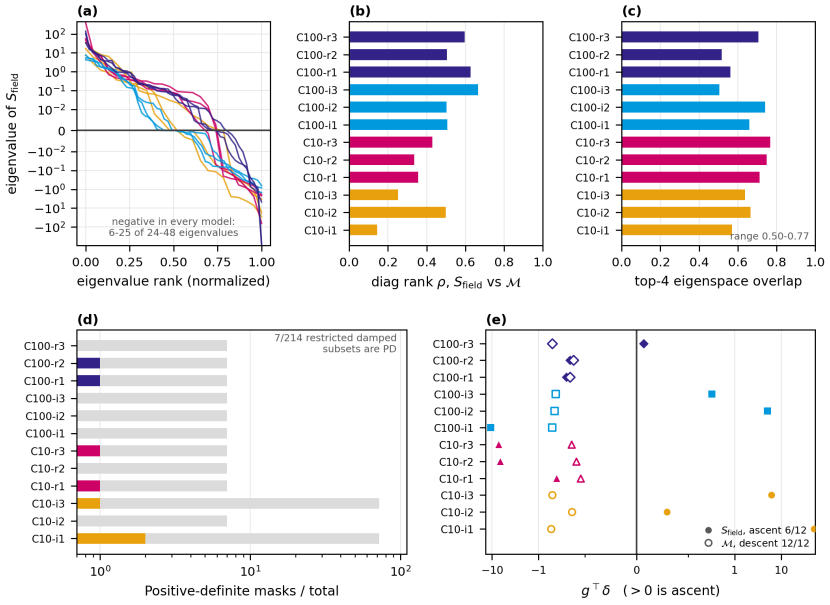

Figure A5: Adaptation-field Jacobian diagnostics across the CIFAR analysis set. Signed spectra of the symmetrized field Jacobian, diagonal-rank agreement with $\mathcal{M}$, leading-subspace overlap, restricted positive definiteness, and solve direction. The field Jacobian is not a Hessian.

The LCT subspace is $3L$-dimensional, so the full Jacobian is formed from $3L$ double-backward passes averaged over ten read-noise draws. This is exact with respect to the surrogate used for hardware-aware training. It is not a second derivative of the physical forward computation or of a scalar deployment objective.

The asymmetry survives the diagnostic controls. It remains under shared-noise evaluation, after excluding saturated coordinates, under central differences, and under an ideal forward pass. The field also differs measurably from the gradient of the evaluated loss, confirming that the asymmetry is not explained by sampling noise or analog nonideality.

The straight-through construction explains the nonconservative field. The forward value uses the nonideal analog product while LCT-coordinate derivatives flow through a digital proxy of the separable map. Mixed derivatives of this assembled field need not commute, so the symmetrized field Jacobian is not a Hessian. Restricted quadratics built from S_{field} are frequently indefinite and can yield ascent directions, whereas the regularized $\mathcal{M}$ response gives descent across the evaluated model set. A trust-region or cubic-regularized construction would define a different intervention and is not evaluated here (Steihaug, 1983; Nesterov & Polyak, 2006).

Predictive success does not identify optimizer mechanism. The optimizer does not apply $\mathcal{M}^{-1}g$. The observed association supports $\mathcal{M}$ as a local intervention-ranking metric, not as the geometry governing the realized training trajectory.

Table A10. Direction cosine against realized displacement. Median cosine for three candidate step directions formed from the stored g and $\mathcal{M}$. Saturated coordinates are excluded.

direction	median cosine, epoch 1	median cosine, endpoint
plain descent $-g$	$+0.392$	$+0.303$
diagonally preconditioned $-g/\sqrt{\operatorname{diag}\mathcal{M}+\mu}$	$+0.565$	$+0.476$
response $-(\mathcal{M}+\mu I)^{-1}g$	$+0.554$	$+0.480$

Metric information helps. Both preconditioned directions beat the raw gradient by 0.17–0.18 at either horizon, consistent with the covariance term carrying the score (G.4). But **the full response is not distinguishable from a diagonal preconditioner**, differing by 0.011 at epoch 1 and 0.004 at the endpoint, so the coupling that Q_{full} exploits for subset selection is not what shapes the realized trajectory.

G.4 Which part of $\mathcal{M}$ carries the score. Figure A6 separates the mean-field and covariance contributions. The rank-one term $gg^\top$ carries 0.006 to 0.317 of the trace of $\mathcal{M}$ across the model set, median 0.081, so most of the metric is the covariance of the field rather than its mean. With $\mathcal{M} = gg^\top + \operatorname{Cov}(z)$, the rank-one term admits a closed form under identity regularization,

$$\tfrac{1}{2}g_S^\top\left(g_S g_S^\top + \mu I\right)^{-1} g_S = \frac{\|g_S\|^2}{2\left(\mu + \|g_S\|^2\right)}. \tag{G.2}$$

This quantity is monotone in $\|g_S\|$, so the rank-one-only score is exactly a ranking by summed squared forcing. The measured Spearman agreement is $+1.000$ in 12 of 12 models, providing an arithmetic check before the component-wise recovery comparison.

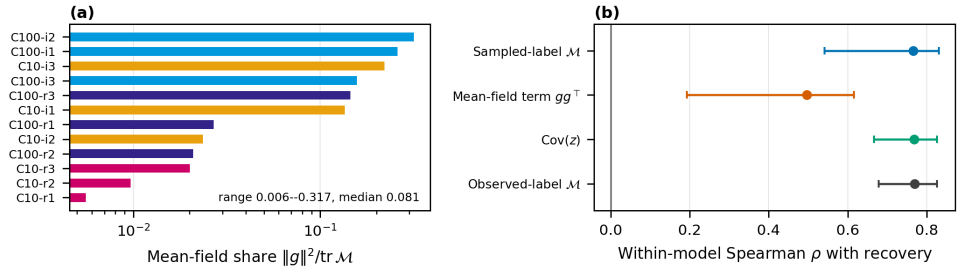

Figure A6: Which component of $\mathcal{M}$ carries the response score. (a) The rank-one trace share $\|g\|^2/\operatorname{tr}\mathcal{M}$. (b) Pooled within-model rank correlation against realized recovery for the deployed metric, covariance term, rank-one term, and sampled-label metric.

Table A11. Which component of the metric predicts recovery. Intervals resample trained models.

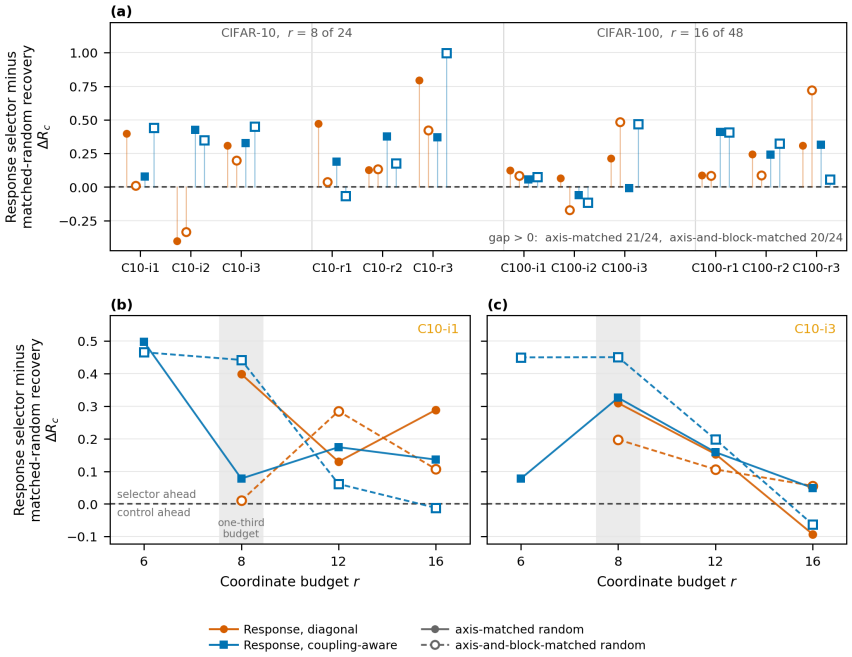

Figure A7: Composition-matched random controls. (a) Difference in R_c between each response criterion and the stronger of three matched random subsets at the one-third coordinate budget across all twelve CIFAR models. Filled markers use axis-matched controls and open markers use axis-and-block-matched controls. Positive values favor the response criterion. (b,c) Budget sweeps for C10-i1 and C10-i3. The shaded region marks the one-third budget used in the primary comparison.

scoring matrix	pooled ρ	95% CI	selected-subset Jaccard against the deployed subset
$\mathcal{M}$, as deployed	$+0.770$	$[+0.679, +0.826]$	n/a
covariance $\mathrm{Cov}(z)$ alone	$+0.769$	$[+0.666, +0.826]$	0.778–1.000 (median 1.000)
rank-one $gg^\top$ alone ($\equiv$ forcing)	$+0.496$	$[+0.193, +0.616]$	0.333–0.882 (median 0.600)
sampled-label metric, observed forcing	$+0.765$	$[+0.542, +0.831]$	0.600–0.882
sampled-label metric *and* sampled forcing	$+0.296$	$[-0.185, +0.519]$	0.231–0.524

The covariance term reproduces both the deployed ranking and, in the median, the deployed subset, whereas the forcing term does not. The score is not a repackaged forcing magnitude. The last two rows are discussed in I.2. Sampling labels strongly attenuates the mean field, so only the middle sampled-label row isolates the metric.

G.5 Greedy near-optimality. Section 4.3 uses greedy search to optimize Q_{full}. On CIFAR-10 the search can be checked exhaustively because 24 coordinates at subset size 8 is $\binom{24}{8} = 735{,}471$ subsets, and all of them are enumerated. Greedy returns the **exact global optimum in 5 of 6 models**. In the sixth exactly one subset beats it, by 0.056% in score, with coordinate overlap 0.875. CIFAR-100 ($\binom{48}{16} \approx 2.25 \times 10^{12}$) is out of reach, so greedy is compared against hill climbing on single swaps and a width-64 beam search. These improve on greedy in 1 of 6 models, by 0.031%. The criterion differences reported in Section 5.3 are not explained by incomplete search. No submodularity or general global-optimality guarantee is inferred from this finite check.

G.6 Selection-criterion comparison at matched one-third subset size. Every model is scored by R_c of Eq. (10) under the same held-out protocol. The ceiling and restricted-selection stages begin from statistically indistinguishable converted states. Normalizing by the LCT-coordinate recovery ceiling avoids conflating selection quality with the amount of recoverable conversion damage and does not change criterion ordering within a model.

A Q criterion is first in 11 of 12 models, the sole exception being C100-i1 where pre-mapping sensitivity leads all five. The CIFAR-10 / CIFAR-100 reversal in the two sensitivity baselines survives the common normalization. Post-conversion sensitivity exceeds pre-mapping sensitivity in five of six CIFAR-10 models (the sixth a tie at 0.511 against 0.514), and pre-mapping sensitivity is at or above post-mapping sensitivity in four of six CIFAR-100 models.

G.7 Composition-matched random controls. Response-based subsets can differ from random subsets in how many coordinates they select from each LCT axis or block. Figure A7 reports both matched-control families and the denser budget sweeps. Two controls isolate this composition effect. The axis-matched control preserves the selected subset's fractional-order, scale, and shear counts. The axis-and-block-matched control preserves both axis and block composition. Three random subsets are drawn for each control, and the strongest realized recovery is used, making the comparison conservative with respect to random variation. Define ΔR_c as the response criterion's R_c minus the corresponding matched-control R_c.

At the one-third coordinate budget, $\Delta R_c > 0$ in 21 of 24 response-criterion comparisons against the axis-matched control and in 20 of 24 against the axis-and-block-matched control across the twelve CIFAR models. Both response criteria and both datasets contribute positive and negative cases, so the control does not support uniform dominance. The predominance of positive gaps shows that the recovery advantage is not explained only by allocating coordinates to favorable axes or blocks.

The denser budget sweeps for C10-i1 and C10-i3 show that the matched-control gap is budget dependent. Both response criteria remain ahead of both control families at the one-third budget used for the primary comparison, whereas some gaps shrink or cross zero at other subset sizes. The supported claim is specific to selection quality at the declared operating budget rather than universal superiority over composition-matched random subsets.

G.8 Seed and run replication. Seeds 43 and 44 include both response criteria, both sensitivity baselines, and all three random controls for every CIFAR model. Comparisons are paired within model and seed. Analyses outside these replication tests use seed 42 unless stated otherwise.

Run-to-run variation at a fixed seed. The experiment reseeds immediately before each condition, so nominally identical conditions should agree up to nondeterministic execution. Six duplicate-condition pairs test that directly. Two criteria happened to select the same coordinate set on the same model at the same budget, and each pair was fine-tuned twice from the same converted state, at the same seed, with pre-fine-tuning losses agreeing to full printed precision.

Table A12. Repeat runs of identical subsets at identical seeds. ΔR_c is the difference in structured-ceiling-normalized recovery between the two runs of one subset.

model	budget	criteria that coincided	Δ loss	ΔR_c	best epoch
C10-r1	8	Q_{diag}, Q_{full}	0.0444	0.269	7 vs 5
C100-r1	16	pre-mapping sensitivity, post-mapping sensitivity	0.0174	0.143	9 vs 18
C10-i1	6	post-mapping sensitivity, random	0.0054	0.008	20 vs 6
C10-i3	16	pre-mapping sensitivity, post-mapping sensitivity	0.0033	0.006	15 vs 16
C10-i1	8	highest-scoring coordinate per block, θ-only	0.0019	0.003	3 vs 11
C100-r2	16	pre-mapping sensitivity, post-mapping sensitivity	0.00003	0.0003	15 vs 15

Duplicate runs at a fixed training seed quantify residual run-to-run variation rather than seed sensitivity. Most duplicate conditions agree closely, while the larger discrepancies coincide with different selected validation epochs, suggesting that flat validation trajectories can amplify small numerical differences through epoch selection.

Comparison with stratified random across seeds. The selected subsets, trained model state, and conversion configuration are held fixed while the fine-tuning seed is varied. Paired within-seed margins remove variation shared by the response-selected and random-control arms. The response criterion retains a positive margin in nearly all model–seed comparisons, although the size of that margin varies across models.

Table A13. Within-seed response-criterion margin over stratified random at seeds 42, 43, and 44. Positive values favor the response criterion. The last column is the range across seeds within each model. Seed 42 uses the subset estimated from its own calibration run, while seeds 43 and 44 reuse the corresponding fixed subset within each comparison arm.

model	seed 42	seed 43	seed 44	spread
C10-i1	$+0.4321$	$+0.4271$	$+0.4292$	0.0050
C10-i2	-0.0072	$+0.0275$	$+0.0290$	0.0363
C10-i3	$+0.2061$	$+0.2109$	$+0.2112$	0.0051
C10-r1	$+0.0924$	$+0.2701$	$+0.0462$	0.2239
C10-r2	$+0.0650$	$+0.0715$	$+0.0427$	0.0288
C10-r3	$+0.1360$	$+0.0781$	$+0.1145$	0.0579
C100-i1	$+0.0216$	$+0.0150$	$+0.0143$	0.0073
C100-i2	$+0.0295$	$+0.0275$	$+0.0287$	0.0020
C100-i3	$+0.1310$	$+0.1268$	$+0.1313$	0.0045
C100-r1	$+0.0313$	$+0.0185$	$+0.0349$	0.0163
C100-r2	$+0.0350$	$+0.0329$	$+0.0018$	0.0332
C100-r3	$+0.0834$	$+0.0657$	$+0.0713$	0.0177

One near-zero case changes sign across seeds, while the larger variability is concentrated in a small number of models and can arise in either comparison arm. The model-set pattern is more stable than the effect size of any individual run.

Comparison with sensitivity across seeds. The same paired construction is used for response- and sensitivity-selected subsets. Each arm keeps its selected subset fixed across the additional fine-tuning seeds so that the within-model spread reflects training variation rather than reselection.

Table A14. Lead of the best response criterion over the best sensitivity baseline, in R_c, at each seed. Positive is the response criterion ahead. The last column is the change across seeds, not a comparison of models.

model	seed 42	seed 43	seed 44	spread
C10-i1	$+0.408$	$+0.413$	$+0.422$	0.015
C10-i2	$+0.736$	$+0.751$	$+0.776$	0.040
C10-i3	$+0.213$	$+0.151$	$+0.153$	0.062
C10-r1	$+0.409$	$+0.326$	$+0.254$	0.155
C10-r2	$+0.386$	$+0.381$	$+0.304$	0.081
C10-r3	$+0.283$	$+0.278$	$+0.221$	0.062
C100-i1	-0.038	-0.026	-0.028	0.012
C100-i2	$+0.108$	$+0.086$	$+0.099$	0.022
C100-i3	$+0.051$	$+0.044$	$+0.057$	0.013
C100-r1	$+0.229$	$+0.231$	$+0.288$	0.059
C100-r2	$+0.086$	$+0.121$	$+0.016$	0.106
C100-r3	$+0.464$	$+0.293$	$+0.305$	0.172

The response-versus-sensitivity difference retains its sign across seeds for every model. One model consistently favors sensitivity, while another has a positive but small margin comparable to its across-seed variation. The remaining models show a sign-stable response advantage.

Criterion-level variation across seeds. Variation is concentrated in a small high-variance tail. The coupling-aware criterion avoids the largest excursions seen for the diagonal criterion, while their

median variation is similar. This supports the narrower claim in Section 5.3 that coupling can improve robustness without consistently increasing median recovery.

Subset changes and training variation are distinct. Runs that retain the same selected subset show smaller typical variation than runs in which the selected set changes by one coordinate. A near-boundary change in subset membership can therefore contribute more variability than changing only the fine-tuning seed.

Variability from re-estimating the selected subset. Independent calibration runs retain very high response-score rank agreement, but subset membership can change near the selection boundary. Observed changes are limited to a single coordinate and occur preferentially when the score gap at the budget boundary is small. Fine-sample variation can therefore alter membership even when the underlying rankings are stable, so reselection variability is kept distinct from fixed-subset seed replication.

H Conversion damage and branch decomposition

H.1 Conversion damage and full-model normalization. Conversion damage is measured from the digital state to the mapped state, while full-model hardware-aware fine-tuning defines the recovery normalization used for the branch comparison. Parameters outside the released adaptation scope remain frozen. In accuracy terms, recovered fraction is $(\text{post} - \text{conv})/(\text{digital} - \text{conv})$. Larger damage leaves more headroom, which is the coarse form of the regime dependence quantified in Appendix I.

H.2 Branch decomposition. For adaptation scope c, define $R_c^{\text{branch}} = (\mathcal{L}_{\text{conv}} - \mathcal{L}_c)/(\mathcal{L}_{\text{conv}} - \mathcal{L}_{\text{full}})$ using held-out validation cross-entropy.

Across all twelve CIFAR models the ranges are channel-branch 0.72–1.02, token-branch 0.49–0.92, and C_{struct} 0.13–0.65. Channel-mixing adaptation recovers the largest share in every model, whereas the LCT-coordinate ceiling varies most strongly across the model set.

H.3 Channel recovery does not make token mixing dispensable. Channel-mixing parameters absorb most conversion damage, but the substitution test in E.3 shows that token mixing remains functionally specific to the trained model. Replacing only the LCT coordinates with those from a separately trained source model reduces accuracy while all other parameters remain frozen. Matching the substituted quadrature matrix to the recipient matrix in Frobenius norm does not remove the degradation. High channel-branch recoverability does not imply that the token-mixing transform is functionally interchangeable.

Trainable counts are $8{,}454{,}272$ for channel mixing, $13{,}504$ for the token branch, and 48 for the LCT coordinates on CIFAR-100. The branch comparison is not a parameter-efficiency test because the adaptation scopes differ greatly in dimension and channel-branch recovery approaches its normalization ceiling. It establishes the location of compensatory capacity, not efficiency per trainable parameter.

I Capacity, selection, and the domain of validity

I.1 The two normalizations. Capacity, $C_{\text{struct}} = (\mathcal{L}_{\text{conv}} - \mathcal{L}_{\text{struct}})/(\mathcal{L}_{\text{conv}} - \mathcal{L}_{\text{full}})$, asks how much of the recoverable loss the full LCT subspace can absorb. Selection quality, $R_c = (\mathcal{L}_{\text{conv}} - \mathcal{L}_c)/(\mathcal{L}_{\text{conv}} - \mathcal{L}_{\text{struct}})$, asks which coordinates to free once adaptation is already restricted to that subspace. They answer different questions and are reported separately. Across the twelve CIFAR models, measured capacity by group is CIFAR-10-imag 0.46–0.65, CIFAR-10-real 0.35–0.49, CIFAR-100-imag 0.21–0.29, CIFAR-100-real 0.13–0.15.

The capacity spread is not attributed to depth because dataset and depth vary together in these experiments. CIFAR-10 (depth 8) and CIFAR-100 (depth 16) differ in task as well as in depth, so the two are confounded and the present experiments do not separate those effects.

I.2 Observed-label inflation of the field second moment. The mapped-to-pre-mapping trace ratio of the observed-label field second moment ranges from $3.4\times$ to $71.7\times$ across models and tracks mapping error rate, damage magnitude, and underprediction. Because $Q_{\mathcal{M}} \propto g^\top \mathcal{M}^{-1} g$, inflation

deflates the score, and the inflation is largest where misclassification is heaviest. Sampling labels from the model separates the label-rule effect from the state-dependent scale of the field second moment. The sampled-label metric still increases by about threefold from pre-mapping to mapped, while observed labels add a further factor of about three. The latter excess scales with damage ($+0.82$) more steeply than the sampled-label component. I.3 tests this explanation by changing the second-moment construction while holding the observed-label forcing fixed. The $gg^\top$ component is separately bounded in G.4 and is too small to account for it.

I.3 Label-rule control with observed forcing. The recomputation in I.2 requires one additional control. Sampling $y \sim p(y \mid x)$ changes *two* quantities. It changes the label rule used in the field second moment and it strongly attenuates the mean field. Because the forward values and backward proxy do not derive from one scalar objective, the score-function zero-mean identity does not hold exactly for this field. Measured directly, $\|g_{\text{sampled}}\|/\|g_{\text{observed}}\|$ is 0.032 to 0.956 with median 0.15. Deployment forcing is a property of the loss actually being minimized and is not a modelling choice. The label convention used to construct the second moment is the quantity being tested. The informative variant keeps g at the observed labels and changes only $\mathcal{M}$.

Table A15. Absolute calibration under three label conventions. Twelve CIFAR models. Predicted is $Q(\text{all LCT coordinates})/(\mathcal{L}_{\text{conv}} - \mathcal{L}_{\text{full}})$ and realized is C_{struct}. Intervals resample models ($B = 5000$).

construction	realized / predicted	ρ(damage, ratio)	cross-model r(predicted, realized)
observed-label $\mathcal{M}$ (deployed)	0.46–4.07	$+0.678$ $[+0.125, +0.950]$	-0.032 $[-0.757, +0.579]$
sampled-label metric *and* forcing	9.4–171.1	$+0.860$ $[+0.540, +0.978]$	$+0.209$ $[-0.423, +0.744]$
sampled-label metric, observed forcing	0.40–1.63	$+0.035$ $[-0.542, +0.634]$	$+0.741$ $[+0.309, +0.910]$

Replacing only the second-moment construction moves the calibration ratio toward one, weakens its dependence on damage, and produces a positive cross-model relation whose interval excludes zero. Within-model ranking is preserved at aggregated within-model $\rho = +0.765$ against the deployed $+0.770$ (Table A9). The fully sampled row is reported only to show why it is unsuitable for calibration. Its ratios are two orders of magnitude off because the sampled-label construction strongly attenuates the forcing, so it is not a fair recovery predictor and is not used as a calibration estimate. Table A15 uses one label draw per model. I.4 provides the replicated analysis used for the main-text claim.

Sampled-label estimator. The estimator performs one hardware-conditioned forward for each example and read-noise draw, samples y from the softmax of *those same* logits, and differentiates through the same straight-through LCT-coordinate path. There is no second forward pass, so the sampled label and the field share one read-noise realization ξ. Writing $z(x, y, \xi)$ for that adaptation field,

$$\mathcal{M}_{\text{samp}} = \mathbb{E}_x \, \mathbb{E}_\xi \, \mathbb{E}_{y \sim p_\theta(\cdot \mid x, \xi)} \left[z(x, y, \xi) z(x, y, \xi)^\top \right]. \tag{I.1}$$

This is a **sampled-label second moment of the same adaptation field**. The conditioning is matched, so the label rule is the only change in the metric construction.

Relation to the model Fisher. Turning a sampled-label outer product into the model Fisher requires the differentiated quantity itself to be $\nabla_p \log p_\theta(y \mid x)$. The field is routed through the straight-through proxy and differs from the gradient of the loss it evaluates by a median relative error of 0.30 (G.3). This comparison isolates label-rule dependence of the adaptation-field second moment. No Fisher or curvature interpretation follows.

I.4 Replication over independent label draws. The comparison was repeated with 24 label draws across 3 calibration seeds for all twelve models. Figure A8 summarizes the label-convention and calibration comparison. The direction is preserved while the estimated magnitude is smaller. Figure 5(b) and Section 5.3 use this replicated analysis. Table A15 retains the single-draw comparison because it includes all three estimator constructions.

Table A16. Replicated calibration, 24 draws × 3 calibration seeds per model. Twelve CIFAR models. Intervals resample models ($B = 5000$).

construction	realized/predicted	spread	ρ(damage, ratio)	cross-model r(pred, realized)
observed-label $\mathcal{M}$	0.45–4.11	9.1×	+0.657 [+0.112, +0.936]	−0.047 [−0.756, +0.579]
sampled metric, observed forcing	0.40–1.64	4.1×	+0.189 [−0.470, +0.693]	+0.680 [+0.219, +0.896]

Changing only the second-moment construction moves the damage dependence from an interval excluding zero to one containing it, and the cross-model relation the other way. Both are weaker than the single draw indicated (+0.035 and +0.741), so the damage dependence is **reduced rather than removed**, and the main text says so. Replicate variability is 5.4–16.1% of the predicted value, and label-draw variation exceeds calibration-seed variation in 11 of 12 models, so the residual uncertainty is in the label sampling rather than in which examples are used.

Better absolute calibration does not imply better subset selection. The sampled metric selects a subset with Jaccard 0.600–0.882 against the deployed one, and whether *that* subset recovers more is a fine-tuning question that has not been run. Section 5.3 continues to use the observed-label metric for selection, and no selection criterion based on the sampled-label metric is evaluated.

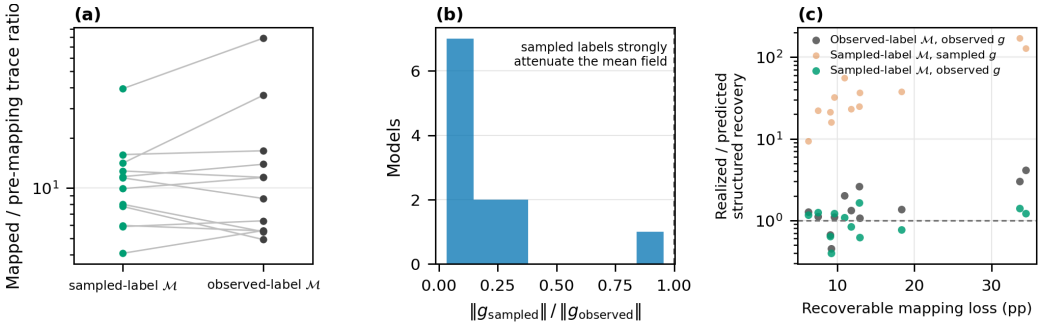

Figure A8: Label convention and absolute calibration on twelve CIFAR models. Panels compare pre-mapping-to-mapped trace inflation, attenuation of the sampled-label mean field, and realized-to-predicted LCT-coordinate recovery under observed-label, fully sampled, and sampled-metric constructions.

J Tiny ImageNet and the larger token grid

Evaluation scope. Tiny ImageNet provides a larger-grid evaluation with 200 classes, 64-pixel input, $N = 256$ tokens on a 16×16 grid, depth 16, and **one trained model per quadrature**. The imaginary model starts from $(\theta, s, k) = (0.25, 2.0, 0)$ at 57.04% and real $(0.25, 1.0, 0)$ at 56.11%. Two models do not support the model-level inference used for the twelve CIFAR models, so the evaluation is excluded from CIFAR model-level statistics.

The Tiny ImageNet protocol uses a shorter hardware-aware fine-tuning schedule because of its higher computational cost and uses the official validation split. Recovery-ceiling evaluations are complete for both quadratures at the same 8-bit ADC and 6-bit DAC operating point. The setting is therefore treated as descriptive rather than as an extension of the CIFAR selection analysis.

Table A17. Tiny ImageNet evaluation with 8-bit ADC and 6-bit DAC. Accuracies are in percent, damage in pp, and ratios are unitless. Two models, one per quadrature, **not comparable to a CIFAR row.**

	imaginary	real
digital baseline	57.04	56.11
converted → full-model	37.03 → 51.28	29.22 → 50.89
conversion damage (digital − converted)	20.01	26.89
recoverable damage (full-model − converted)	14.25	21.67
channel-branch recovery $R_{\mathrm{ch}}^{\mathrm{branch}}$	1.046	1.012
token-branch recovery $R_{\mathrm{tok}}^{\mathrm{branch}}$	0.604	0.851
LCT-subspace capacity C_{struct}	0.277	0.367

The qualitative recovery ordering transfers, not its magnitude. Both quadratures reproduce the CIFAR within-model ordering of channel > token > LCT-coordinate recovery, and fractional order remains the dominant sensitivity axis. Table A17 reports the corresponding values without treating them as directly comparable to CIFAR model-level statistics.

Restricted-subset selection is not evaluated on Tiny ImageNet. The descriptive setting does not provide the replicated model-level units used for the CIFAR selection comparison. Tiny ImageNet is therefore used only to test transfer of the qualitative recovery ordering and quadrature-level differential structure, not to support a separate subset-selection claim.

Quadrature-derivative transfer on the larger grid. At $H = 16$, the finite-grid ratios closely track the asymptotic envelope ($r = +0.980$ for the one-dimensional predictor and $+0.983$ for the separable predictor), while the separable shape factor remains within its analytic bound. The quadrature-level structure of E.1 therefore transfers to the larger token grid without recalibration.