GAUGE-EQUIVARIANT CONDUCTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Internal attribution explains predictions by assigning credit to hidden units. Yet a network function does not determine hidden coordinates: a compensated invertible affine change can preserve outputs while changing coordinate scores. We formulate this as an identifiability problem for function-level interventions. For conductance, we derive the similarity law under the full affine gauge and show that, for nonscalar conductance matrices, every trace-compatible coordinate attribution vector is attainable under a function-preserving gauge; completeness therefore constrains only the trace. We replace coordinate credit with isolated spectral subspaces defined by Riesz projectors, whose interventions commute with the hidden gauge, and evaluate this criterion with a frozen protocol across vision and language modalities. Specifically, gauges change conventional coordinate winners, while spectral scores remain invariant and dominant clusters yield larger target-conditioned effects than matched-rank comparators; they also beat the entire complement on a text holdout. Attack paths show conditional magnitude but mixed direction, with near-modulus ties failing the isolation gate. Thus, spectral subspaces provide function-compatible intervention targets, while coordinate conductance remains parameterization-relative. The code is available at an anonymous repository.

1 INTRODUCTION

Attribution asks which part of a model deserves credit for a prediction. Input coordinates are externally fixed, but a hidden layer is a representation space: an invertible change of hidden coordinates can be canceled downstream without changing any input–output behavior. Internal attribution nevertheless often treats a named neuron as functionally intrinsic. Conductance makes this concrete. It extends Integrated Gradients to hidden neurons, satisfies completeness, and supports ranking, pruning, and mechanistic narratives (Sundararajan et al., 2017; Dhamdhere et al., 2018; Lundstrom et al., 2022); yet completeness fixes only their sum, not whether any individual score belongs to the represented function. That distinction matters whenever a claim uses neuron identity, sign, ordering, or sparsity. Figure 1 visualizes this ambiguity: function-preserving hidden-basis changes move coordinate credit while preserving its total. Thus, in general, coordinate conductance is not a function-level object under the full affine gauge; we therefore seek a gauge-equivariant intervention target.

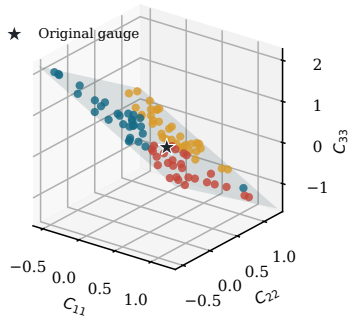


Figure 1: For a nonscalar conductance matrix, coordinate conductance is not determined by the represented function. The axes are its diagonal entries C_{11}, C_{22}, C_{33}; function-preserving hidden-basis changes move between points with the same total $C_{11} + C_{22} + C_{33}$. The star marks the original gauge.

We expose the exact algebra behind this ambiguity. For a hidden path, each coordinate score is one diagonal entry of a conductance matrix. Changing hidden coordinates and compensating downstream leaves model outputs unchanged, but can change these diagonal entries while preserving their sum. Fillmore's theorem (Fillmore, 1969) says that, for a nonscalar matrix, every coordinate attribution vector with the same total can be reached by such a function-preserving change. When the completeness total is nonzero, a coordinate that starts at the bottom can become uniquely top-ranked and absorb essentially all absolute attribution mass while logits remain unchanged.

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 negative result also tells us what a function-level intervention target must preserve: it should transform with the hidden representation rather than depend on its coordinate axes. Similarity does not preserve coordinate axes, but it carries isolated spectral subspaces equivariantly. We therefore use a Riesz projector (Kato, 1995) to select such a subspace. Removing the displacement in that subspace defines a gauge-compatible intervention, including for non-normal matrices when the cluster is isolated. Figure 2[1] depicts this replacement as an equivariant projector and projected displacement.

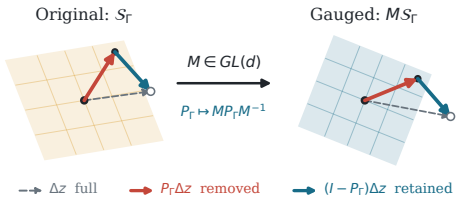


Figure 2: Gauge-equivariant spectral intervention. Gray arrows show the original displacement Δz. Red and blue arrows show the removed $P\Delta z$ and retained $(I - P)\Delta z$ projections. The right panel is the M-transformed decomposition.

This construction gives a concrete methodological criterion: an attribution intended to support a function-level intervention should be gauge-equivariant, and the induced intervention should commute with the gauge. The spectral projector meets that algebraic criterion; the experiments test its numerical stability, output effect, cross-modal replication of the protocol, and behavior on attack-generated endpoints. We evaluate this criterion under a frozen protocol, with a matched-rank comparator and separate tests of gauge validity, intervention utility, and adversarial endpoint stress. In summary, our contributions are as follows:

- Conceptually, we formulate internal attribution as identifiability under the full affine gauge of a hidden representation, separating useful parameterization-relative scores from credit intrinsic to the represented network function, clarifying when internal scores can support function-level claims.
- Theoretically, we prove the nonlinear conductance similarity law and combine it with the prescribed-diagonal theorem to characterize the coordinate orbit: outside the scalar exception, completeness constrains only the trace, thereby ruling out function-intrinsic coordinate rankings.
- Technically, we introduce Riesz-projector interventions that commute with invertible basis changes, and make them auditable through cluster, eigengap, projector-norm, collision, and conditioning diagnostics, with theoretical coverage of non-normal matrices for isolated spectral clusters.
- Empirically, we cover 27 vision and 9 text-environment cells, relocation, 2,592 gauges, conditioning, rank-matched/complement holdouts, and attack-path tests across architectures and non-saturated budgets; magnitude and signed direction are estimands, while isolation is a validity diagnostic.

2 Related Work

Internal attribution and feature interpretation. Integrated Gradients attributes a prediction difference along a baseline-to-input path and motivates implementation invariance for fixed input features (Sundararajan et al., 2017); related methods propagate relevance, localize gradients, or fit local surrogates (Bach et al., 2015; Shrikumar et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017; Ancona et al., 2018). Conductance applies the chain rule along the same path to hidden neurons, with later work studying its axioms and extensions (Dhamdhere et al., 2018; Lundstrom et al., 2022). These guarantees do not imply invariant hidden-coordinate attributions under basis changes. We instead ask whether conductance's coordinate decomposition is identifiable when a compensated affine map is inserted at a hidden cut.

Unit and concept interpretation. Interpretability often studies individual neurons or directions through activation maximization, unit selection, probing, ablation, or concept alignment. Network Dissection aligns units with labeled concepts, while TCAV tests concept directions rather than single units (Bau et al., 2017; Kim et al., 2018); related work studies concept discovery, feature visualization, ablation, and superposition (Ghorbani et al., 2019; Koh et al., 2020; Olah et al., 2017; Bau et al., 2020; Elhage et al., 2022). Sanity checks, removal and faithfulness benchmarks, infidelity analyses, and alternative Shapley formulations show that interpretation depends on estimator, perturbation, and reference choices (Adebayo et al., 2018; Kindermans et al., 2019; Hooker et al., 2019; Yeh et al., 2019; Sundararajan & Najmi, 2020; Edin et al., 2025). Recent work also shows that influence-based attributions can be systematically altered by small model perturbations (Yadav et al., 2026),

[1]Table 7 lists the notation used throughout the paper.

reinforcing the need to state which transformations an attribution claim is meant to survive. Sparse dictionary learning adds an explicit sparsity prior to select a preferred feature basis (Huben et al., 2024); this is a gauge-fixing assumption, not a proof that the selected coordinate is function-intrinsic. Our theorem instead holds the represented function exactly fixed and varies only its hidden basis, so unit-level tools remain useful as parameterization-relative analyses.

Adversarial robustness and endpoint geometry. Adversarial robustness work uses optimization-generated endpoint perturbations to test model behavior, including projected-gradient procedures and strong parameter-free attack suites (Madry et al., 2018; Croce & Hein, 2020). We use these endpoints as controlled stress tests rather than as a robustness benchmark; recent evaluation work emphasizes that local adversarial statistics need not summarize global robustness (Li et al., 2024).

Representation similarity and identifiability. Representation-similarity and identifiability work compares activation spaces or recoverability under restricted transformation classes (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019; Ding et al., 2021; Williams et al., 2021; Roeder et al., 2021). Permutation and broader parameter symmetries also affect model comparison and merging (Entezari et al., 2022; Ainsworth et al., 2023; Godfrey et al., 2022), while other attribution results study rescaling or translation invariance (Xu et al., 2020; Lundstrom et al., 2022). Related identifiability theory for unsupervised concept extraction treats guarantees for latent concepts as a prerequisite for downstream steering and unlearning (Squires, 2026). These lines either compare representations or impose restricted equivalences. Our cut-level functional-equivalence construction permits arbitrary $M \in GL(d)$, including dense basis changes, without requiring M and M^{-1} to be absorbable into the original layer types; we derive its action on a path-integrated attribution matrix and distinguish invariant scalars from equivariant subspaces that can support interventions.

Causal interventions and algebraic scope. Causal abstraction and mediation approaches use internal interventions to connect neural states with higher-level variables (Pearl, 2009; Geiger et al., 2021; Vig et al., 2020; Meng et al., 2022). Our criterion is algebraic rather than semantic: changing coordinates and then intervening must agree with intervening and then changing coordinates, which excludes conclusions created solely by the hidden basis without claiming a causal model. Fillmore's prescribed-diagonal theorem is classical (Fillmore, 1969); we apply it to conductance, derive the nonlinear gauge law, and construct and validate a Riesz-projector intervention on its similarity orbit.

3 METHODOLOGY

Figure 3 traces the matrix, gauge, and equivariant-intervention steps tested across modalities. Table 7 provides a compact reference for the notation and terminology used throughout the paper.

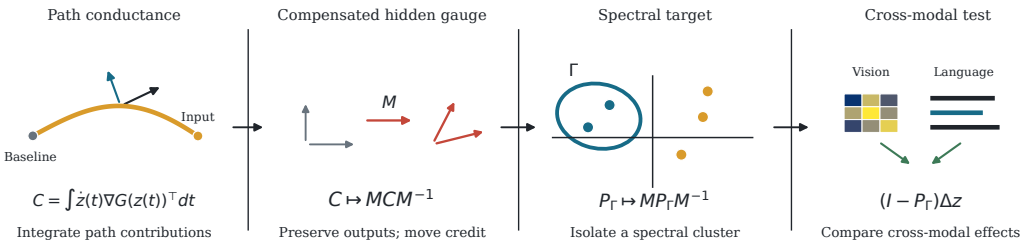


Figure 3: Method overview. Path integration constructs the conductance matrix; a compensated hidden-basis change preserves the network function but moves coordinate credit; an isolated spectral cluster yields an equivariant projector; and the resulting intervention is tested with the same rule in vision and language models.

3.1 CONDUCTANCE IS A MATRIX DIAGONAL

Let a scalar network output factor at a hidden cut be $F = G \circ H$, with $\mathcal{X} \subseteq \mathbb{R}^m$, $H : \mathcal{X} \to \mathbb{R}^d$, and $G : \mathbb{R}^d \to \mathbb{R}$. For a piecewise continuously differentiable input path $\gamma : [0, 1] \to \mathcal{X}$, write

$z(t) = H(\gamma(t))$. Using column-vector gradients, define the conductance matrix

$$C(\gamma; H, G) = \int_0^1 \dot{z}(t) \nabla G(z(t))^\top \, dt. \tag{1}$$

The conventional conductance of hidden coordinate i is $\mathrm{Cond}_i = C_{ii}$. Summing coordinates gives

$$\sum_{i=1}^{d} \mathrm{Cond}_i = \mathrm{tr}(C) = \int_0^1 \nabla G(z(t))^\top \dot{z}(t) \, dt = F(\gamma(1)) - F(\gamma(0)). \tag{2}$$

Thus Eq. 2 is the trace identity for C. Equation 1 retains cross-coordinate transport terms that the conventional diagonal discards. Figure 1 visualizes this fixed-trace coordinate orbit.

3.2 The affine hidden gauge

For $M \in GL(d)$ and $c \in \mathbb{R}^d$, define

$$\widetilde{H}(x) = MH(x) + c, \qquad \widetilde{G}(\widetilde{z}) = G\big(M^{-1}(\widetilde{z} - c)\big). \tag{3}$$

The transformed network is pointwise identical: $\widetilde{G} \circ \widetilde{H} = G \circ H$. This is not an approximation, retraining procedure, or distributional equivalence; it preserves every network output on every input. It is a function-equivalence construction: unless the adjacent layers can absorb M, M^{-1}, and c, it is an equivalent implementation with an inserted affine interface rather than a reparameterization inside the original architecture.

Theorem 1 (Conductance similarity law)**.** *Assume that z is absolutely continuous and that G is locally Lipschitz near the path, with the pathwise chain rule holding almost everywhere; C^1 smoothness is a sufficient special case. Under the compensated affine gauge in Eq. 3,*

$$\widetilde{C} = MCM^{-1}. \tag{4}$$

Proof. Translation vanishes on differentiation, so $\dot{\widetilde{z}} = M\dot{z}$. The a.e. chain rule gives $\nabla_{\widetilde{z}} \widetilde{G} = M^{-\top} \nabla_z G$ wherever the derivatives exist; the exceptional null set does not change the integral. Substitution into Eq. 1 yields $\widetilde{C} = \int M\dot{z} \nabla G(z)^\top M^{-1} dt = MCM^{-1}$. □

The theorem applies to cuts whose pathwise chain rule holds, including the evaluated ReLU paths; for vector outputs, fix a scalar logit, margin, or loss, since target choice is part of the estimand.

3.3 Complete coordinate controllability

Similarity preserves trace, determinant, characteristic polynomial, Jordan structure, and traces of powers. It does not preserve the diagonal.

Theorem 2 (Complete coordinate controllability)**.** *Let $C \in \mathbb{R}^{d \times d}$ be nonscalar. For every $a \in \mathbb{R}^d$ satisfying $\mathbf{1}^\top a = \mathrm{tr}(C)$, there exists $M \in GL(d)$ such that*

$$\mathrm{diag}(MCM^{-1}) = a. \tag{5}$$

If $C = \lambda I$, its diagonal is fixed under every similarity.

Proof. The nonscalar statement is Fillmore's prescribed-diagonal theorem applied to C (Fillmore, 1969). The exception follows immediately from $M(\lambda I)M^{-1} = \lambda I$. □

Corollary 3 (Coordinate claims are not identified)**.** *For nonscalar C, the maximally credited coordinate, coordinate ranking, sign pattern subject to the trace, and coordinate sparsity can all be changed by an exact function-preserving gauge.*

This result is stronger than exhibiting one counterexample. The attainable set in Eq. 5 is the entire affine hyperplane defined by completeness. Declaring a smaller gauge group can restore coordinate structure, but that declaration is an additional modeling assumption. For example, an orthogonal gauge preserves Euclidean geometry but still mixes axes; a permutation gauge preserves the multiset of scores but not neuron names.

3.4 GAUGE-IDENTIFIED SPECTRAL SUBSPACES

To replace coordinate credit, we seek an object that transforms with the representation. Let Λ be a separated cluster of eigenvalues of C and let Γ be a positively oriented contour enclosing only that cluster. Its Riesz projector (Kato, 1995) is defined by Eq. 6:

$$P_\Lambda(C) = \frac{1}{2\pi i} \oint_\Gamma (\xi I - C)^{-1} \, d\xi. \tag{6}$$

This definition remains valid for non-normal matrices and repeated eigenvalues, provided the selected cluster is isolated.

Proposition 4 (Projector equivariance)**.** *For every invertible M,*

$$P_\Lambda(MCM^{-1}) = MP_\Lambda(C)M^{-1}. \tag{7}$$

Proof. The relation in Eq. 7 follows because the resolvent satisfies $(\xi I - MCM^{-1})^{-1} = M(\xi I - C)^{-1}M^{-1}$. Integrating around the same spectral cluster proves the claim. □

Given the endpoint displacement $\Delta z = z(1) - z(0)$, write $P = P_\Lambda(C)$ for a selected cluster and remove its selected spectral projection via

$$\mathcal{I}_P(\Delta z) = (I - P)\Delta z. \tag{8}$$

The corresponding intervened endpoint is $z(0) + \mathcal{I}_P(\Delta z)$. In the transformed basis, $(I - MPM^{-1})M\Delta z = M(I - P)\Delta z$. After mapping back, the intervened hidden state and downstream output are identical in every gauge. Figure 2 depicts commutation. Coordinate ablation fails this commutation relation because a coordinate axis does not generally map to a coordinate axis.

3.5 SELECTION RULE AND NUMERICAL DIAGNOSTICS

Equivariance alone does not choose among spectral clusters: the theory permits any isolated cluster. The protocol takes the largest-modulus candidate as dominant and records frozen isolation diagnostics; utility results require the gate, whereas the attack grid retains near-tie cells as flagged boundary cases. Its utility comparator is the largest remaining isolated cluster of the dominant rank (the same-rank runner-up by modulus). For a real-valued intervention, the selected cluster must be closed under complex conjugation; the implementation enforces this by grouping a non-real eigenvalue with its numerically matched conjugate. It computes the selected projector from left and right eigenvectors, so the evaluated protocol is restricted to simple, numerically diagonalizable eigenvalues or conjugate pairs; the broader Riesz-projector construction is not claimed to be numerically covered for defective matrices or unresolved modulus ties. We track relative eigengap, nearest modulus-tie gap, projector operator norm, covariance error after pullback, and target-logit intervention discrepancy. Utility holdouts enforce these thresholds for both clusters; near-tie attack cells remain flagged boundary cases rather than passing the gate. This is a retained-data criterion rather than a general collision-handling algorithm; a production implementation near a collision should use a Schur- or resolvent-based projector, enlarge the cluster, or abstain. For a selected projector P, intervention utility is measured by the absolute target-logit change between the original endpoint and the endpoint after Eq. 8. This estimand tests whether the chosen invariant subspace carries output effect. It does not assert semantic alignment, human interpretability, unique causal status, or global optimality.

3.6 ADVERSARIAL ENDPOINTS AS A SPECTRAL STRESS TEST

For a clean input x_0, an attack algorithm A, a budget ε, and its perturbed endpoint $x_1^{A,\varepsilon}$, we use the clean-to-perturbed path

$$\gamma_{A,\varepsilon}(t) = x_0 + t\left(x_1^{A,\varepsilon} - x_0\right). \tag{9}$$

The attack-path conductance matrix is Eq. 1 evaluated on this path, with a fixed scalar target G_y given by the true-class logit. Write $z_0 = H(x_0)$, $z_1^{A,\varepsilon} = H(x_1^{A,\varepsilon})$, and $\Delta z_{A,\varepsilon} = z_1^{A,\varepsilon} - z_0$. For its dominant spectral projector $P_{A,\varepsilon}$, we define the restored target-logit effect

$$R_{A,\varepsilon}(x_0; y) = G_y(z_0 + (I - P_{A,\varepsilon})\Delta z_{A,\varepsilon}) - G_y(z_1^{A,\varepsilon}), \tag{10}$$

It is the target-logit change obtained by removing the selected spectral displacement from the attacked endpoint. The primary magnitude estimand is $A_{A,\varepsilon} = |R_{A,\varepsilon}|$; signed $R_{A,\varepsilon}$ is retained as a directional diagnostic.

Proposition 5 (Attack-path equivariance)**.** *For any fixed endpoint produced by an input-space endpoint generator, including PGD and APGD (Madry et al., 2018; Croce & Hein, 2020), the matrix on Eq. 9 obeys the same similarity law as Eq. 4; hence $P_{A,\varepsilon} \mapsto M P_{A,\varepsilon} M^{-1}$ and $R_{A,\varepsilon}$ is unchanged by the compensated hidden gauge.*

Proof. The hidden image of Eq. 9 is an admissible path, so the conductance similarity law applies directly. The displacement transforms as $\Delta\widetilde{z} = M\Delta z$, and therefore $(I - MPM^{-1})M\Delta z = M(I - P)\Delta z$ before applying the compensated downstream map. □

The endpoint target-logit drop $D_{A,\varepsilon} = G_y(z_0) - G_y(z_1^{A,\varepsilon})$ is reported descriptively. We do not divide by $D_{A,\varepsilon}$ as a primary metric because matched random paths can have a near-zero or sign-changing denominator. For each clean sample and budget, an attack endpoint is paired with a box-feasible magnitude-permutation/sign control from the same clean image: its displacement magnitudes are randomly reassigned subject to image-box feasibility, with feasible signs randomized, preserving the attack endpoint's ℓ_2 and ℓ_∞ scales. This is a structured randomized control, not a uniform draw from the feasible box-constrained shell. The attack-success AUC is computed only within a frozen case-control sample containing equal attack-success and attack-failure strata; it is a conditional separation statistic, not a population attack rate or calibrated vulnerability probability. We report both the AUC of $A_{A,\varepsilon}$ and the signed AUC of $R_{A,\varepsilon}$, because a larger output-relevant magnitude need not have a common direction across architectures or attacks. The attack algorithm selects the endpoint, while the conductance matrix and spectral intervention diagnose its hidden displacement; no attack optimization is performed in representation space. This separation makes the endpoint grid a stress test of the proposed equivariant diagnostic: a positive result would show that the same subspace object used for ordinary intervention also resolves structured attack-generated displacement, while a failure would delimit its scope rather than falsify the gauge theorem.

4 Experiments

4.1 Questions and frozen protocol

Experimental questions. We organize the experiments around five questions. **Q1:** Does the implementation preserve the network and obey the exact similarity and projector laws? **Q2:** Does conventional hidden-coordinate conductance expose consequential non-identifiability rather than a formal curiosity? **Q3:** Are spectral interventions stable across architectures, datasets, seeds, and conditioned dense gauges? **Q4:** Does the selected invariant subspace carry more target-logit effect than a matched-rank invariant alternative? **Q5:** Across architectures, attack algorithms, and non-saturated budgets, does the magnitude $|R|$ of the dominant restored effect exceed the matched-random endpoint effect while preserving gauge covariance, and is its signed direction consistent?

Vision matrix. The balanced matrix uses ResNet-18 (He et al., 2016), WRN-28-2 (Zagoruyko & Komodakis, 2016), and our PlainConvNet architecture control on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011). Three independently trained seeds produce 27 architecture–dataset–seed cells and 324 correctly classified input paths. Paths run from a fixed baseline to the input; conductance is evaluated at nonlinear cuts with true-label targets and clean-correct filtering. Each path receives condition-bounded gauges, totaling 2592 trials. Thresholds, model checkpoints, path identities, cluster rules, and seeds are stored in frozen configs.

Utility holdouts. The 48-path development utility study fixes the dominant-versus-runner-up comparison. Five subsequent holdouts contain 24 paths each: ResNet-18/CIFAR-10 seed 2027, WRN-28-2/CIFAR-10 seed 2030, PlainConvNet/CIFAR-10 seed 2030, ResNet-18/CIFAR-100 seed 2030, and ResNet-18/SVHN seed 2030. The sampled paths are retained only when both compared projectors satisfy the frozen numerical-isolation policy; all sampled paths and exclusions remain recorded in the artifacts. No threshold or selection rule is fitted after observing those outcomes; the runner-up comparator and spectral-isolation protocol are frozen in advance.

Text replication. The text replication uses a two-layer, four-head Transformer (Vaswani et al., 2017) trained from scratch on AG News (Zhang et al., 2015), with a 128-dimensional CLS cut and a nonlinear GELU MLP head (Hendrycks & Gimpel, 2016). Conductance follows a null-token embedding baseline rather than an interpolation of discrete token IDs. A seed-2026 pilot freezes the protocol; seeds 2027–2029 are untouched holdouts under clean text, deterministic word dropout, and deterministic character deletion, yielding the nine cells and 144 paths summarized in Table 4.

Evidence hierarchy. These evidence blocks are deliberately non-substitutable: Q1 and Q3 test exact gauge implementation and numerical equivariance, Q2 realizes the conventional-coordinate negative control, Q4 tests paired target-conditioned utility, and Q5 tests conditional endpoint-magnitude separation. Gauge draws validate covariance but are not independent utility replicates; paths are the experimental units, while training seeds and architecture–dataset or text-environment cells are replication axes.

Table 1: Frozen gauge audit. Spectral residuals test exact identities; coordinate variation is a gauge-sensitive negative control, not a utility score.

Quantity	Result
Architectures / datasets / seeds	$3/3/3$
Cells / paths / gauges	27/324/2592
Maximum projector residual	1.16e-11
Maximum spectral-score residual	3.6e-06
Path-cluster bootstrap LB	98.78%
Balanced-matrix maximum coordinate range	22.71 logits
Relocation success	$3/3$ architectures
Utility / complement wins	$5/5$ cells; $72/72$ paths

4.2 Q1: EXACT GAUGE VALIDITY

Table 1 summarizes the implementation audit. The general nonlinear check attains similarity-law, diagonal, and trace errors of 1.03×10^{-13}, 5.33×10^{-14}, and 4.62×10^{-14}; a separate nine-cell holdout preserves logits to 4.04×10^{-13} and complete conductance totals to relative error 2.54×10^{-14}. Across the balanced matrix, projector covariance and spectral-score errors remain at 1.16e-11 and 3.6e-06, respectively. These are implementation checks against exact identities, not empirical approximations to the theorem. The direct relocation pilot uses penultimate linear heads for transparent prescribed-diagonal constructions; separate probes at nonlinear cuts on all three architectures recover full-rank conductance with nontrivial power-trace structure, with worst completeness error 9.74×10^{-3}. We treat those probes as a scope check rather than utility evidence; the balanced matrix supplies the nonlinear-cut gauge and projector tests.

Table 2: Representative direct coordinate-credit relocation pilot on CIFAR-10. The targeted bottom-decile coordinate becomes the unique credit winner while logits remain unchanged; mass is the gauged top-coordinate absolute attribution fraction, and effect is the selected top-coordinate ablation change.

Architecture	d	$\kappa(M)$	Logit res.	Original mass	Gauged mass	Effect
ResNet-18	512	115.6	1.9e-13	3.67%	$> 99.99\%$	0.35→9.44
WRN-28-2	128	31.0	2.8e-14	24.54%	$> 99.99\%$	3.23→13.16
PlainConvNet	256	47.4	9.6e-14	5.52%	$> 99.99\%$	0.92→16.70

4.3 Q2: COORDINATE CREDIT CAN BE MADE CONSEQUENTIALLY WRONG

The controllability theorem predicts more than ranking noise; conventional hidden-coordinate conductance is the gauge-sensitive negative control here. Table 2 reports a CIFAR-10 construction in which a bottom-decile coordinate becomes the top coordinate in all three architectures while the maximum logit difference stays below 2×10^{-13}, captures more than $1 - 2.1 \times 10^{-13}$ of absolute attribution mass, and changes the selected top-coordinate ablation effect from $0.35 \to 9.44$, $3.23 \to 13.16$, and $0.92 \to 16.70$ logits for ResNet-18, WRN-28-2, and PlainConvNet. This realizes the theorem's worst case rather than claiming that training discovers arbitrary gauges; independently sampled gauges still give a path-cluster bootstrap lower bound

Table 3: Disaggregated balanced matrix over three training seeds. Top-change is the seed-wise minimum; range and error are worst cases. Each row contains 36 paths and 288 gauges.

Architecture	Data	Top-change	Coord. range	Proj. error
ResNet-18	CIFAR-10	99.0%	6.20	2.5e-13
ResNet-18	CIFAR-100	100.0%	6.76	1.2e-11
ResNet-18	SVHN	100.0%	3.71	2.4e-13
WRN-28-2	CIFAR-10	97.9%	10.93	2.9e-13
WRN-28-2	CIFAR-100	97.9%	21.16	1.7e-13
WRN-28-2	SVHN	99.0%	5.64	1.4e-13
PlainConvNet	CIFAR-10	97.9%	22.71	6.3e-13
PlainConvNet	CIFAR-100	99.0%	21.38	1.1e-13
PlainConvNet	SVHN	96.9%	14.92	8.0e-14

of 98.78% for winner changes and up to 11.08 logits of intervention variation across five holdouts (the representative pilot is in Table 2; the untouched-seed holdout is in Appendix Table 9).

4.4 Q3: SPECTRAL INTERVENTIONS SURVIVE BROAD GAUGE VARIATION

Vision matrix. Every one of the 27 balanced vision cells passes the frozen covariance and score criteria across residual, wide-residual, and plain convolutional models on CIFAR-10, CIFAR-100, and SVHN; Table 3 reports the disaggregated seed results rather than a pooled average. The conventional coordinate negative control changes winners broadly while projector covariance remains far below tolerance, and the path-cluster bootstrap—not repeated gauge draws—is the primary uncertainty summary.

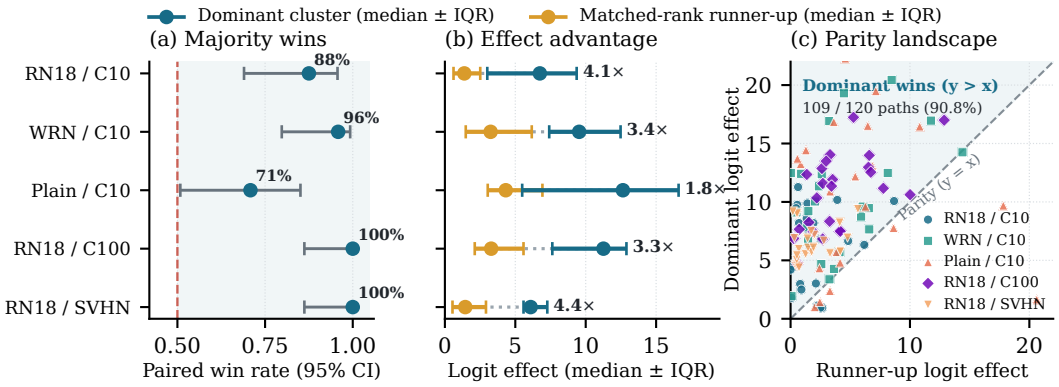


Figure 4: Frozen utility holdouts across five settings (120 paths). (a) Paired-win fraction of the dominant cluster with 95% Wilson confidence intervals against chance (0.5). (b) Absolute true-class-target effects: circles denote medians, horizontal bars denote interquartile ranges (IQR, 25th–75th percentiles), and numbers show median paired effect ratios. (c) Path-level parity landscape ($n = 120$); points above the dashed $y = x$ parity line denote dominant-cluster wins (109/120 paths). Next-class control results are reported in the text.

Text modality. The same gauge/intervention rule is applied to a from-scratch two-layer Transformer on AG News: across 9 seed–environment cells and 144 paths, accuracy remains at least 73.60%, conventional coordinate winners change at least 97.92%, projector covariance is at most 1.64e-14, and the dominant cluster wins the paired absolute target-logit comparison against the matched-rank runner-up on every path. On the fresh seed-2029 text holdout it also beats the entire spectral complement on 72 of 72 paths, with median effects 2.21 versus 0.004 logits; Table 4 reports the seed-level summary and the full environment breakdown is in Appendix Table 16.

Table 4: Seed-level AG News gauge audit in the main text. Acc. is clean accuracy; Coord. is coordinate-winner change; RU win is the dominant-cluster win rate against the matched-rank runner-up.

Seed	Paths	Acc.	Coord.	Proj. err.	RU win
2027	36	73.70%	97.9%	8.8e-15	100.0%
2028	36	73.60%	97.9%	1.6e-14	100.0%
2029	72	73.85%	97.9%	1.5e-14	100.0%

Numerical stress checks. Conditioning stress with $\kappa(M) \leq 100$ still changes every coordinate winner while spectral scores remain identical at reported precision; the minimum dominant eigengap is 0.243 and maximum projector norm is 21.75 (Appendix Table 12). Against a finer quadrature reference, the production integral has maximum relative matrix error 1.18e-03 and maximum intervention-effect error 3.99e-03, separating numerical approximation from the exact gauge claims.

Table 5: Complete frozen vision utility summary. LB is the 95% Wilson lower bound; D/RU gives median dominant/runner-up absolute effects for the true-class target; ratio is the median paired effect ratio.

Setting	Wins/n	LB	D/RU	Ratio
RN18/C10/2027	21/24	69.0%	6.74/1.38	4.07
WRN/C10/2030	23/24	79.8%	9.53/3.25	3.37
Plain/C10/2030	17/24	50.8%	12.64/4.32	1.85
RN18/C100/2030	24/24	86.2%	11.26/3.29	3.35
RN18/SVHN/2030	24/24	86.2%	6.09/1.45	4.37

4.5 Q4: INVARIANT DOES NOT MEAN VACUOUS

This conditional utility test compares the frozen dominant rule with a rank-matched invariant comparator; conventional coordinate conductance is retained separately as a gauge-sensitive negative control. On 47 of 48 development paths, the dominant cluster wins 87.23% with Wilson lower bound (Brown et al., 2001) 74.83% and median effect ratio 2.92; true-class effects are 6.25 versus 1.86 logits. The next-class control favors the dominant cluster on 70.21%, but the weakest holdout is only 50.00%, so the result is target-conditioned. All five holdouts retain majority wins after the pre-registered isolation gates: 120 of 120 paths remain, and the weakest cell reaches 70.83% wins with Wilson lower bound 50.83% and median ratio 1.85 (Table 5). The corresponding paired-effect landscape is shown in Figure 4.

4.6 Q5: ADVERSARIAL ENDPOINTS STRESS-TEST THE EQUIVARIANT DIAGNOSTIC

Q5 applies the same dominant spectral rule used in Q4 to attack-generated endpoints, then compares them with box-feasible magnitude-permutation/sign endpoint controls under a frozen case-control protocol. Table 6 reports the core cell-level results. Its primary endpoint estimand is within-cell separation of the absolute restored effect $|R|$, not attack-success rate or calibrated robustness. The grid uses the frozen 129-point path discretization, so its covariance checks apply to that discretized estimator; the finer quadrature audit in Q3 is a separate clean-to-input audit. The grid covers 12 architecture–attack–budget cells and 576 endpoint paths; all cells pass implementation checks, norm matching is within 7.6e-06 / 3.6e-07, and spectral isolation passes 10 cells, with 2 near-modulus ties retained. Attack median $|R|$ exceeds matched random in 12/12 cells, $AUC_{|R|}$ is above chance in 9 cells, and the paired bootstrap lower bound is positive in 9 cells. This supports a conditional magnitude stress-test claim, not a universal attack mechanism: signed AUC is above chance in 7 cells and signed paired difference is positive in 7 cells, so magnitude does not imply one common direction. The complete cell-level bootstrap intervals, signed effects, and near-tie diagnostics are in Appendix Table 13.

Table 6: Core adversarial endpoint grid. $AUC_{|R|}$ separates attack-success from attack-failure strata using the absolute restored effect; A/R gives median $|R|$ for attack and matched-random endpoints; signed AUC is a secondary directional diagnostic; gap is the minimum dominant modulus gap in the cell.

Architecture	Attack	ε	$AUC_{\|R\|}$	A/R $\|R\|$	$\Delta\|R\|$	AUC_R	Gap
ResNet-18	PGD-40	1/255	0.500	0.02/0.01	0.015	0.535	2.9e-03
ResNet-18	PGD-40	2/255	0.681	0.04/0.02	0.016	0.660	1.3e-03
ResNet-18	APGD-CE	1/255	0.556	0.02/0.02	0.012	0.528	3.4e-04
ResNet-18	APGD-CE	2/255	0.590	0.02/0.01	0.019	0.576	1.4e-02
WRN-28-2	PGD-40	1/255	0.674	0.04/0.02	0.013	0.403	6.5e-03
WRN-28-2	PGD-40	2/255	0.451	0.05/0.03	0.025	0.340	2.0e-03
WRN-28-2	APGD-CE	1/255	0.792	0.04/0.02	0.016	0.514	4.8e-04
WRN-28-2	APGD-CE	2/255	0.861	0.02/0.02	0.003	0.500	9.2e-03
PlainConvNet	PGD-40	1/255	0.500	0.11/0.04	0.079	0.583	1.3e-02
PlainConvNet	PGD-40	2/255	0.576	0.18/0.08	0.095	0.583	5.1e-03
PlainConvNet	APGD-CE	1/255	0.694	0.10/0.04	0.064	0.396	4.0e-03
PlainConvNet	APGD-CE	2/255	0.882	0.16/0.08	0.050	0.333	3.4e-03

5 CONCLUSION

Taken together, these results delimit what internal attribution can support under the hidden-basis gauge. Coordinate conductance is a useful parameterization-relative negative control, but cannot support function-level claims about neuron identity, ordering, sign, or sparsity: completeness preserves only the trace and the full coordinate orbit remains attainable. A spectral projector supports a narrower claim: its intervention commutes with the gauge and, under the frozen selection rule, carries more target-logit effect than the matched-rank comparator across the retained vision and text holdouts.

For practice, coordinate scores should be treated as parameterization-relative diagnostics unless the basis is externally privileged or explicitly gauge-fixed. When it is not, the paper's replacement is to report the selected equivariant subspace together with its separation diagnostics and a paired intervention effect; this makes the intervention claim well-defined on the gauge orbit without turning the subspace into a semantic oracle. The empirical scope is conditional: the adversarial grid is a stress test of the same diagnostic, not a robustness benchmark. Its results support a bounded attack-path diagnostic, with mixed directional effects and reduced reliability near spectral ties; they do not establish a universal mechanism, semantic interpretation, causal claim, or robustness predictor.

AI USE STATEMENT

Generative AI tools were used only in an auxiliary capacity for limited drafting and language editing, manuscript organization and formatting, reproducibility documentation, consistency checks of reported artifacts, and editing implementation scripts and scientific figures. They were not used to generate data, formulate the central mathematical claims, develop or prove the theory, design the experiments, or determine the interpretation and conclusions. All AI-assisted text, code, and figure changes were reviewed, tested, and revised by the authors, who take full responsibility for the final manuscript, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs for the gauge and projector results are given in the Methodology and Appendix. The experimental sections specify the models, datasets and preprocessing, baselines, path and intervention definitions, selection rules, thresholds, random seeds, and evaluation metrics; supplementary materials and code include the implementation and analysis scripts, frozen configurations, and input indices needed to reproduce the reported analyses.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001. doi: 10.1214/ss/1009213286.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2206–2216, 2020.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? *arXiv preprint arXiv:1805.12233*, 2018.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in neural information processing systems*, 34:1556–1568, 2021.

Joakim Edin, Andreas Geert Motzfeldt, Casper L. Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. Normalized AOPC: Fixing misleading faithfulness metrics for feature attributions explainability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1730. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.86.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

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

Peter A. Fillmore. On similarity and the diagonal of a matrix. *The American Mathematical Monthly*, 76(2):167–169, 1969. doi: 10.1080/00029890.1969.12000162.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11893–11905, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Tosio Kato. *Perturbation Theory for Linear Operators*, volume 132. Springer, 1995.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Zaitang Li, Pin-Yu Chen, and Tsung-Yi Ho. GREAT Score: Global robustness evaluation of adversarial perturbation using generative models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39158–39189, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pp. 14485–14508. PMLR, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Chandler Squires. A unifying framework for unsupervised concept extraction. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4393–4401. PMLR, 2026.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017.

Lloyd N Trefethen. Spectra and pseudospectra: The behaviour of non-normal matrices and operators. In *The graduate student's guide to numerical analysis' 98: Lecture notes from the VIII EPSRC Summer School in Numerical Analysis*, pp. 217–250. Springer, 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9680–9689, 2020.

Chhavi Yadav, Ruihan Wu, and Kamalika Chaudhuri. Influence attributions can be systematically altered by model manipulation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1396–1404. PMLR, 2026.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657, 2015.

A Appendix

A.1 Notation and terminology

Symbol / term	Meaning
F	Scalar network output factor at the hidden cut, with $F = G \circ H$.
H	Hidden representation map $H : \mathcal{X} \to \mathbb{R}^d$ from the input space.
G	Scalar downstream map that turns the hidden state into the selected target output.
C^1	Continuously differentiable; a sufficient regularity condition for the conductance theorem.
Locally Lipschitz	Weaker regularity of G used in the theorem, together with the standard chain rule almost everywhere along the path.
Absolutely continuous	Regularity of $z(t)$ that permits an almost-everywhere derivative and the matrix path integral.
γ	Input path $\gamma : [0, 1] \to \mathcal{X}$ used for path integration; t is its path parameter.
$z(t)$	Hidden trajectory, defined by $z(t) = H(\gamma(t))$ in a d-dimensional hidden space.
Δz	Endpoint displacement in hidden space, $z(1) - z(0)$.
C	Path-integrated conductance matrix $\int \dot{z}(t) \nabla G(z(t))^\top dt$; i, j index its hidden coordinates.
C_{ij}	(i, j) entry of C; off-diagonal entries retain cross-coordinate transport discarded by coordinate scores.
$\mathrm{diag}(C)$	Vector of diagonal entries of C, i.e., the coordinate conductance scores.
Cond_i	Conventional conductance of coordinate i, defined as $\mathrm{Cond}_i = C_{ii}$.
$\mathrm{tr}(C)$	Sum of coordinate conductances; completeness identifies it with the selected output difference.
M	Invertible linear change of hidden coordinates in the compensated affine gauge.
c	Constant hidden translation paired with M; it does not affect C because differentiation removes it.
$GL(d)$	Group of invertible real $d \times d$ matrices allowed for M.
$\widetilde{\cdot}$	Transformed quantity under the compensated hidden-coordinate gauge, e.g., $\widetilde{C} = MCM^{-1}$.
a	Prescribed diagonal vector in the controllability theorem.
$\mathbf{1}$	All-ones vector in $\mathbf{1}^\top a = \mathrm{tr}(C)$.
Λ	Selected isolated cluster of eigenvalues of C.
Γ	Contour in the complex plane enclosing Λ and no other eigenvalues.
$P_\Lambda(C)$	Riesz projector onto the generalized eigenspace associated with the isolated cluster Λ.
P	Shorthand for the selected projector $P_\Lambda(C)$ in the intervention definition.
I	Identity matrix of the hidden-space dimension.
$\mathcal{S}_\Gamma$	Invariant spectral subspace $\mathrm{im}\, P_\Lambda(C)$ selected by Γ.
$\mathcal{I}_P(\Delta z)$	Projector intervention $(I - P)\Delta z$, which removes the selected spectral component; $\|P\|_2$ is monitored for oblique-projector sensitivity.
$x_1^{A,\varepsilon}$	Endpoint from generator A (such as PGD/APGD) at budget ε, starting from clean input x_0; $\kappa(M)$ is the gauge condition number.
$R_{A,\varepsilon}$	Signed restored target-logit effect of the dominant spectral intervention.
$A_{A,\varepsilon}$	Absolute magnitude estimand, defined by $A_{A,\varepsilon} = \|R_{A,\varepsilon}\|$.
Gauge	Function-preserving affine change of hidden coordinates, possibly implemented at an inserted cut.
Invariant	A scalar or matrix property unchanged by the hidden-coordinate gauge.
Equivariant	An object that transforms covariantly, such as $P \mapsto MPM^{-1}$, so its abstract meaning is preserved.
Estimand	A quantity specified in advance as the target of an analysis, rather than a post hoc summary.

Table 7: Notation and terminology used throughout the paper.

A.2 Additional Theory and Scope

A.2.1 Assumptions and architectural scope

The hidden space is finite-dimensional and real, $z(t) \in \mathbb{R}^d$; the target G is scalar and locally Lipschitz on a neighborhood of the hidden path, differentiable along the path almost everywhere with the standard a.e. chain rule; z is absolutely continuous (piecewise C^1 is sufficient); and the entries of $\dot{z}(t)\nabla G(z(t))^\top$ are integrable. For the evaluated ReLU paths, the path composition is absolutely continuous and piecewise smooth, and the conductance integrand uses a specified almost-everywhere gradient representative. The gauge matrix $M \in GL(d, \mathbb{R})$ is constant along the path. These assumptions are sufficient for the a.e. chain rule, the matrix integral, and the similarity action used above; C^1 smoothness remains a sufficient special case. A vector output must first be reduced to a fixed scalar logit, margin, or loss.

Equation 3 may be implemented by inserting an affine map and its inverse around a cut. It need not be absorbable into the original layer types. When the adjacent layers are affine, it can instead be absorbed into their parameters. The theorem concerns equality of the represented function, not membership of the transformed parameters in a narrower architectural parameterization.

A.2.2 Translations, targets, and paths

Translation does not enter C because $\dot{\widetilde{z}} = M\dot{z}$. The inverse translation in $\widetilde{G}$ restores z before differentiation, so the full affine gauge has the same matrix action as its linear part. The input path itself need not be linear; only piecewise continuous differentiability is required. For vector outputs, fixing a scalar target before constructing C is essential because different logits, margins, or losses generally induce different similarity orbits.

A.2.3 Real prescribed diagonals and the linear-head boundary

The controllability theorem is used over the real field: for real nonscalar C and real a with $\mathbf{1}^\top a = \mathrm{tr}(C)$, the similarity matrix can be chosen in $GL(d, \mathbb{R})$ (Fillmore, 1969). Complex eigenvalues enter only the spectral construction, where conjugate pairs are grouped to retain a real invariant subspace. They are not needed for prescribing the conductance diagonal.

For a linear downstream head $G(z) = w^\top z + b$,

$$C = \Delta z\, w^\top, \qquad \mathrm{rank}(C) \le 1, \qquad \mathrm{tr}(C) = w^\top \Delta z. \tag{11}$$

If this matrix is nonscalar, Fillmore's theorem still realizes every real diagonal with the required sum. Thus a zero output difference permits any zero-sum coordinate-credit vector, including nonzero positive and negative entries. The only fixed case is $C = \lambda I$; for rank-one C and $d > 1$, this reduces to $C = 0$. The targeted relocation experiments use this transparent rank-one boundary, while the nonlinear probes verify that the general matrix law is not confined to it.

A.2.4 Spectral clusters versus eigenvectors

An eigenvector is scale ambiguous, may be complex, and is unstable or undefined as an individual object at repeated eigenvalues. A Riesz projector identifies the entire generalized eigenspace associated with an isolated cluster and is defined by the resolvent even for defective matrices. The numerical experiments use the simpler left/right-eigenvector formula for simple eigenvalues and conjugate pairs, so defective matrices and unresolved modulus ties remain outside the evaluated implementation. If a desired cluster collides with its complement, no continuous rule can preserve the original separation; a robust extension should merge clusters or abstain rather than select an unstable individual eigenvector.

A.2.5 Invariant scalars versus equivariant interventions

Traces of powers, determinant, and characteristic-polynomial coefficients are similarity invariants. They summarize an orbit but do not specify where to intervene. A projector is equivariant rather than invariant: its matrix changes as $P \mapsto MPM^{-1}$, while its image denotes the same abstract subspace. This covariance is what makes Eq. 8 operational.

A.2.6 Endpoint commutation

Let $z_1 = z(1)$, $z_0 = z(0)$, and $z_1^{(-P)} = z_0 + (I - P)(z_1 - z_0)$. In the transformed coordinates, $\widetilde{z}_j = Mz_j + c$, $\widetilde{\Delta} z = M\Delta z$, and $\widetilde{P} = MPM^{-1}$. Therefore

$$\widetilde{z}_1^{(-\widetilde{P})} = \widetilde{z}_0 + (I - \widetilde{P})\widetilde{\Delta} z \tag{12}$$

$$= M\{z_0 + (I - P)\Delta z\} + c = Mz_1^{(-P)} + c. \tag{13}$$

Pullback by $M^{-1}(\cdot - c)$ returns exactly $z_1^{(-P)}$, so the compensated downstream maps produce identical outputs. This proves endpoint equality, not only covariance of a matrix score.

A.2.7 Cluster collisions and non-normal conditioning

The evaluated scripts sort simple eigenvalue or conjugate-pair clusters by modulus and require the aggregate minimum relative separation to exceed the frozen threshold 10^{-3}. They do not dynamically merge all modulus ties or return per-path abstentions; no evaluated path falls below the retained-data threshold. The broader collision-handling rule remains a methodological extension rather than an executed part of this study.

For non-normal C, eigengap alone does not control sensitivity. The resolvent may be large far from the spectrum, and the oblique Riesz projector can have $\|P\|_2 \gg 1$ (Kato, 1995; Trefethen, 2005; Stewart & Sun, 1990). We therefore monitor both relative separation and $\|P\|_2$, together with a direct pullback covariance residual. The selected subspace varies continuously only while the contour remains in the resolvent set; no method can continuously preserve the same split through a cluster collision.

A.3 Complete Experimental Results

A.3.1 Numerical theorem identities

Table 8 reports every residual in the general nonlinear theorem check. These are implementation tests against exact identities, not sampling-based estimates.

Table 8: General conductance-similarity implementation checks.

Identity	Maximum residual	Tolerance
Similarity law	1.03e-13	1.0e-10
Conductance diagonal	5.33e-14	1.0e-10
Completeness / trace	4.62e-14	1.0e-10
Normalized power traces	8.13e-14	1.0e-10

A.3.2 Targeted coordinate-credit relocation

The pilot deliberately selects an originally bottom-decile coordinate and prescribes it as the unique credit winner. The representative three-architecture pilot is reported in Table 2 in the main text. Table 9 repeats direct prescription on two untouched seeds in each architecture–dataset cell.

Table 9: Complete untouched-seed direct-prescription holdout. Logit, attribution, and total residuals are reported separately for every seed; mass is the gauged top-coordinate absolute attribution fraction.

Architecture	Dataset	Seed	$\kappa(M)$	Logit res.	Attr. res.	Gauged mass	Total res.
ResNet-18	CIFAR-10	2029	114.2	4.0e-13	2.3e-13	$> 99.99\%$	2.5e-14
ResNet-18	CIFAR-10	2030	115.2	2.1e-14	1.4e-14	$> 99.99\%$	1.1e-15
WRN-28-2	CIFAR-10	2029	30.3	2.3e-14	2.0e-14	$> 99.99\%$	4.0e-16
WRN-28-2	CIFAR-10	2030	30.0	3.3e-14	1.1e-14	$> 99.99\%$	6.6e-16
PlainConvNet	CIFAR-10	2029	48.6	4.3e-14	5.7e-15	$> 99.99\%$	2.2e-16
PlainConvNet	CIFAR-10	2030	47.3	9.2e-14	1.4e-14	$> 99.99\%$	2.2e-16
ResNet-18	CIFAR-100	2029	54.8	3.1e-14	7.5e-15	$> 99.99\%$	5.8e-16
ResNet-18	CIFAR-100	2030	59.3	2.1e-14	7.1e-15	$> 99.99\%$	3.4e-16
WRN-28-2	CIFAR-100	2029	21.3	4.6e-14	1.8e-14	$> 99.99\%$	1.2e-15
WRN-28-2	CIFAR-100	2030	25.5	3.2e-14	2.1e-14	$> 99.99\%$	8.7e-16
PlainConvNet	CIFAR-100	2029	33.0	5.0e-14	1.1e-14	$> 99.99\%$	6.3e-16
PlainConvNet	CIFAR-100	2030	42.1	5.7e-14	4.6e-15	$> 99.99\%$	1.4e-16
ResNet-18	SVHN	2029	102.9	1.4e-14	7.1e-15	$> 99.99\%$	9.4e-16
ResNet-18	SVHN	2030	109.1	1.4e-14	6.5e-15	$> 99.99\%$	6.0e-16
WRN-28-2	SVHN	2029	32.9	2.1e-14	7.1e-15	$> 99.99\%$	9.1e-16
WRN-28-2	SVHN	2030	34.0	2.3e-14	8.9e-15	$> 99.99\%$	1.1e-15
PlainConvNet	SVHN	2029	59.1	4.3e-14	1.5e-14	$> 99.99\%$	6.8e-16
PlainConvNet	SVHN	2030	59.4	5.0e-14	1.8e-14	$> 99.99\%$	1.4e-15

A.3.3 Nonlinear conductance matrices

Table 10 separates the exact synthetic similarity check from quadrature-based trained-network probes. Every path has rank above one and a nontrivial spectrum. The trained rows report completeness error, whereas the synthetic row reports the maximum invariant residual.

A.3.4 Balanced matrix and conditioning

Figure 5 visualizes the nine architecture–dataset aggregates without pooling away weak cells; Table 3 gives their values. Table 11 exposes all 27 independently trained seed cells underlying that summary. Each row aggregates exactly 12 paths; all 324 path rows and 2,592 gauge trials remain in the named JSON source artifacts. The complete condition-number-100 path-level stress rows are reported in Appendix Table 12, alongside Figure 6 and the protocol details.

Table 10: Nonlinear-cut matrix structure across synthetic and trained paths.

Setting	Paths	Rank range	Nontrivial spectrum	Max. residual
Synthetic nonlinear	24	2	100%	3.2e-14
ResNet-18	8	256–256	100%	8.05e-03
WRN-28-2	8	64–64	100%	9.74e-03
PlainConvNet	8	127–128	100%	2.44e-03

Table 11: Complete seed-level visual gauge matrix. Each cell has 12 paths and 96 dense-gauge trials. Change is the trial-level coordinate-winner change rate; range and residual are path maxima; gap is the minimum dominant relative eigengap and $\|P\|$ the maximum projector norm.

Architecture	Dataset	Seed	Paths	Change	Range	Proj. residual	Min. gap	Max. $\|P\|$
ResNet-18	CIFAR-10	2027	12	100.0%	4.45	1.9e-14	0.488	7.41
ResNet-18	CIFAR-10	2028	12	99.0%	1.89	1.4e-14	0.628	6.98
ResNet-18	CIFAR-10	2030	12	100.0%	6.20	2.5e-13	0.243	21.75
ResNet-18	CIFAR-100	2027	12	100.0%	4.90	1.2e-11	0.070	64.47
ResNet-18	CIFAR-100	2028	12	100.0%	6.76	2.4e-14	0.558	8.27
ResNet-18	CIFAR-100	2030	12	100.0%	5.64	2.5e-13	0.250	16.11
ResNet-18	SVHN	2027	12	100.0%	3.23	2.4e-13	0.316	17.31
ResNet-18	SVHN	2028	12	100.0%	3.42	1.8e-14	0.452	7.08
ResNet-18	SVHN	2030	12	100.0%	3.71	3.0e-14	0.568	8.78
WRN-28-2	CIFAR-10	2027	12	99.0%	10.55	2.4e-13	0.305	11.41
WRN-28-2	CIFAR-10	2028	12	99.0%	10.93	8.0e-14	0.290	10.55
WRN-28-2	CIFAR-10	2030	12	97.9%	10.90	2.9e-13	0.263	11.19
WRN-28-2	CIFAR-100	2027	12	99.0%	19.54	9.3e-14	0.323	7.83
WRN-28-2	CIFAR-100	2028	12	97.9%	21.16	1.7e-13	0.387	12.59
WRN-28-2	CIFAR-100	2030	12	99.0%	18.15	8.0e-14	0.264	8.63
WRN-28-2	SVHN	2027	12	99.0%	5.02	1.4e-13	0.121	5.76
WRN-28-2	SVHN	2028	12	99.0%	4.95	4.4e-14	0.208	6.57
WRN-28-2	SVHN	2030	12	100.0%	5.64	8.8e-14	0.290	11.24
PlainConvNet	CIFAR-10	2027	12	100.0%	22.71	6.3e-13	0.213	26.16
PlainConvNet	CIFAR-10	2028	12	100.0%	13.05	2.3e-13	0.141	23.27
PlainConvNet	CIFAR-10	2030	12	97.9%	19.10	3.4e-14	0.368	10.28
PlainConvNet	CIFAR-100	2027	12	100.0%	19.42	1.1e-13	0.234	13.77
PlainConvNet	CIFAR-100	2028	12	99.0%	21.38	7.8e-14	0.341	12.66
PlainConvNet	CIFAR-100	2030	12	99.0%	19.97	3.7e-14	0.343	9.55
PlainConvNet	SVHN	2027	12	96.9%	8.59	8.0e-14	0.228	11.56
PlainConvNet	SVHN	2028	12	99.0%	10.37	7.3e-14	0.165	9.34
PlainConvNet	SVHN	2030	12	100.0%	14.92	3.7e-14	0.335	11.31

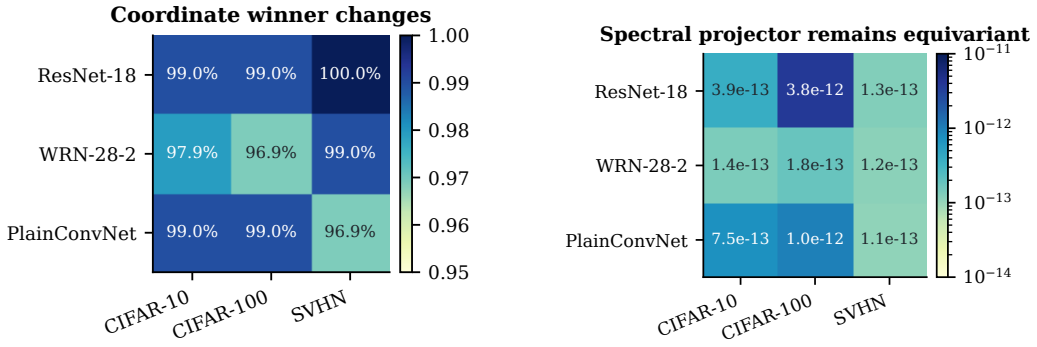
A.3.5 CORE ADVERSARIAL ENDPOINT GRID

The core grid evaluates the frozen CIFAR-10 seed-2026 checkpoints with PGD-40 and APGD-CE at $1/255$ and $2/255$. Each cell selects 12 clean-correct attack successes and 12 clean-correct attack failures, then pairs every attack endpoint with the scale-matched randomized displacement from the same clean image. The labels define a conditional case-control analysis; they do not estimate the dataset-level attack success rate.

The artifact status is FALSIFIED because the frozen all-cell spectral-isolation gate is falsified, while the implementation status is PASS: every cell meets the path-count, matched-norm, projector-covariance, and pulled-back intervention-score criteria. The isolation gate passes 10 of 12 cells; the remaining cells are retained to expose the near-tie boundary. The absolute-effect bootstrap intervals resample endpoint observations within the balanced attack-success and attack-failure strata, and paired intervals resample clean-image pairs. They do not treat the four gauge trials per path as independent evidence.

Table 12: Path-level condition-number-100 stress results. Each row is one frozen path; spectral score error is zero at reported precision on every row.

Path	Rank	Rel. gap	$\|P\|_2$	Coord. range	Proj. residual
9967	1	0.759	4.23	8.89	1.9e-13
6907	2	0.831	4.94	8.00	1.6e-12
7413	2	0.619	6.25	8.41	6.7e-13
2648	1	0.243	16.73	7.89	1.3e-11
5113	1	0.524	21.75	10.23	9.5e-12
3463	2	0.589	14.09	10.52	3.7e-12
3098	2	0.653	5.96	5.84	1.8e-13
7427	1	0.568	5.16	7.28	2.9e-13
1769	1	0.616	4.53	8.36	1.2e-13
2091	1	0.777	4.30	13.66	9.0e-14
4132	1	0.508	3.34	3.73	5.9e-14
3399	1	0.775	3.93	5.95	5.0e-14



(a) Minimum winner-change rate over seeds.

(b) Worst projector residual over seeds.

Figure 5: Balanced visual gauge matrix by architecture and dataset.

A.3.6 Historical two-resample PGD pilot

We reuse the frozen ResNet-18/CIFAR-10 checkpoint and existing matched PGD/random image artifacts, but change the input path from clean-to-input to clean-to-perturbed. The primary setting is ℓ_∞ $2/255$, where the PGD success labels are not saturated. For each of two independent selection seeds, we sample 32 clean-correct PGD-success and 32 clean-correct PGD-failure examples, then evaluate both the PGD and matched random endpoint for every selected sample. Thus the balanced PGD success/failure sample is case-control by construction; its 50% success rate is not an estimate of the dataset-level attack rate. Table 14 reports the two historical resamples.

Both resamples pass the exact implementation checks: the maximum projector covariance error is 4.7e-12 and the maximum pulled-back spectral-score discrepancy is 4.8e-07. Across the two resamples, PGD success-stratum AUC ranges from 0.952–0.959 (pooled AUC 0.956), while PGD median R_P ranges from 2.68–2.81 logits and matched random medians remain near zero (-0.014 to +0.024 logits). The mean paired PGD-minus-random effect is 3.53 logits. This historical pilot motivated the broader endpoint grid, but its signed separation result is not pooled with the core magnitude estimand. Neither artifact establishes a universal adversarial mechanism, cross-architecture transfer beyond the grid, or calibrated attack-success prediction outside the frozen case-control protocols.

A.3.7 Comparator and ablation rationale

The intervention comparison must respect the same gauge symmetry as the proposed method. Conventional hidden-coordinate conductance remains a parameterization-relative diagnostic; the Q2 orbit and relocation experiments use it as a gauge-sensitive negative control, not a gauge-invariant utility

Table 13: Complete adversarial endpoint grid. $AUC_{|R|}$ is the attack-success/failure separation AUC for absolute restored effect, with a bootstrap interval over endpoint strata; $\Delta|R|$ is the paired attack-minus-random mean with its bootstrap interval. Signed quantities are secondary diagnostics. The final column is the maximum paired relative ℓ_2 norm mismatch; the two smallest modulus gaps identify near-tie cells.

Architecture	Attack	ε	$AUC_{\|R\|}$ [95%]	Attack $\|R\|$	Random $\|R\|$	$\Delta\|R\|$ [95%]	AUC_R	ΔR	Eigengap	Modulus gap	Norm err.
ResNet-18	PGD-40	1/255	0.500 [0.264, 0.729]	0.02	0.01	0.015 [0.005, 0.028]	0.535	2.16e-02	0.200	2.9e-03	3.6e-07
ResNet-18	PGD-40	2/255	0.681 [0.438, 0.889]	0.04	0.02	0.016 [-0.004, 0.039]	0.660	4.84e-02	0.212	1.3e-03	2.9e-07
ResNet-18	APGD-CE	1/255	0.556 [0.319, 0.778]	0.02	0.02	0.012 [0.001, 0.023]	0.528	1.63e-02	0.093	3.4e-04	2.2e-07
ResNet-18	APGD-CE	2/255	0.590 [0.340, 0.806]	0.02	0.01	0.019 [0.001, 0.040]	0.576	1.33e-02	0.271	1.4e-02	2.3e-07
WRN-28-2	PGD-40	1/255	0.674 [0.444, 0.903]	0.04	0.02	0.013 [-0.003, 0.026]	0.403	1.01e-02	0.130	6.5e-03	2.9e-07
WRN-28-2	PGD-40	2/255	0.451 [0.215, 0.688]	0.05	0.03	0.025 [0.008, 0.042]	0.340	-4.44e-04	0.194	2.0e-03	3.1e-07
WRN-28-2	APGD-CE	1/255	0.792 [0.590, 0.965]	0.04	0.02	0.016 [0.006, 0.027]	0.514	5.47e-03	0.172	4.8e-04	3.0e-07
WRN-28-2	APGD-CE	2/255	0.861 [0.674, 1.000]	0.02	0.02	0.003 [-0.008, 0.014]	0.500	-4.56e-03	0.232	9.2e-03	1.6e-07
PlainConvNet	PGD-40	1/255	0.500 [0.250, 0.729]	0.11	0.04	0.079 [0.060, 0.099]	0.583	-9.20e-03	0.112	1.3e-02	2.1e-07
PlainConvNet	PGD-40	2/255	0.576 [0.340, 0.806]	0.18	0.08	0.095 [0.058, 0.132]	0.583	5.41e-02	0.171	5.1e-03	2.3e-07
PlainConvNet	APGD-CE	1/255	0.694 [0.465, 0.882]	0.10	0.04	0.064 [0.044, 0.088]	0.396	-5.17e-02	0.112	4.0e-03	2.2e-07
PlainConvNet	APGD-CE	2/255	0.882 [0.715, 0.993]	0.16	0.08	0.050 [0.006, 0.092]	0.333	-3.00e-02	0.131	3.4e-03	2.2e-07

Table 14: Historical two-resample PGD pilot on a frozen ResNet-18/CIFAR-10 checkpoint at $2/255$. Paths gives PGD/random counts; AUC is computed only within the balanced PGD success/failure selection; R_P is the dominant restored target-logit effect from Eq. 10.

Resample	Paths	PGD AUC	PGD median R_P	Random median R_P	Paired mean difference
Resample 1	64/64	0.952	2.81	-0.014	3.41
Resample 2	64/64	0.959	2.68	0.024	3.65

comparator. We therefore use the runner-up spectral cluster not as a competing attribution method, but as a same-rank, gauge-equivariant spectral comparator computed from the same matrix. Selecting the largest remaining modulus within that rank-matched set makes the comparison deterministic and prevents tuning a weak comparator after seeing outcomes.

This tests whether the dominant-cluster rule contributes beyond spectral invariance alone. The comparison is local rather than exhaustive: random invariant subspaces require an extra gauge-dependent metric, whereas coordinate conductance answers the negative-control question above. Future comparisons should include other analytic matrix functions and task-supervised invariant subspaces, with their selection information and rank controlled explicitly.

Our ablation removes the projected endpoint displacement while retaining the same path baseline and downstream function. Absolute target-logit change is paired within each path, avoiding cross-example scale comparisons in the win metric. The effect ratio supplies magnitude, while Wilson intervals expose uncertainty from the finite number of paths. No result is filtered by whether the dominant cluster wins.

A.3.8 Nonlinear cuts and scope of evidence

The theorem allows arbitrary differentiable downstream networks, while some direct controllability experiments use penultimate linear heads for exact prescription and transparent ablation. Separate trained nonlinear-cut probes on all three architectures recover nontrivial, full-rank conductance matrices and nontrivial power-trace information beyond completeness. Their worst numerical completeness error is 9.74×10^{-3}. These probes establish that the matrix object is not restricted to a linear head, but we do not count them as intervention-utility evidence. The balanced spectral matrix supplies the nonlinear-cut gauge tests; broader nonlinear downstream utility remains an important extension.

A.3.9 Vision intervention utility

Table 5 now reports the complete five-cell utility summary, including both median effect magnitudes rather than ratios alone. The five source artifacts retain all sampled paired path outcomes, including

comparator wins and isolation exclusions; the aggregate artifact records the exact per-cell requested and usable counts and Wilson intervals.

A.3.10 AG NEWS ENVIRONMENTS AND COMPLEMENT HOLDOUT

We reuse the locally audited AG News corpus but train a new two-layer, four-head Transformer from scratch, with no pretrained data. The 128-dimensional CLS cut feeds a nonlinear GELU MLP head. Conductance follows a differentiable path from a null-token embedding baseline through the Transformer, rather than interpolating discrete token IDs. A seed-2026 pilot freezes the protocol; seeds 2027–2029 are untouched holdouts evaluated on clean text, deterministic word dropout, and deterministic character deletion. Table 15 reports the seed-level worst-environment summary.

Table 15: AG News holdouts. Accuracy and coordinate change are worst-environment values; RU denotes wins over the matched-rank runner-up.

Seed	Paths	Acc.	Coord.	Proj. err.	RU
2027	36	73.70%	97.9%	8.8e-15	100.0%
2028	36	73.60%	97.9%	1.6e-14	100.0%
2029	72	73.85%	97.9%	1.5e-14	100.0%

Across 9 seed–environment cells and 144 paths, the weakest environment accuracy is 73.60% and conductance matrices have numerical rank up to 22. The worst completeness error is 1.74e-03, maximum projector covariance error is 1.64e-14, and spectral score discrepancy is zero at reported precision. The minimum environment-level coordinate-winner change rate is 97.92%, while the dominant cluster wins the paired absolute target-logit comparison against the matched-rank runner-up on every path. In the fresh seed-2029 holdout, it also beats the higher-capacity entire spectral complement on 100.0% of 72 paths; median absolute effects are 2.21 versus 0.004 logits.

Table 16: Complete AG News environment holdouts. Each row is one seed–environment cell; Acc. is accuracy, Coord. is coordinate-winner change, RU is the dominant-versus-runner-up win rate, and Ratio is the median paired effect ratio.

Seed	Environment	Paths	Accuracy	Coord. change	Proj. residual	RU	Effect ratio
2027	Clean	12	74.5%	100.0%	8.8e-15	100%	237.5
2027	Word dropout	12	74.0%	97.9%	3.2e-15	100%	451.9
2027	Character deletion	12	73.7%	100.0%	4.0e-15	100%	938.7
2028	Clean	12	75.4%	100.0%	4.0e-15	100%	195.4
2028	Word dropout	12	73.9%	100.0%	1.6e-14	100%	488.3
2028	Character deletion	12	73.6%	97.9%	7.3e-15	100%	351.5
2029	Clean	24	74.9%	100.0%	5.2e-15	100%	974.5
2029	Word dropout	24	73.9%	97.9%	1.5e-14	100%	232.7
2029	Character deletion	24	74.3%	99.0%	1.0e-14	100%	1456.1

Seeds 2027–2028 use the matched-rank runner-up; seed 2029 additionally tests the entire spectral complement on 24 paths per environment. This separately frozen comparison is reported here rather than extrapolated to the two earlier holdouts. This result is a modality replication, not merely another image dataset: the encoder, inputs, perturbations, and nonlinear head all differ. It remains one text corpus and one Transformer family, and the very small complement effect makes universal superiority an inappropriate claim.

A.4 COMPLETE EXPERIMENTAL PROTOCOL

A.4.1 BALANCED MATRIX

The primary matrix contains 12 paths per architecture–dataset–seed cell and four dense gauges per path, with both direct and composed checks contributing to the recorded gauge-trial total. Independent

seeds repeat model training and the complete matrix rather than merely resampling gauges. Paths are restricted to correctly classified examples to avoid conflating gauge behavior with an incorrect endpoint prediction.

For every trial we verify original and transformed logits, the conductance similarity residual, projector covariance after pullback, equality of spectral intervention scores, coordinate top selection, and coordinate-selected intervention range. We also record dominant-cluster rank, relative eigengap, and projector norm. A cell passes only when all frozen criteria pass.

A.4.2 Intervention-utility protocol

The development comparison contains 48 paths. The comparator is the largest remaining isolated spectral cluster with the dominant cluster's rank (the same-rank runner-up). Primary effects use the true-class logit on clean-correct paths; the next class modulo the output dimension is evaluated as a frozen target-control diagnostic. Holdout paths, checkpoints, cluster rules, outcome metrics, and thresholds were frozen before execution. We report path-level win fractions with Wilson intervals because each holdout contains 24 independent paired paths; effect ratios summarize magnitude but are not treated as calibrated probabilities.

A.4.3 Attack-path protocol

The core attack grid uses frozen seed-2026 ResNet-18, WRN-28-2, and PlainConvNet CIFAR-10 checkpoints, PGD-40 and APGD-CE, and $1/255$ and $2/255$ budgets. For each of 12 cells, selection scans clean-correct examples until it obtains 12 attack successes and 12 attack failures; each selected attack endpoint is paired with a box-feasible randomized assignment/sign control that preserves its displacement magnitudes. The same IDs pair each endpoint type; no endpoint or label is regenerated during conductance evaluation. We compute 129-point clean-to-endpoint paths, use the true-class logit as the scalar target, select the largest-modulus cluster, and apply four condition-bounded dense gauges per endpoint path. The primary estimand is $|R|$, with signed R retained as a directional diagnostic. Both AUCs are within-selection descriptive separation statistics, not population attack-success estimates; bootstrap intervals resample endpoint strata or clean-image pairs, never gauge trials. The historical $4/255$ pilot is retained only as supplementary magnitude context because its PGD labels were saturated.

A.4.4 Evidence ladder and statistical design

Table 18 separates algebraic verification, adversarial severity, replication, numerical stress, and utility. Accumulating gauge trials cannot strengthen an exact theorem, while utility cannot prove the gauge law; each block therefore has a distinct falsification target.

Table 17 records how the numerical tables relate to the frozen artifacts in `experiments/results`. PDF tables expose every independent seed or environment cell and all stress paths; the larger utility, matrix, and NLP artifacts retain the paired path rows and gauge draws rather than treating them as independent replications.

Table 17: Numerical provenance and aggregation. "Worst" denotes the maximum error/range or minimum success statistic, as appropriate. Braced names denote one artifact per listed cell.

Evidence block	PDF rows	Retained rows	Source artifact	Reported aggregation
Direct relocation	3 pilot + 18 holdout	21 seed trials	`direct_conductance_gauge_{intervention,holdout}_v1.*`	per seed; pilot per architecture
Balanced visual matrix	27	324 paths / 2,592 gauges	`conductance_spectral_matrix_{arch}_{data}_seed{s}_v1.json`	per seed cell; path worst cases and trial rates
Quadrature convergence	3	9 paths / 3 resolutions	`conductance_quadrature_convergence_v1.json`	production-vs-reference matrix and effect errors
Conditioning stress	12	12 paths	`conductance_spectral_stability_stress_v1.json`	every path, no pooling
Vision utility	5	120 paired paths	`conductance_spectral_causal_utility_{cell}_v1.json`	wins, Wilson LB, and paired medians per cell
Attack endpoint grid	12 cells	288 attack + 288 random paths	`conductance_attack_path_grid_v1.json`	magnitude AUC, paired absolute effects, signed diagnostics, and isolation gates
Historical attack pilot	2 resamples	128 PGD + 128 random paths	`conductance_attack_path_n64_{,resample2}_v1.json`	supplementary signed AUC and paired restored effects
NLP replication	9	144 paired paths	`nlp_conductance_gauge_seed{s}*.json`	per seed–environment cell; seed worst cases in main text

Invariance reports worst-case residuals because one failed path falsifies the implementation claim; coordinate instability uses per-cell minima and path-cluster bootstrap lower bounds, while gauge-trial Wilson intervals are descriptive. Utility uses paired path-level wins and 95% Wilson intervals. Input

Table 18: Evidence ladder. Counts are not pooled across incompatible estimands.

Block	Scale	Role
Nonlinear theorem check	general downstream map	similarity, diagonal, and trace implementation
Targeted credit relocation	3 architectures	realize a bottom-decile unit as the credit winner
Direct frozen holdout	9 arch.–data cells	preserve logits and complete totals on untouched seeds
Balanced spectral matrix	27 cells; 324 paths; 2,592 gauges	replicate instability and projector equivariance
Quadrature convergence	3 cells; 9 paths; 3 resolutions	check the production path-integral discretization
Conditioning stress	$\kappa(M) \leq 100$	expose eigengap and projector-norm sensitivity
Vision utility holdouts	5 settings; 120 paths	test generalization without retuning
Attack endpoint grid	12 cells; 576 endpoint paths	test conditional PGD/APGD magnitude separation and gauge covariance
Historical attack pilot	2 resamples; 256 retained paths	supplementary PGD signed-effect replication
NLP holdouts	9 cells; 144 paths	test cross-modal replication and text-shift behavior

paths are experimental units, training seeds and architecture–dataset or text-environment cells are replication axes, and gauge draws are not independent utility evidence.

A.5 Discussion and Scope

Even apart from feature superposition (Elhage et al., 2022), function-preserving affine gauges make coordinate attribution non-identifiable; sparse dictionary learning can act as a *gauge-fixing mechanism*. Coordinate attribution therefore requires a privileged or explicitly gauge-fixed basis; otherwise, the fixed spectral rule supplies a gauge-equivariant intervention target.

The evidence supports four bounded claims: theory and trained models permit coordinate reassignment without function change; spectral interventions obey the gauge law and the dominant cluster has a larger target-conditioned effect than the matched-rank comparator; and the attack grid compares restored-effect magnitude with matched-random endpoint controls. We do not establish natural dense gauges, useless neuron analyses, semantic or universally optimal clusters, or coverage of every subspace family. Attack evidence is conditional on PGD/APGD, three seed-2026 CIFAR-10 checkpoints, two non-saturated budgets, and the clean-correct case-control protocol; it does not claim universal robustness prediction or transfer. Magnitude is positive in 12/12 cells with positive paired bootstrap lower bounds in 9 cells; signed direction is mixed and 2 cells approach the modulus-tie boundary. The scope is bounded by the $GL(d)$ functional equivalence and reported separation/conditioning, with priorities of semantic causal ground truth, multiple layer depths, and low-rank approximations for wide layers.

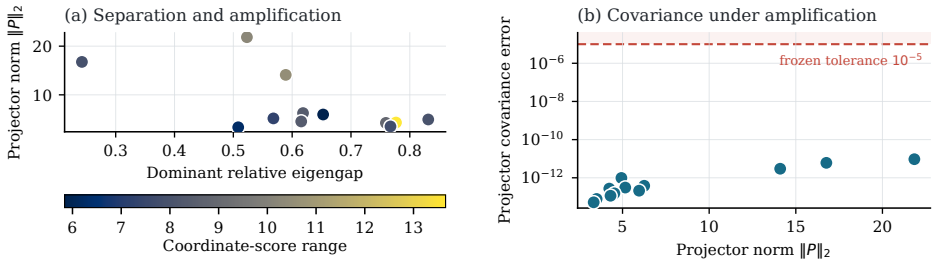


Figure 6: Conditioning stress. Projector covariance error remains below the frozen 10^{-5} tolerance at condition number 100.

A.6 SCOPE

The supported scope is exact gauge non-identifiability, numerical spectral equivariance, and conditional utility in frozen vision and text settings; it does not imply useless neuron analyses, naturally occurring arbitrary gauges, uniquely optimal projectors, or semantic causal discovery.