BEYOND STATIC EMBEDDINGS: FINITE-QUERY CERTIFICATES FOR LOCAL REPRESENTATION STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The geometry of finitely observed embeddings does not determine the local stability of the representation map. We investigate this information gap using a finite-query framework. Within the class of continuously differentiable functions, identical activation observations can correspond to arbitrarily large local sensitivity. Moreover, perturbation queries can leave directions outside their span unidentified. Using a classical anchored interpolation operator, we can recover the Jacobian restricted to the covered subspace, accounting for a curvature-conditioning error that relies on an independently provided local smoothness bound. This analysis yields two-sided sensitivity estimates, finite-scale displacement bounds, and a prediction-invariance condition that accounts for the readout and decision margin. We present a 1D construction that meets the recovery bound for a one-sided endpoint query. This operator analysis leads to Secant-Frame Regularization (SFR), an input-Jacobian-free objective that targets the largest reconstructed logit response. Through experiments, we instantiate the conditional bounds. On the CIFAR-10 dataset, SFR demonstrates improved robustness in evaluated covered transformations compared to matched-query pairwise consistency. Our findings link finite-representation observations to a defined local stability analysis and distinguish conditional certification from empirical regularization.

1 INTRODUCTION

Representation learning enables neural networks to transform observations into feature spaces that enhance prediction and transfer capabilities. Analyzing representation similarity (Kornblith et al., 2019), effective rank (Garrido et al., 2023), class concentration (Papyan et al., 2020), and alignment and uniformity (Wang & Isola, 2020) has been crucial for understanding these spaces. Learning objectives like Barlow Twins and VICReg shape representation geometry by promoting consistency between augmented views while controlling redundancy and preventing collapse (Zbontar et al., 2021; Bardes et al., 2022a;b). Recent advancements have also extended this analysis to language model representations (Pan & Woodard, 2026), with theoretical connections established between joint embedding pretext objectives and downstream prediction losses (Wasserer et al., 2026). A useful representation may not remain stable despite high nominal accuracy, as vulnerabilities to adversarial attacks (Madry et al., 2019) and common corruptions (Hendrycks & Dietterich, 2019) can coexist. Local representation stability indicates the magnitude of feature changes from permitted input perturbations. Recent studies show that representational and functional similarities may diverge (Braun et al., 2025), and the relationship between representational similarity and task behavior can vary by evaluation dataset (Ciernik et al., 2025). This leads to the crucial question: *What information, alongside which regularity assumptions, can sufficiently establish an upper bound on local representation sensitivity based on finite function evaluations?*

This question matters because descriptive evidence and deterministic stability guarantees can lead to different conclusions. A finite activation matrix may indicate feature redundancy but does not provide insight into changes between observed inputs. Additionally, small responses to limited perturbations only reveal information about tested directions, not the broader input space. Differentiating these cases is important when designing learning objectives: while suppressing nuisance variation can be beneficial, suppressing task-relevant variation may erase useful discriminative information. Hence, a stability analysis should specify the examined perturbation directions and quantify the uncertainty of its conclusions. Existing research offers valuable insights for this analysis. Shared-invariance measures compare models based on perturbation behavior (Nanda et al., 2022), while neural-frame analyses assess changes in local directions as they move through networks (Kvinge et al., 2024). Jacobian regularization manages input-output sensitivity, with recent theories addressing its training dynamics in the infinite-width limit (Jakubovitz & Giryes, 2018; Kim & Yang, 2024). Compositional curvature bounds provide additional robustness certification information for continuously

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.

differentiable networks (Entesari et al., 2024). Our aim is to characterize what finite representation evaluations can establish about local stability, which directions remain unconstrained, and what independent regularity information is required to obtain a bound. Figure 1 summarizes the framework. The conditional certificates and the SFR objective serve different roles: the former require established smoothness bounds, whereas the latter can be evaluated and optimized without certifying those bounds.

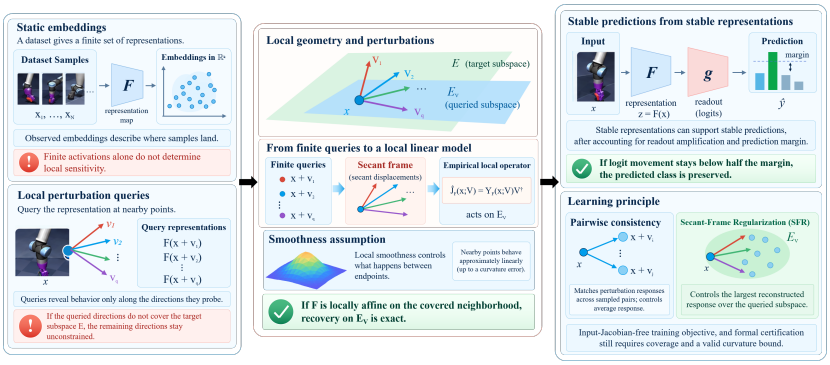

Figure 1: Overview of the finite-query stability framework.

Our ***key contributions*** are: (1) We formalize an information boundary between finite activation geometry and local functional stability, identifying the limitations of activation-only observations and under-spanning perturbation queries over the stated function classes. (2) We connect finite secant observations to restricted-Jacobian recovery and conditions for finite-scale prediction, highlighting the roles of coverage, conditioning, curvature, readout amplification, and margin. A scalar construction demonstrates sharpness in the specified one-sided query setting. (3) We formulate a task-anchored learning objective that targets the largest reconstructed response and quantify its distinction from fixed-frame average squared consistency, while separating empirical regularization from the assumptions required for certification.

2 Related Work

Representation geometry and functional behavior. Similarity, alignment, effective rank, and linear decodability define observed representation structures (Kornblith et al., 2019; Wang & Isola, 2020; Garrido et al., 2023; Harvey et al., 2024). Methods like Barlow Twins, VICReg, and VICRegL focus on shaping agreement and redundancy (Zbontar et al., 2021; Bardes et al., 2022b;a). Recent work emphasizes collapse prevention (Sansone et al., 2025) and pretext-to-downstream guarantees under specific assumptions (Wasserer et al., 2026). Neural collapse highlights the alignment of concentration with discriminative class geometry (Papyan et al., 2020). Research distinguishes representational similarity from task performance, addressing robust model comparisons and limitations in dataset dependence and model stitching (Jones et al., 2022; Balogh & Jelasity, 2023; Ciernik et al., 2025; Braun et al., 2025; Smith et al., 2025). STIR investigates shared invariances (Nanda et al., 2022), while tangent propagation and manifold methods direct transformation paths (Simard et al., 1991; Rifai et al., 2011). Neural frames analyze how local perturbations affect representations (Kvinge et al., 2024). These approaches motivate the distinction between observed embedding structure and functional behavior. We focus on finite-query identifiability: which local-sensitivity claims are supported by a given observation record, on which subspace, and under what independent regularity assumptions.

Interpolation, sensitivity bounds, and regularization. Anchored linear models, vector-valued residual interpolation, and subspace reconstruction are vital to derivative-free optimization (DFO) (Conn et al., 2009; Cartis & Roberts, 2019; 2023), and have been enhanced by advances in constrained model construction and finite-difference trust-region methods (Roberts, 2025; Davar & Grapiglia, 2025). We use this established interpolation construction as an observation operator for representation stability. Input-gradient and Jacobian penalties regulate derivatives (Ross & Doshi-Velez, 2018; Jakubovitz & Giryes, 2018), while Parseval constraints influence layer operators (Cisse et al., 2017). Additional research focuses on Jacobian-regularized training and low-rank Jacobian structures (Kim & Yang, 2024; Scarvelis & Solomon, 2024). Lipschitz-margin methods convert sensitivity into predictive criteria (Tsuzuku et al., 2018; Sokolić et al., 2017), and analyses of local linearity and second-order effects address nonlinear variations (Qin et al., 2019; Singla & Feizi, 2020; Entesari et al., 2024). Techniques like Mean Teacher, virtual adversarial training, and shift consistency utilize finite perturbation responses for signal learning (Tarvainen & Valpola, 2018; Miyato et al., 2018; Zhang et al., 2021). Our method fits a vector-valued operator on a prescribed query span and penalizes its largest reconstructed gain. The accompanying fixed-frame comparison

quantifies the difference between average squared response and worst-direction control. More related work is offered in Appendix B.

3 Problem Setup and Secant-Frame Formulation

Throughout, we equip vectors with the Euclidean norm $\|\cdot\|_2$. For a matrix A, $\|A\|_{\mathrm{op}}$, $\|A\|_{\mathrm{F}}$, and $A^\dagger$ denote its spectral norm, Frobenius norm, and Moore-Penrose pseudoinverse. For every nonzero matrix A, $\sigma^+_{\min}(A)$ denotes its smallest nonzero singular value. The representation-sensitivity quantities use fixed Euclidean input and representation metrics, which remain unchanged under orthogonal coordinate transformations of the map and the perturbation subspace. In contrast, prediction-level statements rely on a specified logit norm and fixed class coordinates. Thus, arbitrary orthogonal mixing of logits does not preserve the norm or the predicted label.

3.1 Local representation and prediction stability

Let $\Omega \subseteq \mathbb{R}^d$ be open and let $F : \Omega \to \mathbb{R}^p$ be a representation map. For prediction-level statements, write $f = g \circ F : \Omega \to \mathbb{R}^C$, where g maps the representation to C logits. This decomposition separates stability of an intermediate representation from stability of the final predictor. To state either notion relative to a specified perturbation model, we first define the neighborhood over which perturbations are allowed.

Definition 3.1 (Directional Neighborhood). *For $x \in \Omega$, a nonzero linear subspace $E \subseteq \mathbb{R}^d$, and $\rho > 0$, define $\mathbb{B}_E(x,\rho) := \{x + \delta : \delta \in E,\ \|\delta\|_2 \le \rho\}$. The triple (x, E, ρ) is admissible if $\mathbb{B}_E(x,\rho) \subseteq \Omega$.*

The subspace E is part of the stability claim: it may be the full input space, a tangent or low-frequency subspace, or the coordinate space of a differentiable transformation family. More generally, if $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ is open, $0 \in \mathcal{A}$, and $\phi_x : \mathcal{A} \to \Omega$ is differentiable with $\phi_x(0) = x$, then the definitions below apply to the composite map $H_x(a) := F(\phi_x(a))$ at the anchor $a = 0$, where d_a is the coordinate-space dimension. Having fixed the admissible perturbation domain, we next quantify representation movement within it.

Definition 3.2 (Directional Sensitivity). *For an admissible triple (x, E, ρ), the anchored finite-scale directional sensitivity is*

$$\mathsf{ASens}_\rho(F; x, E) := \sup_{\substack{\delta \in E \\ 0 < \|\delta\|_2 \le \rho}} \frac{\|F(x+\delta) - F(x)\|_2}{\|\delta\|_2}. \tag{1}$$

If F is Fréchet differentiable at x, its infinitesimal directional sensitivity is

$$\mathsf{DSens}(F; x, E) := \|J_F(x) P_E\|_{\mathrm{op}}, \tag{2}$$

where $J_F(x) \in \mathbb{R}^{p \times d}$ is the Jacobian and P_E is the orthogonal projection onto E. Moreover, $\lim_{\rho \downarrow 0} \mathsf{ASens}_\rho(F; x, E) = \mathsf{DSens}(F; x, E)$.

Eq. 1 captures nonlinear behavior over a nonzero radius, whereas Eq. 2 is the first-order gain at the anchor. Both quantities are restricted to E. To translate a bound on representation movement into a statement about the classifier, we also require the separation between the winning and competing logits.

Definition 3.3 (Prediction Margin). *Assume that $c_f(x) := \arg\max_{1 \le c \le C} f_c(x)$ is unique. The predicted logit margin at x is $\mathsf{mar}_f(x) := f_{c_f(x)}(x) - \max_{c \ne c_f(x)} f_c(x) > 0$.*

If $x + \delta \in \Omega$ and $\|f(x+\delta) - f(x)\|_\infty < \mathsf{mar}_f(x)/2$, then $c_f(x+\delta) = c_f(x)$. Thus, the margin measures decision separation at the anchor, but it supports prediction stability only when perturbation-induced logit movement is also controlled.

3.2 Finite observations

The local quantities above depend on F values away from the anchor. To isolate what is available without such off-sample evaluations, let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subseteq \Omega \times \mathcal{Y}$ have pairwise distinct inputs and formalize statistics that use the corresponding finite representation values.

Definition 3.4 (Activation-Only Statistic). *The finite activation observation is $\mathscr{O}_{\mathcal{D}_N}(F) := (F(x_1), \ldots, F(x_N))$. A statistic S is activation-only on $\mathcal{D}_N$ if $\mathsf{S}(F; \mathcal{D}_N) = \Psi_{\mathcal{D}_N}(\mathscr{O}_{\mathcal{D}_N}(F))$ for some map $\Psi_{\mathcal{D}_N}$. The map may use labels or other fixed metadata in $\mathcal{D}_N$, but it does not evaluate F outside $\{x_1, \ldots, x_N\}$.*

Two maps are observationally equivalent on $\mathcal{D}_N$, written $F \equiv_{\mathcal{D}_N} \widetilde{F}$, if $F(x_i) = \widetilde{F}(x_i)$ for every i. Every activation-only statistic is constant on each observational-equivalence class. Such statistics can describe the organization of observed embeddings.

3.3 Secant frames and covered subspaces

Activation-only information contains no direct observation of local representation movement. Hence, we augment an anchor with finite perturbation queries and retain the subspace spanned by those queries. Fix an admissible triple (x, E, ρ) and let $V = [v_1, \dots, v_q] \in \mathbb{R}^{d\times q}$, $v_j \in E \setminus \{\mathbf{0}\}$, $\|v_j\|_2 \le \rho, \quad j = 1, \dots, q$. Define $E_V := \mathrm{range}(V)$. The frame V covers a prescribed subspace E if $E_V = E$. Since V is nonzero, $\sigma^+_{\min}(V) > 0$, and $\|V^\dagger\|_{\mathrm{op}} = \frac{1}{\sigma^+_{\min}(V)}$. Thus, weak representation of a direction at the queried scales amplifies errors when finite differences are reconstructed into an operator. With the queries and their covered subspace fixed, we collect the corresponding function values and representation differences into a local observation.

Definition 3.5 (Local Secant Frame)**.** *The local query observation, secant matrix, and local secant frame are* $\mathcal{Q}_{x,V}(F) := (F(x), F(x + v_1), \dots, F(x + v_q))$; $Y_F(x; V) := [F(x + v_1) - F(x), \dots, F(x + v_q) - F(x)] \in \mathbb{R}^{p\times q}$; *and* $\mathfrak{S}_F(x; V) := (V, Y_F(x; V))$. *The associated empirical secant operator is* $\widehat{J}_F(x; V) := Y_F(x; V)V^\dagger \in \mathbb{R}^{p\times d}$.

Because $VV^\dagger = P_{E_V}$, the empirical operator satisfies $\widehat{J}_F(x; V) = \widehat{J}_F(x; V)P_{E_V}$ and therefore acts only on the covered subspace. It is the unique minimum-Frobenius-norm element among the minimizers of $\min_{A\in\mathbb{R}^{p\times d}} \|AV - Y_F(x; V)\|_{\mathrm{F}}$. Accordingly, it is the canonical anchored linear least-squares model fitted to the observed input-representation differences. Such local interpolation models are standard in DFO (Conn et al., 2009; Cartis & Roberts, 2019; 2023). Here, the same algebra serves as an observation operator for representation stability. A secant frame contains only finitely many endpoint evaluations. To control behavior between those endpoints and relate the fitted operator to the true directional Jacobian, we impose an independent local regularity condition.

Assumption 3.6 (Local Directional Smoothness)**.** *The map F is continuously differentiable on an open neighborhood of $\mathbb{B}_E(x, \rho)$, and there exists $\beta \ge 0$ such that $\|(J_F(z) - J_F(z'))P_E\|_{\mathrm{op}} \le \beta\|z - z'\|_2$ for all $z, z' \in \mathbb{B}_E(x, \rho)$.*

Assumption 3.6 is a directional $C^{1,1}$ condition: it controls variation of the Jacobian only on E. For every $v \in E$ with $\|v\|_2 \le \rho$, the integral form of Taylor's theorem gives $F(x+v)-F(x)-J_F(x)v = \int_0^1 \big(J_F(x+tv)-J_F(x)\big)v\,dt$, and therefore $\|F(x+v)-F(x)-J_F(x)v\|_2 \le \frac{\beta}{2}\|v\|_2^2$. The empirical secant operator supplies information about the first-order response, while β is an independently established bound on Jacobian variation throughout the neighborhood containing all queried and certified inputs. For certification, the documented curvature upper bound must be valid throughout a directional neighborhood containing all queried inputs and all perturbed inputs for which stability is claimed. For a smoothness-based certificate in transformation coordinates, Assumption 3.6 must hold for the composite map, including the transformation and preprocessing operations.

4 Finite-Scale Stability Theory and Learning

Section 3 separated finite activation values, local perturbation responses, and regularity between queried points. We now establish what each information level can support. Complete proofs are provided in Appendix D.

4.1 A finite-query information hierarchy

We begin with the weakest observation model: F is evaluated only on the finite dataset, without any off-sample queries around the observed inputs.

Theorem 4.1 (Activation-Only Non-Identifiability)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be open, let $\mathcal{D}_N = \{(x_j, y_j)\}_{j=1}^N$ have pairwise distinct inputs, fix $i \in \{1, \dots, N\}$, and let $E \subseteq \mathbb{R}^d$ be a nonzero linear subspace. For every $F_0 \in C^1(\Omega; \mathbb{R}^p)$ and every $M > 0$, there exists $F_M \in C^1(\Omega; \mathbb{R}^p)$ such that $F_M \equiv_{\mathcal{D}_N} F_0$ and $\mathsf{DSens}(F_M; x_i, E) > M$. Moreover, F_M can be chosen so that there exists $r_i > 0$ with $\mathbb{B}_{\mathbb{R}^d}(x_i, r_i) \subseteq \Omega$ and*

$$J_{F_M}(z) - J_{F_M}(z') = J_{F_0}(z) - J_{F_0}(z') \tag{3}$$

for all $z, z' \in \mathbb{B}_{\mathbb{R}^d}(x_i, r_i)$.

Theorem 4.1 distinguishes the geometry of sampled representations from the local behavior of the representation map. Maps that have identical finite activations can show significantly different

directional sensitivities at the selected anchor. Thus, no activation-only statistic can provide a uniformly applicable finite upper bound across the specified C^1 function class. This does not imply that static statistics are useless or that such constructions can be realized within every fixed neural architecture. Eq. 3 clarifies the distinction: the absolute first-order gain can increase while the Jacobian variation remains unchanged within a sufficiently small neighborhood around the anchor. This preservation is local and may not apply to a larger region that includes other observed inputs. Next, we explore whether a finite number of local perturbation observations can constrain sensitivity throughout a defined subspace.

Proposition 4.2 (Uncovered Directions Remain Non-Identifiable). *Let $V = [v_1, \dots, v_q]$ be a perturbation frame at x, set $G := E_V$, and let $E \subseteq \mathbb{R}^d$ be a linear subspace with $G \subsetneq E$. For every map $F : \Omega \to \mathbb{R}^p$ differentiable at x and every $M > 0$, there exists a map $\widetilde{F}_M : \Omega \to \mathbb{R}^p$, differentiable at x, such that $\mathscr{Q}_{x,V}(\widetilde{F}_M) = \mathscr{Q}_{x,V}(F)$ and $\mathsf{DSens}(\widetilde{F}_M; x, E) > M$. If F is continuously differentiable on a neighborhood $\mathcal{U}$ of x and satisfies a Jacobian-difference bound there, $\widetilde{F}_M$ can be chosen to satisfy the same bound on $\mathcal{U}$.*

Proposition 4.2 establishes a second information barrier: exact, noiseless secants on V do not contain any deterministic information about directions in $E \cap E_V^{\perp}$. A finite-query certificate must be stated on E_V, unless a separate structural or probabilistic assumption connects the observed and unobserved directions. The number of probes alone is not sufficient: while $q \geq \dim(E)$ is necessary for full coverage, linear dependence can still result in $\operatorname{rank}(V) < \dim(E)$. For representation learning, this result limits the applicability of sparse directional diagnostics. A small set of probes in a high-dimensional input space cannot guarantee unrestricted robustness in the ambient space. However, it can still support valid claims for a fully covered low-dimensional perturbation model, such as transformation coordinates, a prescribed frequency subspace, or a locally estimated tangent space.

4.2 Coverage-aware recovery and finite-scale stability

With the query frame fixing the target perturbation subspace, the analysis shifts from identifiability to estimation. The observed secants determine an anchored linear model on E_V, but for a nonlinear representation this model need not equal $J_F(x)P_{E_V}$. Finite probe lengths create Taylor remainders, and pseudoinverse reconstruction can amplify those remainders when V is poorly conditioned. For a nonzero frame $V = [v_1, \dots, v_q]$, define the curvature-conditioning error $\eta_\beta(V) := \frac{\beta}{2\,\sigma^+_{\min}(V)} \left(\sum_{j=1}^q \|v_j\|_2^4\right)^{1/2}$. This bounds the uncertainty created by finite-scale querying under Assumption 3.6. The numerator aggregates the columnwise Taylor remainders, while $1/\sigma^+_{\min}(V) = \|V^\dagger\|_{\mathrm{op}}$ measures their worst amplification.

Theorem 4.3 (Coverage-Aware Secant Recovery). *Let (x, E, ρ) be admissible and satisfy Assumption 3.6 with constant β. Let $V = [v_1, \dots, v_q]$ be nonzero, with $v_j \in E \setminus \{\mathbf{0}\}$ and $\|v_j\|_2 \leq \rho$ for every j, and set $E_V = \operatorname{range}(V)$. Then*

$$\|\widehat{J}_F(x; V) - J_F(x)P_{E_V}\|_{\mathrm{op}} \leq \eta_\beta(V). \tag{4}$$

Consequently,

$$\Big[\|\widehat{J}_F(x; V)\|_{\mathrm{op}} - \eta_\beta(V)\Big]_+ \leq \mathsf{DSens}(F; x, E_V) \leq \|\widehat{J}_F(x; V)\|_{\mathrm{op}} + \eta_\beta(V). \tag{5}$$

Moreover, every $\delta \in E_V$ with $\|\delta\|_2 \leq \rho$ satisfies

$$\|F(x+\delta) - F(x)\|_2 \leq \Big(\|\widehat{J}_F(x; V)\|_{\mathrm{op}} + \eta_\beta(V)\Big)\,\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2. \tag{6}$$

Hence, for every $0 < \rho' \leq \rho$, $\mathsf{ASens}_{\rho'}(F; x, E_V) \leq \|\widehat{J}_F(x; V)\|_{\mathrm{op}} + \eta_\beta(V) + \frac{\beta}{2}\rho'$.

Theorem 4.3 separates the estimated response from the strength of the evidence supporting that estimate. The center of the interval in Eq. 5 ($\|\widehat{J}_F(x; V)\|_{\mathrm{op}}$) is the largest gain of the fitted secant operator. The radius ($\eta_\beta(V)$) reflects finite probe scale, local nonlinearity, and frame conditioning. A large $\eta_\beta(V)$ makes this recovery bound uninformative about local sensitivity under the available assumptions. By itself, it does not prove instance-specific ambiguity. In particular, small observed secants accompanied by poor conditioning do not establish stability. The recovered object is $J_F(x)P_{E_V}$, not the full Jacobian unless $E_V = \mathbb{R}^d$. This measures how the selected family of input variations is

locally transported into feature space. Small restricted sensitivity is not, by itself, a complete measure of representation quality: a constant representation is stable but non-discriminative, and stability on E_V says nothing about task-relevant directions outside that subspace. Hence, prediction-level conclusions require the downstream readout and the decision margin.

Eq. 6 separates first-order gain from finite-radius curvature. Two representations with similar Jacobians at the anchor can behave differently at deployment scale if their local curvature differs. The guarantee is deterministic when β is a valid bound throughout the stated neighborhood. Estimating curvature from a few additional queries is diagnostic. The interpolation algebra is classical in DFO, while this theorem identifies the subspace-specific representation sensitivity supported by finite queries and translates its reconstruction error into a stability statement. It upper-bounds the finite-query uncertainty. Next, we investigate whether the dependence on curvature and one-sided probe scale is a proof artifact.

Proposition 4.4 (One-Dimensional Sharpness of the Curvature-Scale Term)**.** *Let $d = p = 1$, $x = 0$, $V = [h]$ with $h > 0$, and $\beta > 0$. Define $F_\pm : \mathbb{R} \to \mathbb{R}$ by $F_\pm(t) := \pm\frac{\beta}{2}t(h - t)$. Then, $F_\pm(0) = F_\pm(h) = 0$, each derivative $F'_\pm$ is β-Lipschitz, $F'_+(0) = \frac{\beta h}{2}$, and $F'_-(0) = -\frac{\beta h}{2}$. Consequently, every deterministic estimator $\mathfrak{A}$ that depends only on $\mathscr{Q}_{0,[h]}(F)$ satisfies $\max_{\nu\in\{+,-\}} \left|\mathfrak{A}\big(\mathscr{Q}_{0,[h]}(F_\nu)\big) - F'_\nu(0)\right| \geq \frac{\beta h}{2}$. For the empirical secant estimator, $\widehat{J}_{F_\pm}(0;[h]) = 0$ and $\eta_\beta([h]) = \frac{\beta h}{2}$, so Eq. 4 is attained with equality.*

Proposition 4.4 demonstrates that complete one-dimensional coverage and noiseless endpoint values do not fully eliminate the uncertainty introduced by a finite query scale. Both functions converge to the same endpoint value, which means that a single endpoint-consistency penalty reflects zero discrepancy, even though the derivative at the anchor is nonzero. Consequently, endpoint agreement does not guarantee stability along the entire connecting path within the specified function class. The sharpness claim indicates that the factor of $1/2$ and the linear dependence on βh cannot be uniformly improved for a one-dimensional, one-sided endpoint query under the assumption of a Lipschitz derivative. Furthermore, in stochastic or finite-precision scenarios, making h arbitrarily small can amplify noise, leading to a distinct bias-noise trade-off.

4.3 From representation stability to prediction stability

Theorem 4.3 controls movement in representation space, whereas a classifier acts through the composition $f = g \circ F$. Therefore, a task-level statement must propagate representation movement through the readout and compare the resulting logit change with the decision separation at the anchor.

Corollary 4.5 (Subspace Prediction Stability)**.** *Suppose that the conditions of Theorem 4.3 hold and that V covers the target subspace E, so $E_V = E$. Let $\mathcal{Z}_{x,E,\rho} := F(\mathbb{B}_E(x,\rho)) \subseteq \mathbb{R}^p$. Assume that the readout $g : \mathbb{R}^p \to \mathbb{R}^C$ satisfies $\|g(z) - g(z')\|_\infty \leq L_g\|z - z'\|_2$ for all $z, z' \in \mathcal{Z}_{x,E,\rho}$ and some finite $L_g \geq 0$. Define $B_F(x;V) := \|\widehat{J}_F(x;V)\|_{\mathrm{op}} + \eta_\beta(V)$. Then every $\delta \in E$ with $\|\delta\|_2 \leq \rho$ satisfies $\|f(x+\delta) - f(x)\|_\infty \leq L_g\left(B_F(x;V)\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2\right)$. Consequently, if $c_f(x)$ is unique and $L_g\left(B_F(x;V)\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2\right) < \frac{\mathsf{mar}_f(x)}{2}$, then $c_f(x+\delta) = c_f(x)$. More generally, any $\varepsilon \in (0,\rho]$ satisfying $L_g\left(B_F(x;V)\varepsilon + \frac{\beta}{2}\varepsilon^2\right) < \frac{\mathsf{mar}_f(x)}{2}$ certifies $c_f(x+\delta) = c_f(x)$ for every $\delta \in E$ with $\|\delta\|_2 \leq \varepsilon$.*

Corollary 4.5 identifies three task-dependent factors: the representation displacement bound ($B_F(x;V)\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2$), the readout amplification L_g, and the margin $\mathsf{mar}_f(x)$. Small movement in an intermediate representation is not enough if the readout strongly amplifies it, while a large margin is not enough if local logits vary rapidly. The scalar L_g is generally conservative because it does not exploit alignment between the actual feature displacement and directions to which the readout is insensitive. The conclusion is specific to local and subspace contexts. It preserves the model's current prediction, meaning a misclassified anchor can still receive a certificate indicating the same incorrect class. Evaluations should differentiate between prediction invariance and robust correctness. When applying the result to the logit map, we set $F = f$ and g to the identity, thereby avoiding arbitrary rescaling of an intermediate representation and its inverse transformation. However, a formal certificate requires a valid curvature bound for the entire neighborhood and complete coverage of the claimed perturbation subspace. This margin-sensitivity structure aligns with established robustness

analyses (Sokolić et al., 2017; Tsuzuku et al., 2018). This corollary specializes this concept into a representation bound reconstructed from finite queries within a specified subspace.

4.4 Operator control and Secant-Frame Regularization (SFR)

Corollary 4.5 identifies restricted first-order gain as one contributor to prediction instability. This motivates reducing the empirical secant operator during training. Before defining the objective, we distinguish operator control from independently averaging consistency penalties over the queried directions.

Proposition 4.6 (Pairwise Consistency vs. Worst-Direction Control)**.** *Let $U \in \mathbb{R}^{d\times r}$ have columns $u_1, \dots, u_r$ and satisfy $U^\top U = I_r$. For $h > 0$, set $V = hU$. Assume that $x + hu_j \in \Omega$ for every j, and define*

$$\mathcal{R}_{\mathrm{pair}}(F; x, V) := \frac{1}{r}\sum_{j=1}^{r} \frac{\|F(x+hu_j) - F(x)\|_2^2}{h^2}. \tag{7}$$

Then, $\mathcal{R}_{\mathrm{pair}}(F; x, V) = \frac{1}{r}\|\widehat{J}_F(x; V)\|_{\mathrm{F}}^2$ and $\sqrt{\mathcal{R}_{\mathrm{pair}}(F; x, V)} \le \|\widehat{J}_F(x; V)\|_{\mathrm{op}} \le \sqrt{r\,\mathcal{R}_{\mathrm{pair}}(F; x, V)}$. The right inequality is attained when $\widehat{J}_F(x; V)$ has rank one. If $p \ge r$, the left inequality is attained when it has r equal nonzero singular values. Affine representation maps can realize both cases.

The pairwise score is the mean squared response along the queried orthonormal basis, whereas $\|\widehat{J}_F(x; V)\|_{\mathrm{op}} = \sup_{\substack{u\in E_V \\ \|u\|_2=1}} \|\widehat{J}_F(x; V)u\|_2$ measures the largest reconstructed gain over all unit linear combinations in the covered subspace. At a fixed pairwise score, this worst-direction response can vary by a factor of $\sqrt{r}$ when $p \ge r$. For affine maps, the distinction is for the true restricted Jacobian. For nonlinear maps, it concerns the empirical operator until Theorem 4.3 is invoked. This proposition does not show that pairwise consistency is ineffective: driving the pairwise score to zero also drives the empirical operator norm to zero. It identifies a quantitative difference in emphasis. Average squared control may dilute a concentrated sensitive direction, whereas operator control directly targets the largest reconstructed response. For nonorthogonal frames, the pseudoinverse additionally accounts for directional scaling and dependence, while poor conditioning remains visible through the recovery uncertainty. The comparison is specific to the squared Euclidean fixed-frame objective in Eq. 7. This distinction motivates a task-anchored objective that directly penalizes the empirical operator norm on a chosen perturbation subspace.

Definition 4.7 (SFR)**.** *Let $f_\theta : \Omega \to \mathbb{R}^C$ be the logit map. For each training input x, fix an admissible triple (x, E_x, ρ_x) and choose $V_x = [v_{x,1}, \dots, v_{x,q_x}]$, with $q_x \ge 1$, $v_{x,j} \in E_x \setminus \{\mathbf{0}\}$, $\|v_{x,j}\|_2 \le \rho_x$, and $\sigma^+_{\min}(V_x) \ge \sigma_0$ for a fixed $\sigma_0 > 0$. Set $E_{V_x} := \mathrm{range}(V_x)$. The frame is selected independently of θ or treated as stop-gradient; V_x and $V_x^\dagger$ are held fixed during parameter differentiation. Define*

$$\widehat{S}_\theta(x; V_x) := \|Y_{f_\theta}(x; V_x)V_x^\dagger\|_{\mathrm{op}}. \tag{8}$$

For a finite, nonempty mini-batch $\mathcal{B}$, threshold $\zeta \ge 0$, and weight $\lambda_{\mathrm{SF}} \ge 0$, define $\mathcal{R}_{\mathrm{SF}}(\theta) := |\mathcal{B}|^{-1}\sum_{(x,y)\in\mathcal{B}}\left[\widehat{S}_\theta(x; V_x) - \zeta\right]_+^2$ and $\mathcal{L}(\theta) := \mathcal{L}_{\mathrm{task}}(\theta) + \lambda_{\mathrm{SF}}\mathcal{R}_{\mathrm{SF}}(\theta)$.

Eq. 8 measures the largest reconstructed logit gain on E_{V_x}. Conditioning does not imply coverage: $\sigma^+_{\min}(V_x)$ excludes zero singular values. Applying Corollary 4.5 to the prescribed E_x requires $E_{V_x} = E_x$, equivalently $\mathrm{rank}(V_x) = \dim(E_x)$, together with a valid neighborhood-wide smoothness bound for the final logit map and the stated margin condition. Under Assumption 3.6, Theorem 4.3 connects the score to the true restricted-Jacobian norm. SFR penalizes only the empirical term. Complementary to Jacobian and input-gradient regularization (Jakubovitz & Giryes, 2018; Ross & Doshi-Velez, 2018; Liu et al., 2024), SFR uses finite evaluations and ordinary parameter backpropagation: it is input-Jacobian-free. The threshold permits nonzero empirical gain, but penalizing label-changing directions can suppress useful information despite the task loss. Logit anchoring avoids ambiguity regarding the internal representation and readout rescaling, while the score remains dependent on the overall logit scale and must be interpreted alongside margins and task performance. Because SFR acts on raw logits, it can penalize input-dependent class-independent logit shifts that do not change probabilities or decisions.

5 Experiments

The experiments are categorized into three evidence types. Experiments 1-6 conduct controlled audits of the information hierarchy, operator recovery, and prediction conditions. Experiment 7

applies the certificate to trained smooth classifiers in standardized PCA coordinates. Experiments 8-9 evaluate SFR and representation diagnostics on CIFAR-10, focusing on empirical robustness. In the certificate studies, certified prediction invariance counts anchors that meet the sufficient condition for label preservation, while certified accuracy requires those anchors to be correctly classified. Conditional certificate coverage uses correctly classified anchors as the denominator. For CIFAR-10, evaluated joint robust correctness combines clean accuracy with label preservation under the examined transformations. Comparisons for CIFAR-10 utilize five seeds $\{42, 128, 999, 1022, 3407\}$ and related prepared splits. Implementation details and additional results are provided in Appendix E.

Activation-only non-identifiability (Experiment 1). We apply Theorem 4.1 by introducing a localized smooth term that vanishes on the protected observation set, including the paired views employed by static diagnostics, while enhancing the directional derivative at the anchor. In our analysis of 135 geometries and 810 amplitude-level measurements, we observe a finite-observation mismatch of zero and a maximum invariance/analytic residual of 4.55×10^{-13}. At $\alpha = 1000$, the measured directional sensitivity reaches 1000 with unchanged finite activations. These findings demonstrate the theorem's worst-case information limitation.

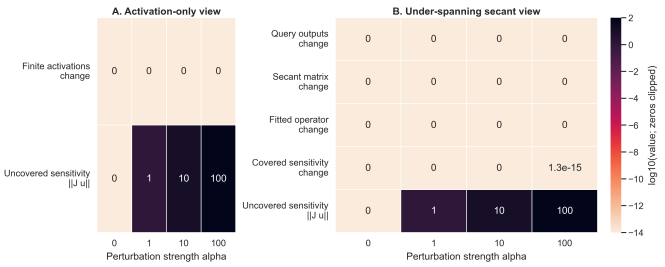

Figure 2: Under-covered secants leave omitted directions unidentified.

Coverage as a necessary condition (Experiment 2). To isolate directional coverage, we use affine maps with an adjustable gain in a direction orthogonal to the query span. Across 350 reported geometries and 1,400 amplitude-level measurements, the original finite observations, secants, and empirical operators remain unchanged within the stated numerical tolerance while the omitted-direction sensitivity increases to 100. Appending the missing direction reveals a recovered response of norm 100, with a maximum invariance/recovery residual of 2.21×10^{-13}. The negative and positive controls in Figure 2 distinguish missing directional information from a failure of the reconstruction algebra. They do not extend the guarantee to directions outside the augmented frame's span.

Recovery and one-sided sharpness (Experiments 3-4). Having established the limitation of under-covered queries, we next test whether covering secants recover local sensitivity within the uncertainty predicted by Theorem 4.3. Experiment 3 uses quadratic representation maps $F(z) = Az + \frac{1}{2}b(z^\top Qz)$, with $Q = Q^\top$ and $\|b\|_2 = 1$. Their Jacobians $J_F(x) = A + b(Qx)^\top$ and valid global smoothness constants $\beta = \|Q\|_{\mathrm{op}}$ are analytically available. This allows a direct comparison between the measured operator error and $\eta_\beta(V)$, as well as a check that the interval in Eq. 5 contains the true sensitivity. Across 24,192 reported frame settings varying probe scale, curvature, rank, redundancy, and conditioning, neither check fails beyond the configured tolerances. These results support the quantitative recovery statement on the tested nonlinear maps: finite-query estimates remain consistent with the true restricted sensitivity after accounting for the prescribed reconstruction uncertainty. Experiment 4 addresses the complementary question of whether the curvature-scale allowance can be attained. Across 125 settings of the scalar construction in Proposition 4.4, the two functions have identical endpoint observations and zero empirical secants, but opposite anchor derivatives of magnitude $\beta h/2$. The measured recovery error equals $\eta_\beta([h]) = \beta h/2$, reproducing the proposition's equality case. For representation analysis, this illustrates how perfect endpoint agreement can conceal nonzero first-order sensitivity, and why the curvature correction cannot be discarded. Symmetric endpoint queries recover the derivative for these quadratics up to numerical tolerance, highlighting that the sharpness claim concerns the specified one-sided observation model.

Average vs. worst-direction response (Experiment 5). We compare affine maps whose queried responses have the same normalized squared Frobenius norm but either an isotropic or a rank-one singular-value spectrum. The pairwise/Frobenius identity has a reported residual of 5.68×10^{-14}, and the ratio of the rank-one to isotropic operator norms follows the predicted $\sqrt{r}$ factor, reaching 5.66 at $r = 32$. This realizes the equality cases of Proposition 4.6: the same nonzero average squared response can coexist with different largest directional gains. Appendix E.6 provides a learned stress test that illustrates the objectives' spectral behavior.

Analytic prediction-certificate audit (Experiment 6). We evaluate Corollary 4.5 on $F(x) = Ax + b_0 + \frac{1}{2}(x^\top Qx)b$, where $b_0, b \in \mathbb{R}^p$, with a linear readout $g(z) = Wz + c$. The bound uses the

valid global constants $\beta = \|b\|_2\|Q\|_{\mathrm{op}}$, and $L_g = \max_k \|W_{k,:}\|_2$. The curvature expression is for symmetric Q. For general Q, it is a valid upper bound because only its symmetric part contributes to $x^\top Qx$. The query frames cover the claimed subspaces, and the 960-anchor audit spans ranks one through four with radius cap $\rho = 0.5$. The mean and median certified radii are 0.26 and 0.24. Independent boundary checks are in the stated 1D cases. Higher-dimensional comparisons use attack-derived upper bounds on adversarial radii. Table 1 illustrates that as the target radius increases, fewer anchors can be certified, but all reported target radii maintain zero observed attack violations, aligning with the corollary.

Table 1: Prediction-certificate test at representative target radii.

Target Radius	Prediction Invariance ↑	Robust Correct ↑	Coverage Given Correct ↑	Attack Violations ↓
0.05	0.88	0.71	0.87	0
0.10	0.75	0.62	0.76	0
0.20	0.56	0.46	0.56	0
0.35	0.34	0.28	0.34	0
0.50	0.21	0.17	0.21	0

Certifications on trained smooth models (Experiment 7). To instantiate Corollary 4.5 beyond analytic constructions, we evaluate one-hidden-layer unthresholded Softplus classifiers trained on MNIST (LeCun et al., 2010) in standardized PCA coordinates. For the specified one-hidden-layer unthresholded Softplus logit map, the certificate uses the analytic bound $\overline{\beta}_f = (\tau/4)\|W_2\|_{\mathrm{op}}\|W_1\|_{\mathrm{op}}\|W_1\|_{2\to\infty}$, derived in Appendix D.8, and the identity readout with $L_g = 1$. Table 15 compares the MNIST PCA-16 models. The evaluations yield positive certified radii and include a comparison with exact-Jacobian regularization. The findings demonstrate a trained-model instantiation of the conditional framework. The bound derivation, certificate results, and evaluation details are provided in Appendix E.8.

Empirical covered-transformation robustness (Experiment 8). We compare ERM, matched-query pairwise consistency, subspace TRADES (Zhang et al., 2019), and SFR on CIFAR-10 (Krizhevsky et al., 2009) with ResNet-18. SFR improves evaluated joint robust correctness over pairwise by 2.46 percentage points (95% CI [2.26,2.64]) and prediction invariance by 2.85 points ([2.63,3.07]), with a clean accuracy decrease of 0.30 points ([-0.50,-0.13]). All five paired seeds favor SFR on the two covered robustness metrics. SFR also has a lower mean secant score and empirical finite-radius worst gain than pairwise. However, subspace TRADES has the highest mean evaluated joint robust correctness. The paired SFR-minus-TRADES effect is -0.73 percentage points with interval [-1.22,-0.30]. Thus, SFR improves the fixed-query consistency baseline but does not uniformly dominate the alternatives. The separation between gain scores and robust accuracy also shows why sensitivity must be interpreted together with task margins and the perturbation protocol. Table 2 shows that SFR improves evaluated covered robustness and lowers measured sensitivity relative to pairwise consistency at a small clean accuracy cost.

Table 2: CIFAR-10 ResNet-18 comparison.

Method	Clean Accuracy ↑	Evaluated Joint Robust Correctness ↑	Prediction Invariance ↑	Secant Norm ↓	Empirical Worst Gain ↓
ERM	95.1[95.0, 95.2]	48.1[46.5, 50.2]	50.8[49.1, 52.9]	8.96	5.69
Pairwise	95.5[95.4, 95.6]	82.6[82.3, 82.9]	86.3[85.9, 86.6]	1.36	1.29
Subspace TRADES	95.3[95.1, 95.4]	85.8[85.5, 86.1]	89.7[89.4, 90.0]	4.10	2.05
SFR	95.2[95.1, 95.3]	85.1[84.9, 85.3]	89.1[88.9, 89.3]	1.13	1.05

6 Conclusion and Limitation

This work develops a finite-query framework that distinguishes between descriptive embedding geometry and certifiable local representation stability. We identify the limitations of observations based on activation and incomplete directional queries. Furthermore, we show that secant frames, combined with independently established smoothness, can support representation bounds and conditions for prediction invariance through readout control and margins. This analysis introduces SFR, which penalizes the largest reconstructed input-to-logit gain. Controlled experiments with trained models illustrate the benefits and trade-offs of operator-based regularization within this conditional framework. The guarantees are local, metric-dependent, and restricted to the covered perturbation subspace. They rely on regularity information that finite queries alone do not certify and do not automatically extend to nonsmooth models or unrestricted perturbations. The empirical evidence collected is limited to the evaluated models and perturbation settings, and it does not establish universal superiority over alternative objectives. Broader impacts are provided in Appendix F.

REFERENCES

András Balogh and Márk Jelasity. On the functional similarity of robust and non-robust neural representations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1614–1635. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/balogh23a.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8799–8810. Curran Associates, Inc., 2022a. doi: 10.52202/068431-0640. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39cee562b91611c16ac0b100f0bc1ea1-Paper-Conference.pdf.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022b. URL https://arxiv.org/abs/2105.04906.

Lukas Braun, Erin Grant, and Andrew M Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5355–5382. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/braun25a.html.

Coralia Cartis and Lindon Roberts. A derivative-free gauss–newton method. *Mathematical Programming Computation*, 11(4):631–674, 2019.

Coralia Cartis and Lindon Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization: C. cartis, l. roberts. *Mathematical Programming*, 199(1):461–524, 2023.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets, 2025. URL https://arxiv.org/abs/2411.05561.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/cisse17a.html.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, 2009. doi: 10.1137/1.9780898718768. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898718768.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/croce20b.html.

D. Davar and G. N. Grapiglia. Trfd: A derivative-free trust-region method based on finite differences for composite nonsmooth optimization. *SIAM Journal on Optimization*, 35(3):1792–1821, 2025. doi: 10.1137/24M1701587. URL https://doi.org/10.1137/24M1701587.

Taha Entesari, Sina Sharifi, and Mahyar Fazlyab. Compositional curvature bounds for deep neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Sarah E Harvey, David Lipshutz, and Alex H Williams. What representational similarity measures imply about decodable information. In Marco Fumero, Clementine Domine, Zorah Lähner, Donato Crisostomi, Luca Moschella, and Kimberly Stachenfeld (eds.), *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 140–151. PMLR, 14 Dec 2024. URL https://proceedings.mlr.press/v285/harvey24a.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Kai Hu, Andy Zou, Zifan Wang, Klas Leino, and Matt Fredrikson. Unlocking deterministic robustness certification on imagenet. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 42993–43011. Curran Associates, Inc., 2023. doi: 10.52202/075280-1863. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/863da9d40547f1d1b18859519ce2dee4-Paper-Conference.pdf.

Kai Hu, Haoqi Hu, and Matt Fredrikson. Lipnext: Scaling up lipschitz-based certified robustness to billion-parameter models, 2026. URL https://arxiv.org/abs/2601.18513.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J. Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22745–22757. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c055dcc749c2632fd4dd806301f05ba6-Paper.pdf.

Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 525–541, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01258-8.

Haydn T. Jones, Jacob M. Springer, Garrett T. Kenyon, and Juston S. Moore. If you've trained one you've trained them all: inter-architecture similarity increases with robustness. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 928–937. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/jones22a.html.

Taeyoung Kim and Hongseok Yang. An infinite-width analysis on the Jacobian-regularised training of a neural network. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24584–24657. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kim24ah.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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

Henry Kvinge, Grayson Jorgenson, Davis Brown, Charles Godfrey, and Tegan Emerson. Internal representations of vision models through the lens of frames on data manifolds. In Sophia Sanborn,

Christian Shewmake, Simone Azeglio, and Nina Miolane (eds.), *Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 228 of *Proceedings of Machine Learning Research*, pp. 75–115. PMLR, 16 Dec 2024. URL https://proceedings.mlr.press/v228/kvinge24a.html.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 535–557, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3485258. URL https://doi.org/10.1145/3460120.3485258.

Deyin Liu, Lin Yuanbo Wu, Bo Li, Farid Boussaid, Mohammed Bennamoun, Xianghua Xie, and Chengwu Liang. Jacobian norm with selective input gradient regularization for interpretable adversarial defense. *Pattern Recognition*, 145:109902, 2024. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109902. URL https://www.sciencedirect.com/science/article/pii/S0031320323006003.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning, 2018. URL https://arxiv.org/abs/1704.03976.

Vedant Nanda, Till Speicher, Camila Kolling, John P Dickerson, Krishna Gummadi, and Adrian Weller. Measuring representational robustness of neural networks through shared invariances. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16368–16382. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nanda22a.html.

Tianyu Pan and Damon L. Woodard. GeoLAN: Geometric learning of latent explanatory directions in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5493–5521, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.271. URL https://aclanthology.org/2026.findings-acl.271/.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0defd533d51ed0a10c5c9dbf93ee78a5-Paper.pdf.

Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/d1f44e2f09dc172978a4d3151d11d63e-Paper.pdf.

Lindon Roberts. Model construction for convex-constrained derivative-free optimization. *SIAM Journal on Optimization*, 35(2):622–650, 2025. doi: 10.1137/24M1649113. URL https://doi.org/10.1137/24M1649113.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Emanuele Sansone, Tim Lebailly, and Tinne Tuytelaars. Collapse-proof non-contrastive self-supervised learning, 2025. URL https://arxiv.org/abs/2410.04959.

Christopher Scarvelis and Justin Solomon. Nuclear norm regularization for deep learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop - a formalism for specifying selected invariances in an adaptive network. In J. Moody, S. Hanson, and R.P. Lippmann (eds.), *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/65658fde58ab3c2b6e5132a39fae7cb9-Paper.pdf.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8981–8991. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/singla20a.html.

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: examining model stitching. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *Trans. Sig. Proc.*, 65(16):4265–4280, August 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2708039. URL https://doi.org/10.1109/TSP.2017.2708039.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018. URL https://arxiv.org/abs/1703.01780.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: scalable certification of perturbation invariance for deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6542–6551, Red Hook, NY, USA, 2018. Curran Associates Inc.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Theresa Wasserer, Maximilian Fleissner, and Debarghya Ghoshdastidar. Pac-bayesian analysis of the surrogate relation between joint embedding and supervised downstream losses. In Matus Telgarsky and Jonathan Ullman (eds.), *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–33. PMLR, 23–26 Feb 2026. URL https://proceedings.mlr.press/v313/wasserer26a.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

Shufei Zhang, Zhuang Qian, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Xinping Yi. Towards better robust generalization with shift consistency regularization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12524–12534. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhang21p.html.

Appendices

A Declaration: Use of LLMs **17**

B More Related Work **17**

B.1 Static Representation Geometry and Collapse 17
B.2 Representational Similarity vs. Functional Behavior 17
B.3 Perturbation-Aware Geometry and Neural Frames 18
B.4 Jacobian and Input-Gradient Regularization 18
B.5 Lipschitz, Margin, and Curvature-Based Robustness 18
B.6 Derivative-Free Interpolation and Finite-Difference Models 18
B.7 Consistency and Finite-Perturbation Regularization 19
B.8 Randomized Smoothing and Transformation-Specific Certificates 19
B.9 Deterministic Lipschitz Certification 19

C Notation Summary **20**

D Proofs **22**

D.1 Supporting Results for the Problem Setup 22
D.2 Proof of Theorem 4.1 23
D.3 Proof of Proposition 4.2 24
D.4 Proof of Theorem 4.3 25
D.5 Proof of Proposition 4.4 26
D.6 Proof of Corollary 4.5 27
D.7 Proof of Proposition 4.6 28
D.8 Analytic Constants Used in the Certificate Experiments 29

E Additional Experimental Details **32**

E.1 Setup and Metrics 32
E.2 Experiment 1: Activation-Only Non-identifiability 32
E.3 Experiment 2: Uncovered-Direction Non-identifiability 33
E.4 Experiment 3: Coverage-Aware Secant Recovery 33
E.5 Experiment 4: Sharpness of the Curvature-Scale Term 34
E.6 Experiment 5: Pairwise Consistency vs. Worst-Direction Operator Control 34
E.7 Experiment 6: Analytic Prediction-Certificate Audit 35
E.8 Experiment 7: Certified Neural Bridge on Real Data 36
E.9 Experiment 8: CIFAR-10 Covered Transformation Training 37
E.10 Experiment 9: Checkpoint-Zoo Diagnostic Evidence 38
E.11 Summary of Experimental Evidence 38

E.12 Computing Hardware 39

F Limitation and Broader Impact **40**

A Declaration: Use of LLMs

We restricted LLMs to aiding or polishing writing, while human authors conceived all novel theoretical and empirical claims.

B More Related Work

Our work focuses on the intersection of representation analysis, local robustness, finite-difference approximation, and consistency regularization. These areas address complementary aspects of learned representations, including how finite sets of embeddings are organized, how representations respond to perturbations, how to approximate local derivatives from function evaluations, and how to control sensitivity during training.

B.1 Static Representation Geometry and Collapse

A large body of literature examines neural representations using statistics derived from activations based on a limited set of inputs. Representation-similarity methods, such as canonical correlation analysis (CCA) and centered kernel alignment (CKA), evaluate how activations are organized across different layers, architectures, and training runs (Kornblith et al., 2019). Alignment and uniformity describe, respectively, the closeness of positive-pair embeddings and the overall distribution of normalized features on the hypersphere (Wang & Isola, 2020). Effective-rank criteria, such as RankMe (Garrido et al., 2023), use the singular-value spectrum of an activation matrix as an unsupervised indicator of downstream representation quality. Additional research links representation-similarity measures to the information recoverable through linear readouts (Harvey et al., 2024). While these methods offer valuable insights into the geometry of the observed embeddings, they focus primarily on finite activation matrices rather than the behavior of the underlying representation map in neighborhoods around the observed inputs.

The closely related objectives shape the geometry of finite samples during training. Barlow Twins (Zbontar et al., 2021) encourages agreement between augmented views while reducing redundancy across coordinates. In contrast, VICReg (Bardes et al., 2022b;a) combines invariance, variance, and covariance terms to avoid constant or dimensionally collapsed solutions. More recent research has focused on analyzing and preventing various forms of collapse (Sansone et al., 2025) and on establishing theoretical relationships between joint-embedding pretext objectives and downstream losses under specific assumptions (Wasserer et al., 2026). However, neural collapse shows that concentration is not always problematic. Late-stage supervised training can reduce within-class variability and produce a highly structured geometry of class means and classifiers (Papyan et al., 2020). Therefore, high effective rank or broad dispersion should not be universally interpreted as indicators of representation quality. In this paper, we do not argue that static geometric statistics are uninformative, nor that contraction is necessarily undesirable. We ask a narrower question: whether finite activation values alone determine a deterministic upper bound on local directional sensitivity.

B.2 Representational Similarity vs. Functional Behavior

Many studies have shown that representational similarity and functional behavior are not interchangeable concepts. STIR (Nanda et al., 2022) enhances traditional similarity analysis by incorporating perturbations that preserve the reference network's representation. This approach measures shared invariances rather than evaluating agreement based on the original samples. Analyses of both robust and non-robust networks have shown that representational relationships depend on the behavioral criterion considered (Jones et al., 2022; Balogh & Jelasity, 2023). Recent research has further highlighted this limitation. For instance, Ciernik et al. (2025) shows that representational similarity across models can vary by dataset, suggesting that its relationship with task performance is also dataset-dependent. Braun et al. (2025) draws an analytical distinction between representational and functional similarity, noting that similar representations may produce different functions, or that different representations may realize similar functions. Additionally, Smith et al. (2025) demonstrates that successful model stitching does not necessarily mean that the stitched representations convey equivalent information. These findings emphasize the need for perturbation-aware analysis but address questions different from those this paper explores. While representational similarity methods compare two systems or two activation spaces, this paper focuses on the activation-only non-identifiability result. This result concerns how much local sensitivity information a single representation map determines when evaluated on a finite set.

B.3 Perturbation-Aware Geometry and Neural Frames

The closest conceptual antecedents to our study examine how a network transforms specific perturbation directions. Tangent propagation (Simard et al., 1991) defines desired invariance by controlling output derivatives along known transformation directions. Meanwhile, the Manifold Tangent Classifier (Rifai et al., 2011) estimates local tangent directions and promotes insensitivity along the learned data manifold. Kvinge et al. (2024) introduces neural frames constructed from perturbation vectors associated with points on a data manifold and examines how these vectors change as they pass through vision models. This perspective aligns with our motivation, as it views representation geometry as the movement of meaningful local directions rather than the arrangement of activation points. However, our paper differs in both objectives and inferential scope. We treat finite endpoint evaluations as an observation model and investigate which restricted Jacobian these observations identify. This approach allows us to clearly distinguish the target subspace (i.e., the span actually covered by the query frame) from the conditioning of that frame.

B.4 Jacobian and Input-Gradient Regularization

Local derivatives have long been utilized to characterize or control the sensitivity of neural networks. Tangent-based methods penalize derivatives along selected invariance directions (Simard et al., 1991; Rifai et al., 2011). Input-gradient regularization penalizes the sensitivity of a scalar loss or prediction score to input perturbations (Ross & Doshi-Velez, 2018). In contrast, Jacobian regularization directly penalizes the norms of a vector-valued input-output Jacobian (Jakubovitz & Giryes, 2018). Parseval-style constraints control layer operators by restricting network weights (Cisse et al., 2017). Recent research (Kim & Yang, 2024) analyzed Jacobian-regularized training in the infinite-width limit and experimentally investigated the relevance of their results to wide finite networks. New methods explored alternative Jacobian geometries. Specifically, Jacobian nuclear-norm regularization (Scarvelis & Solomon, 2024) encourages locally low-rank mappings and develops scalable approximations for representation learning applications. These methods highlight the significance of local functional sensitivity, but they differ from the finite-query problem addressed in this paper. An exact Jacobian penalty assumes access to derivatives and does not require a query framework. Hence, its validity is not constrained by the span or conditioning of finite perturbation directions. In contrast, this paper reconstructs the empirical secant operator from function evaluations and only approximates the Jacobian within the queried subspace. The proposed regularizer should also be seen as complementary to exact or stochastic Jacobian regularization.

B.5 Lipschitz, Margin, and Curvature-Based Robustness

A related body of work establishes prediction guarantees based on bounds concerning network sensitivity and decision margins. Lipschitz-margin methods combine a bound on output variation with the separation between the winning class and competing classes (Tsuzuku et al., 2018; Sokolić et al., 2017). Local-linearity regularization minimizes nonlinear variations in the loss around training examples and was partly developed to prevent gradient obfuscation during efficient adversarial training (Qin et al., 2019). Second-order certificates utilize bounds on the Hessian or curvature to enhance first-order robustness analysis (Singla & Feizi, 2020). More recently, compositional methods have derived scalable analytical curvature bounds for continuously differentiable deep networks, applying these bounds for certification and regularization (Entesari et al., 2024). This paper adheres to the established principle of margin sensitivity, aiming to translate a finite-query representation bound into a sufficient condition at the task level. The distinction here lies in the source and scope of the sensitivity estimate: the recovered operator is conditioned on finite secant observations and applies only to the covered perturbation subspace.

B.6 Derivative-Free Interpolation and Finite-Difference Models

Model-based derivative-free optimization (DFO) has a well-established theory for constructing local models based on function evaluations. It effectively manages the geometry of interpolation sets and controls model error within trust regions (Conn et al., 2009). In nonlinear least-squares optimization, the DFO-GN method constructs linear interpolation models of vector-valued residuals and uses them to build a local Gauss-Newton model (Cartis & Roberts, 2019). Subspace variants optimize these models in either user-defined or randomly chosen low-dimensional subspaces, reducing the number of required interpolation points and the associated computational cost (Cartis & Roberts, 2023). Recent research has continued to improve the construction of feasible interpolation sets (Roberts, 2025) and has developed finite-difference trust-region methods for vector-valued black-box maps with Lipschitz-continuous Jacobians (Davar & Grapiglia, 2025). The key distinction between DFO

and our work lies in how we treat the local model. In DFO, the model serves as an algorithmic tool for selecting a step and demonstrating convergence to a stationary point. In contrast, this paper treats the model as an observation operator used to infer the local sensitivity of a learned representation or predictor.

B.7 Consistency and Finite-Perturbation Regularization

Consistency regularization encourages a model to produce compatible outputs when faced with perturbations, stochastic augmentations, or predictions from a teacher model. Mean Teacher utilizes an exponential moving average of model parameters to establish consistency targets (Tarvainen & Valpola, 2018), while virtual adversarial training penalizes alterations along locally adversarial perturbations (Miyato et al., 2018). Similarly, Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022b) enforce agreement among augmented views while incorporating mechanisms to prevent representational collapse. In adversarial learning, shift-consistency regularization manages discrepancies between latent shifts induced by both natural and adversarial inputs (Zhang et al., 2021). These methods illustrate that finite pairs of perturbations can serve as effective learning signals, even without a local certificate. This paper uses the general idea of model evaluations on perturbed inputs but approaches aggregation differently. For an orthonormal frame, the average squared pairwise response equals the normalized squared Frobenius norm of the empirical secant operator, while the proposed score focuses on its operator norm. The average squared response measures the response across the queried basis, whereas the operator norm identifies the largest reconstructed gain from all unit linear combinations within that span. This distinction is quantitatively significant. Lowering the pairwise score towards zero will also diminish the empirical operator norm. For nonorthogonal frames, the pseudoinverse considers the directions and scales represented by the query set, with its sensitivity to poor conditioning becoming evident in the recovery error. This paper also distinguishes itself from finite-difference methods that regularize parameter-space gradients or approximate Hessian-gradient products. Although these methods use similar computational tools, they focus on optimization geometry in parameter space rather than the operator mapping from input to representation or input to logit within a specified perturbation subspace.

B.8 Randomized Smoothing and Transformation-Specific Certificates

Cohen et al. (2019) certify a classifier defined by the most probable base-classifier prediction under Gaussian input noise. Their guarantee covers an entire ambient ℓ_2 ball around an individual input without requiring differentiability of the base classifier. When the relevant class probabilities are bounded exactly, the prediction-invariance statement is deterministic; practical Monte Carlo certification instead establishes the required probability bounds at a prescribed confidence level. The resulting confidence concerns the certification procedure, not merely the fraction of perturbations on which the classifier behaves correctly. Our certificate concerns the specified map $f = g \circ F$, rather than replacing its decision rule by a smoothed classifier, and uses exact finite responses together with independently established regularity and readout bounds. Transformation-specific smoothing also addresses structured perturbation domains: TSS develops certificates for semantic transformations and accounts for interpolation effects in its treatment of differentially resolvable transformations (Li et al., 2021). Accordingly, transformation-coordinate certification is not unique to our formulation. The distinction is the finite-query reconstruction and its conditional error bound, not the mere choice of a low-dimensional transformation family.

B.9 Deterministic Lipschitz Certification

Hu et al. (2023) introduce LiResNet residual blocks designed to admit tighter, efficiently computed Lipschitz bounds, together with Efficient Margin Maximization (EMMA), a loss using Lipschitz-based bounds on adverse logit-margin changes across competing classes. They demonstrate Lipschitz-based certification on ImageNet. More recently, LipNeXt combines orthogonal parameterization and spatial-shift operations in a scalable Lipschitz-controlled architecture (Hu et al., 2026). These approaches obtain sensitivity bounds from network structure and parameters. A global Lipschitz bound is input-uniform, but the resulting certified radius still depends on the prediction margin at the individual input. Therefore, global does not mean prediction invariance throughout input space. Moreover, local Lipschitz certification already exists, including bounds exploiting activation behavior within an input neighborhood (Huang et al., 2021). Our contribution is not the margin-sensitivity principle or locality itself. It is the characterization of the restricted first-order response supported by a finite query record, conditional on separate control of Jacobian variation.

C NOTATION SUMMARY

Notation	Meaning
Core spaces, maps, and norms	
$\mathbb{R}^d, \mathbb{R}^p, \mathbb{R}^C$	Input, representation, and logit spaces; d, p, and C are their dimensions.
$\Omega \subseteq \mathbb{R}^d$	Open input domain on which the representation or predictor is defined.
x, δ	Anchor input and perturbation vector.
$F : \Omega \to \mathbb{R}^p$	Representation map. In certificate experiments this may also be taken as the logit map.
$g : \mathbb{R}^p \to \mathbb{R}^C$	Readout map from representation space to logits.
$f = g \circ F$	Final predictor/logit map.
f_θ	Parametric logit map used in the SFR training objective.
$\|\cdot\|_2$	Euclidean vector norm.
$\|\cdot\|_\infty$	Maximum absolute component of a vector; used on logits in prediction bounds and on intermediate vectors in the analytic-constant calculations.
$\|A\|_{\mathrm{op}}$	Spectral/operator norm of a matrix A.
$\|A\|_{\mathrm{F}}$	Frobenius norm of a matrix A.
$A^\dagger$	Moore–Penrose pseudoinverse of A.
$\sigma^+_{\min}(A)$	Smallest nonzero singular value of a nonzero matrix A.
Directional neighborhoods and finite observations	
$E \subseteq \mathbb{R}^d$	Claimed perturbation subspace. All stability claims are restricted to this subspace unless explicitly stated otherwise.
$\rho > 0$	Perturbation radius for a local directional neighborhood.
$\mathbb{B}_E(x, \rho)$	Directional neighborhood $\{x + \delta : \delta \in E, \|\delta\|_2 \le \rho\}$.
(x, E, ρ) admissible	The directional neighborhood $\mathbb{B}_E(x, \rho)$ lies inside Ω.
P_E	Orthogonal projection onto E.
$J_F(x)$	Jacobian of F at x.
$\mathsf{ASens}_\rho(F; x, E)$	Anchored finite-scale directional sensitivity of F around x on E.
$\mathsf{DSens}(F; x, E)$	Infinitesimal directional sensitivity $\|J_F(x)P_E\|_{\mathrm{op}}$.
$c_f(x)$	Predicted class $\arg\max_{1 \le c \le C} f_c(x)$, assumed unique when the margin is used.
$\mathsf{mar}_f(x)$	Prediction margin $f_{c_f(x)}(x) - \max_{c \ne c_f(x)} f_c(x)$.
$\mathcal{D}_N$	Finite dataset $\{(x_i, y_i)\}_{i=1}^N$.
$\mathscr{O}_{\mathcal{D}_N}(F)$	Finite activation observation $(F(x_1), \dots, F(x_N))$.
$\mathsf{S}, \Psi_{\mathcal{D}_N}$	Activation-only statistic and the map that computes it from finite observations.
$F \equiv_{\mathcal{D}_N} \widetilde{F}$	Observational equivalence on $\mathcal{D}_N$: the two maps agree on all dataset inputs.
Secant frames and covered recovery	
$V = [v_1, \dots, v_q] \in \mathbb{R}^{d \times q}$	Finite perturbation query frame at anchor x; q is the number of queried directions.
v_j	Individual perturbation direction, with $v_j \in E \setminus \{\mathbf{0}\}$ and $\|v_j\|_2 \le \rho$.
E_V	Covered subspace $\mathrm{range}(V)$.
V covers E	Coverage condition $E_V = E$.
$\mathscr{Q}_{x,V}(F)$	Local query observation $(F(x), F(x + v_1), \dots, F(x + v_q))$.
$Y_F(x; V)$	Endpoint-difference matrix $[F(x + v_1) - F(x), \dots, F(x + v_q) - F(x)]$.
$\mathfrak{S}_F(x; V)$	Local secant frame $(V, Y_F(x; V))$.
$\widehat{J}_F(x; V)$	Empirical secant operator $Y_F(x; V)V^\dagger$.
β	Directional Jacobian-Lipschitz/curvature bound in the theory. Distinct from the TRADES hyperparameter.
$\eta_\beta(V)$	Curvature-conditioning uncertainty $\frac{\beta}{2\sigma^+_{\min}(V)} \left(\sum_{j=1}^q \|v_j\|_2^4\right)^{1/2}$.
$[a]_+$	Positive part $\max\{a, 0\}$.
ρ'	Secondary radius used when converting the recovery bound into finite-scale sensitivity.
Prediction certificate and operator-vs-pairwise comparison	
$\mathcal{Z}_{x,E,\rho}$	Representation image $F(\mathbb{B}_E(x, \rho))$ used to state a readout Lipschitz condition.
L_g	Lipschitz constant of the readout g on $\mathcal{Z}_{x,E,\rho}$ from representation movement to logit movement.
$B_F(x; V)$	Certificate shorthand $\|\widehat{J}_F(x; V)\|_{\mathrm{op}} + \eta_\beta(V)$.
ε	Candidate certification or empirical evaluation radius. For Corollary 4.5, it satisfies $0 < \varepsilon \le \rho$ and yields a certificate only when the stated strict margin condition holds.
$U = [u_1, \dots, u_r]$	Orthonormal basis frame used to compare pairwise consistency and operator control.
r	Rank or covered dimension in operator-vs-pairwise comparisons; proof-local radii use decorated variants such as r_i.
h	Probe scale for an orthonormal or one-dimensional frame.
$\mathcal{R}_{\mathrm{pair}}(F; x, V)$	Average squared pairwise endpoint response over the queried orthonormal directions.
$F_\pm$	One-dimensional sharpness construction attaining the curvature-scale error.
$\mathfrak{A}$	Generic deterministic estimator in the one-dimensional sharpness proposition.
SFR learning objective	
E_x, ρ_x, V_x	Training-time anchor-specific perturbation subspace, radius, and query frame.
q_x	Number of training-time query directions for anchor x.
σ_0	Lower bound imposed on $\sigma^+_{\min}(V_x)$ for selected nonzero singular values.
E_{V_x}	Covered subspace $\mathrm{range}(V_x)$ for the training-time frame.
$\widehat{S}_\theta(x; V_x)$	SFR score $\|Y_{f_\theta}(x; V_x)V_x^\dagger\|_{\mathrm{op}}$, the reconstructed largest logit gain.

Continued on next page

Notation	Meaning
$\mathcal{B}$	Mini-batch used to define the empirical SFR penalty.
ζ	Nonnegative threshold below which the SFR score is not penalized.
λ_{SF}	Weight on the SFR regularizer.
$\mathcal{R}_{\mathrm{SF}}(\theta)$	Mini-batch SFR penalty $\lvert\mathcal{B}\rvert^{-1}\sum_{(x,y)\in\mathcal{B}}[\widehat{S}_\theta(x;V_x)-\zeta]_+^2$.
$\mathcal{L}_{\mathrm{task}}(\theta)$	Standard task loss before adding SFR.
$\mathcal{L}(\theta)$	Total training objective $\mathcal{L}_{\mathrm{task}}(\theta)+\lambda_{\mathrm{SF}}\mathcal{R}_{\mathrm{SF}}(\theta)$.
Construction, proof, and experiment constants	
α	Amplitude parameter in the activation-only non-identifiability construction and Experiment 1.
ψ, u, a	Local cutoff, input direction, and output direction used in the constructive activation-only perturbation.
A, b, Q	Linear/quadratic-map parameters in recovery and certificate experiments; in the quadratic audit, $\beta=\lVert Q\rVert_{\mathrm{op}}$ when $\lVert b\rVert_2=1$.
W, c	Linear readout weights and bias in readout-norm calculations.
W_1, W_2, τ	Softplus classifier weights and temperature used for analytic curvature constants.
L	Generic local Lipschitz or induced-norm constant. Use L_g for the certificate readout constant.
$\overline{\beta}_{\mathrm{quad}}, \beta^\star_{\mathrm{quad}}, \overline{\beta}_f$	Valid curvature constants for the quadratic-map and Softplus-logit certificate calculations.
λ	Generic regularization weight in some experimental configurations; use λ_{SF} for the SFR objective.
β_{TRADES}	TRADES loss hyperparameter.
$d_a, \mathcal{A}, a$	Dimension, domain $\mathcal{A}\subseteq\mathbb{R}^{d_a}$, and coordinates of a transformation or standardized PCA-coordinate model.
ϕ_x, H_x	Input transformation ϕ_x with $\phi_x(0)=x$ and composite representation $H_x(a)=F(\phi_x(a))$.
ρ_{S}	Spearman rank correlation in the checkpoint analysis.
$\gamma>0$	Common scalar amplitude in the affine equality-case constructions, for which the pairwise score is γ^2.

D PROOFS

D.1 SUPPORTING RESULTS FOR THE PROBLEM SETUP

We establish the auxiliary facts used in Section 3: the infinitesimal limit of anchored sensitivity, the minimum-norm characterization of secant fitting, and the directional Taylor remainder. The first result requires only differentiability at the anchor, the second is purely linear-algebraic, and the third invokes Assumption 3.6. All dimensions below are positive integers, and I_m denotes the identity on $\mathbb{R}^m$.

Lemma D.1 (Infinitesimal Limit of Anchored Directional Sensitivity)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be open, let $F : \Omega \to \mathbb{R}^p$, and fix an admissible triple (x, E, ρ_0). If F is Fréchet differentiable at x, then $\mathsf{ASens}_\rho(F; x, E)$ is finite for all sufficiently small $\rho > 0$, and*

$$\lim_{\rho \downarrow 0} \mathsf{ASens}_\rho(F; x, E) = \|J_F(x) P_E\|_{\mathrm{op}} = \mathsf{DSens}(F; x, E). \tag{9}$$

Proof. Set $J := J_F(x)$ and $\ell := \|JP_E\|_{\mathrm{op}}$. Since P_E is a contraction and is the identity on E, $\ell = \max_{u \in E,\, \|u\|_2 = 1} \|Ju\|_2$. Indeed, projecting any ambient unit vector gives a vector in the unit ball of E, while every unit vector in E is unchanged by the projection. Homogeneity gives the equality of the suprema. The maximum is attained because E is a nonzero finite-dimensional subspace and its unit sphere is compact. Choose a maximizing unit vector $u_\star \in E$. For $\delta \in E$ with $\|\delta\|_2 \le \rho_0$, write $e_x(\delta) := F(x + \delta) - F(x) - J\delta$. Fréchet differentiability implies that, for every $\epsilon > 0$, there exists $\rho_\epsilon \in (0, \rho_0]$ such that $\|e_x(\delta)\|_2 \le \epsilon \|\delta\|_2$ for all $\delta \in E$ with $0 < \|\delta\|_2 \le \rho_\epsilon$. The same ρ_ϵ applies simultaneously to all directions in E. Fix $0 < \rho \le \rho_\epsilon$. For every admissible nonzero δ in the radius-ρ ball, the triangle inequality gives $\|F(x + \delta) - F(x)\|_2 \le (\ell + \epsilon)\|\delta\|_2$. For the particular displacement $\delta = \rho u_\star$, the reverse triangle inequality gives $\|F(x + \rho u_\star) - F(x)\|_2 \ge \rho(\ell - \epsilon)$. Taking the supremum in the definition of anchored sensitivity therefore yields

$$\ell - \epsilon \le \mathsf{ASens}_\rho(F; x, E) \le \ell + \epsilon, \qquad 0 < \rho \le \rho_\epsilon. \tag{10}$$

Eq. 10 establishes local finiteness. Since $\epsilon > 0$ is arbitrary, it also proves Eq. 9. □

Lemma D.1 justifies the limit asserted in Definition 3.2 without requiring continuous differentiability around the anchor. It is a qualitative shrinking-radius statement: it supplies no explicit error bound at a prescribed nonzero radius. The finite-scale estimates below use the additional directional smoothness assumption. We next determine the linear operator selected by a finite query record. The following result permits redundant or rank-deficient frames and arbitrary response matrices. Exact interpolation is not assumed.

Lemma D.2 (Pseudoinverse Geometry and Minimum-Norm Secant Fitting)**.** *Let $V \in \mathbb{R}^{d \times q}$ be nonzero, set $E_V := \mathrm{range}(V)$, and let $Y \in \mathbb{R}^{p \times q}$ be arbitrary. Define $\widehat{A} := YV^\dagger$. Then*

$$VV^\dagger = P_{E_V}, \qquad \|V^\dagger\|_{\mathrm{op}} = \frac{1}{\sigma^+_{\min}(V)}, \qquad \widehat{A} = \widehat{A} P_{E_V}. \tag{11}$$

Writing $\mathcal{M}(V, Y) := \arg\min_{A \in \mathbb{R}^{p \times d}} \|AV - Y\|_{\mathrm{F}}$, the full set of least-squares minimizers is

$$\mathcal{M}(V, Y) = \big\{ \widehat{A} + Z(I_d - P_{E_V}) : Z \in \mathbb{R}^{p \times d} \big\}. \tag{12}$$

Every minimizer satisfies $AP_{E_V} = \widehat{A}$, and $\widehat{A}$ is the unique minimum-Frobenius-norm member of $\mathcal{M}(V, Y)$.

Proof. Let $r := \mathrm{rank}(V) \ge 1$ and take a compact singular value decomposition $V = U_r \Sigma_r W_r^\top$, where $U_r \in \mathbb{R}^{d \times r}$ and $W_r \in \mathbb{R}^{q \times r}$ have orthonormal columns and $\Sigma_r = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \ge \cdots \ge \sigma_r > 0$. The Moore–Penrose inverse is $V^\dagger = W_r \Sigma_r^{-1} U_r^\top$. Consequently, $VV^\dagger = U_r U_r^\top = P_{E_V}$, and the nonzero singular values of $V^\dagger$ are the reciprocals of those of V. Thus $\|V^\dagger\|_{\mathrm{op}} = 1/\sigma_r = 1/\sigma^+_{\min}(V)$. Moreover, $V^\dagger P_{E_V} = V^\dagger$, so $\widehat{A} P_{E_V} = \widehat{A}$. This proves Eq. 11. Define $\Pi_V := V^\dagger V = W_r W_r^\top$, the orthogonal projection in $\mathbb{R}^q$ onto $\mathrm{range}(V^\top)$. For every $A \in \mathbb{R}^{p \times d}$, $AV - Y = (AV - Y\Pi_V) - Y(I_q - \Pi_V)$. The two terms are orthogonal in the Frobenius

inner product: the rows of $AV - Y\Pi_V$ lie in $\mathrm{range}(V^\top)$, whereas the rows of $Y(I_q - \Pi_V)$ lie in its orthogonal complement. Hence

$$\|AV - Y\|_{\mathrm{F}}^2 = \|AV - Y\Pi_V\|_{\mathrm{F}}^2 + \|Y(I_q - \Pi_V)\|_{\mathrm{F}}^2. \tag{13}$$

Since $\widehat{A}V = Y\Pi_V$, Eq. 13 shows that the minimum residual norm is $\|Y(I_q - \Pi_V)\|_{\mathrm{F}}$ and that A is a minimizer exactly when $AV = Y\Pi_V$. Minimizing the norm and its square gives the same set of minimizers. Write $A = \widehat{A} + H$. The minimizing condition is equivalent to $HV = 0$. By Eq. 11, $HV = 0$ implies $HP_{E_V} = HVV^\dagger = 0$. Conversely, $HP_{E_V} = 0$ implies $HV = 0$ because $P_{E_V}V = V$. Thus A is a minimizer exactly when $H = H(I_d - P_{E_V})$, giving Eq. 12 and $AP_{E_V} = \widehat{A}$. Finally, for any such H, the rows of $\widehat{A} = \widehat{A}P_{E_V}$ and $H = H(I_d - P_{E_V})$ lie in orthogonal subspaces. Therefore $\|\widehat{A} + H\|_{\mathrm{F}}^2 = \|\widehat{A}\|_{\mathrm{F}}^2 + \|H\|_{\mathrm{F}}^2$. This expression is minimized uniquely at $H = 0$, proving the minimum-norm assertion. □

With $Y = Y_F(x; V)$, Lemma D.2 justifies the empirical operator in Definition 3.5. Its zero action on $E_V^\perp$ is a minimum-norm fitting convention, not evidence that the representation map is insensitive there. An exact fit exists only when $Y = YV^\dagger V$; otherwise the secants are fitted in the least-squares sense. If the queried responses satisfy $Y = A_0V$ for some linear map A_0, then Eq. 11 gives $\widehat{A} = A_0P_{E_V}$, establishing exact recovery on the covered subspace in that case. The preceding lemma concerns only the finite linear fitting problem. To relate its observations to a nonlinear representation map, we now establish the remainder estimate supplied by local directional smoothness.

Lemma D.3 (Directional Taylor Remainder)**.** *Let (x, E, ρ) be admissible and suppose that F satisfies Assumption 3.6 with constant $\beta \geq 0$. For every $v \in E$ with $\|v\|_2 \leq \rho$, define $e_x(v) := F(x + v) - F(x) - J_F(x)v$. Then*

$$e_x(v) = \int_0^1 \big(J_F(x + tv) - J_F(x)\big)v\, dt, \tag{14}$$

$$\|e_x(v)\|_2 \leq \frac{\beta}{2}\|v\|_2^2. \tag{15}$$

Proof. If $v = 0$, both assertions hold because $e_x(0) = 0$. Otherwise, for every $t \in [0, 1]$, the point $x + tv$ belongs to $\mathbb{B}_E(x, \rho)$ since $tv \in E$ and $\|tv\|_2 \leq \|v\|_2 \leq \rho$. The assumed continuous differentiability on an open neighborhood of this set implies that $t \mapsto F(x + tv)$ is continuously differentiable along the entire segment, with derivative $J_F(x + tv)v$. The fundamental theorem of calculus, applied componentwise, gives $F(x + v) - F(x) = \int_0^1 J_F(x + tv)v\, dt$. Subtracting $J_F(x)v = \int_0^1 J_F(x)v\, dt$ proves Eq. 14. Because $v = P_E v$, the smoothness assumption yields $\|(J_F(x + tv) - J_F(x))v\|_2 \leq \beta t\|v\|_2^2$ for every $t \in [0, 1]$. Taking norms in Eq. 14 and integrating this upper bound gives $\|e_x(v)\|_2 \leq \beta\|v\|_2^2 \int_0^1 t\, dt = (\beta/2)\|v\|_2^2$, which proves Eq. 15. The argument also covers $\beta = 0$. □

Lemma D.3 applies to every admissible displacement in E, not only the queried columns. It uses the directional Jacobian-difference bound and does not require a Hessian or quantitative control of Jacobian variation in transverse directions. The constant β remains an input supplied by Assumption 3.6. The lemma does not estimate or certify it from endpoint observations.

D.2 Proof of Theorem 4.1

We construct a localized perturbation that preserves all observed activations but is affine near the selected anchor. Its constant Jacobian on that neighborhood changes the first-order gain without changing Jacobian differences.

Proof. Fix the dataset $\mathcal{D}_N$, the index i, the nonzero subspace E, $F_0 \in C^1(\Omega; \mathbb{R}^p)$, and $M > 0$ as in the theorem. The observed inputs belong to Ω, as stipulated in Section 3.2.

Localization. Since Ω is open and the observed inputs are pairwise distinct, choose $R_i > 0$ such that $\mathbb{B}_{\mathbb{R}^d}(x_i, R_i) \subseteq \Omega$ and $\|x_j - x_i\|_2 > R_i$ for every $j \neq i$. Such a choice is possible because x_i

is an interior point and, when $N > 1$, its distances to the finitely many other inputs have a positive minimum. When $N = 1$, the latter condition is vacuous.

For an explicit smooth cutoff, define $\omega(t) := \exp(-1/t)$ for $t > 0$ and $\omega(t) := 0$ for $t \le 0$. This function is C^∞ on $\mathbb{R}$: all of its derivatives on $t > 0$ tend to zero as $t \downarrow 0$, being finite sums of powers of $1/t$ multiplied by $\exp(-1/t)$. Define $\chi(s) := \omega(1-s)/(\omega(1-s) + \omega(s-1/4))$ for $s \in \mathbb{R}$. The denominator is strictly positive for every s, so χ is smooth, takes values in $[0,1]$, equals one for $s \le 1/4$, and equals zero for $s \ge 1$. Set $\psi_i(z) := \chi(\|z - x_i\|_2^2 / R_i^2)$ for $z \in \mathbb{R}^d$. Then ψ_i is smooth, equals one whenever $\|z - x_i\|_2 \le R_i/2$, and vanishes whenever $\|z - x_i\|_2 \ge R_i$. In particular, its support is a compact subset of Ω. Let $r_i := R_i/4$. The function ψ_i is identically one on an open neighborhood of $\mathbb{B}_{\mathbb{R}^d}(x_i, r_i)$, and this closed ball is contained in Ω.

Preservation of the finite observations. Choose unit vectors $u \in E$ and $a \in \mathbb{R}^p$, which exist because $E \neq \{0\}$ and $p \ge 1$. For each $\alpha \ge 0$, define the map on Ω by

$$F^{(\alpha)}(z) := F_0(z) + \alpha \psi_i(z) \langle u, z - x_i \rangle a. \tag{16}$$

The added term is smooth, so $F^{(\alpha)} \in C^1(\Omega; \mathbb{R}^p)$. At $z = x_i$, it vanishes because $\langle u, x_i - x_i \rangle = 0$. At every other observed input x_j, it vanishes because $\|x_j - x_i\|_2 > R_i$ implies $\psi_i(x_j) = 0$. Thus Eq. 16 gives $F^{(\alpha)}(x_j) = F_0(x_j)$ for every j, and hence $F^{(\alpha)} \equiv_{\mathcal{D}_N} F_0$ for every $\alpha \ge 0$.

Arbitrarily large directional sensitivity. On an open neighborhood of $\mathbb{B}_{\mathbb{R}^d}(x_i, r_i)$, the added term in Eq. 16 is the affine map $z \mapsto \alpha \langle u, z - x_i \rangle a$. Consequently,

$$J_{F^{(\alpha)}}(z) = J_{F_0}(z) + \alpha a u^\top, \qquad z \in \mathbb{B}_{\mathbb{R}^d}(x_i, r_i). \tag{17}$$

Choose $\alpha_M := M + \|J_{F_0}(x_i)u\|_2 + 1$ and set $F_M := F^{(\alpha_M)}$. Since $P_E u = u$ and $\|u\|_2 = \|a\|_2 = 1$, Eq. 17 and the reverse triangle inequality imply

$$\begin{aligned} \mathsf{DSens}(F_M; x_i, E) &= \|J_{F_M}(x_i) P_E\|_{\mathrm{op}} \\ &\ge \|J_{F_M}(x_i) u\|_2 \\ &= \|J_{F_0}(x_i) u + \alpha_M a\|_2 \\ &\ge \alpha_M - \|J_{F_0}(x_i) u\|_2 = M + 1 > M. \end{aligned} \tag{18}$$

Eq. 18 proves the required sensitivity bound, while the preceding observation argument gives $F_M \equiv_{\mathcal{D}_N} F_0$.

Preservation of local Jacobian differences. For any $z, z' \in \mathbb{B}_{\mathbb{R}^d}(x_i, r_i)$, subtracting Eq. 17 at these two points with $\alpha = \alpha_M$ cancels the constant matrix $\alpha_M a u^\top$. This proves Eq. 3 for every such pair. The radius r_i was chosen before α_M and can therefore be kept fixed as M varies. All conclusions hold for the arbitrary baseline F_0 and arbitrary $M > 0$, completing the proof. □

D.3 Proof of Proposition 4.2

An affine perturbation suffices when the query frame omits a subspace direction: the perturbation can vanish at every queried input while adding a constant Jacobian in an unqueried direction. This also allows preservation of an existing Jacobian-difference bound on its original neighborhood.

Proof. Fix x, the frame V, the subspaces $G := \mathrm{range}(V) \subsetneq E$, a map $F : \Omega \to \mathbb{R}^p$ differentiable at x, and $M > 0$ as in the proposition. Since V is a perturbation frame at x, all queried points $x + v_j$ belong to Ω.

Choice of an unqueried direction. Choose $w \in E \setminus G$ and let P_G denote the orthogonal projection onto G. The vector $w - P_G w$ is nonzero, belongs to E because $G \subseteq E$, and is orthogonal to G. Thus $u := (w - P_G w)/\|w - P_G w\|_2$ is a unit vector in $E \cap G^\perp$. In particular, $P_E u = u$ and $\langle u, v_j \rangle = 0$ for every j.

Construction and preservation of the query record. Choose a unit vector $a \in \mathbb{R}^p$ and set $\lambda_M := M + \|J_F(x)u\|_2 + 1$. Define

$$\widetilde{F}_M(z) := F(z) + \lambda_M \langle u, z - x \rangle a, \qquad z \in \Omega. \tag{19}$$

The added term in Eq. 19 vanishes at $z = x$ and at every $z = x + v_j$, since $\langle u, 0\rangle = 0$ and $\langle u, v_j\rangle = 0$. Consequently, $\widetilde{F}_M(x) = F(x)$ and $\widetilde{F}_M(x + v_j) = F(x + v_j)$ for every j. By Definition 3.5, $\mathscr{Q}_{x,V}(\widetilde{F}_M) = \mathscr{Q}_{x,V}(F)$.

Arbitrarily large sensitivity on the target subspace. The added term in Eq. 19 is affine on $\mathbb{R}^d$, with constant Jacobian $\lambda_M a u^\top$. Therefore $\widetilde{F}_M$ is differentiable at x and $J_{\widetilde{F}_M}(x) = J_F(x) + \lambda_M a u^\top$. Using $P_E u = u$, $\|u\|_2 = \|a\|_2 = 1$, and the reverse triangle inequality, we obtain

$$\begin{aligned}\mathsf{DSens}(\widetilde{F}_M; x, E) &= \|J_{\widetilde{F}_M}(x) P_E\|_{\mathrm{op}} \\ &\geq \|J_{\widetilde{F}_M}(x) u\|_2 \\ &= \|J_F(x) u + \lambda_M a\|_2 \\ &\geq \lambda_M - \|J_F(x) u\|_2 = M + 1 > M. \end{aligned} \tag{20}$$

Eq. 20 proves the required strict inequality. This part of the argument uses only differentiability of F at x.

Preservation of the Jacobian-difference bound. Now suppose additionally that F is continuously differentiable on the neighborhood $\mathcal{U} \subseteq \Omega$ specified in the proposition. Since the added term is affine, $\widetilde{F}_M$ is continuously differentiable on the same neighborhood, and $J_{\widetilde{F}_M}(z) = J_F(z) + \lambda_M a u^\top$ for every $z \in \mathcal{U}$. Hence

$$\begin{aligned} J_{\widetilde{F}_M}(z) - J_{\widetilde{F}_M}(z') &= \big(J_F(z) + \lambda_M a u^\top\big) - \big(J_F(z') + \lambda_M a u^\top\big) \\ &= J_F(z) - J_F(z'), \qquad z, z' \in \mathcal{U}. \end{aligned} \tag{21}$$

Eq. 21 preserves the Jacobian differences as matrices, not merely their norms. Thus any existing bound on these differences, including one after a fixed directional projection, holds with exactly the same constants and on the same neighborhood $\mathcal{U}$. No shrinking of $\mathcal{U}$ is required. This establishes the final assertion and completes the proof. □

D.4 Proof of Theorem 4.3

We first bound the nonlinear remainder in each observed secant, then propagate these remainders through the pseudoinverse on the actual query span. The resulting operator estimate yields both the infinitesimal sensitivity interval and the finite-scale bounds for arbitrary admissible displacements in that span.

Proof. Fix (x, E, ρ), F, $\beta \geq 0$, and V satisfying the hypotheses of the theorem. Write $J := J_F(x)$, $Y := Y_F(x; V)$, and $\widehat{J} := \widehat{J}_F(x; V) = YV^\dagger$. Since V is nonzero and its columns belong to E, $E_V := \mathrm{range}(V)$ is a nonzero subspace of E. Consequently, $\mathbb{B}_{E_V}(x, \rho) \subseteq \mathbb{B}_E(x, \rho) \subseteq \Omega$. We do not assume $E_V = E$ or that the columns of V are linearly independent.

Columnwise remainders. For each $j \in \{1, \ldots, q\}$, define $r_j := F(x + v_j) - F(x) - Jv_j \in \mathbb{R}^p$. The segment $x + tv_j$, $0 \leq t \leq 1$, lies in $\mathbb{B}_E(x, \rho)$ because $v_j \in E$ and $\|v_j\|_2 \leq \rho$. Lemma D.3 therefore applies and gives $\|r_j\|_2 \leq (\beta/2)\|v_j\|_2^2$. With $R := [r_1, \ldots, r_q] \in \mathbb{R}^{p \times q}$, the columnwise identities imply

$$Y = JV + R. \tag{22}$$

The definition of the Frobenius norm and the remainder bounds give

$$\begin{aligned} \|R\|_{\mathrm{F}} &= \left(\sum_{j=1}^{q} \|r_j\|_2^2\right)^{1/2} \\ &\leq \frac{\beta}{2}\left(\sum_{j=1}^{q} \|v_j\|_2^4\right)^{1/2}. \end{aligned} \tag{23}$$

Recovery of the restricted Jacobian. By Lemma D.2, $VV^\dagger = P_{E_V}$ and $\|V^\dagger\|_{\mathrm{op}} = 1/\sigma^+_{\min}(V)$, where $\sigma^+_{\min}(V) > 0$ is the smallest nonzero singular value. Using Eq. 22, we obtain

$$\begin{aligned}\widehat{J} - JP_{E_V} &= YV^\dagger - JVV^\dagger \\ &= (Y - JV)V^\dagger = RV^\dagger. \end{aligned} \tag{24}$$

Apply submultiplicativity of the operator norm, $\|R\|_{\mathrm{op}} \le \|R\|_{\mathrm{F}}$, and Eq. 23 to Eq. 24:

$$\begin{aligned}\|\widehat{J} - JP_{E_V}\|_{\mathrm{op}} &\le \|R\|_{\mathrm{op}}\|V^\dagger\|_{\mathrm{op}} \\ &\le \frac{\|R\|_{\mathrm{F}}}{\sigma^+_{\min}(V)} \\ &\le \frac{\beta}{2\,\sigma^+_{\min}(V)} \left(\sum_{j=1}^{q} \|v_j\|_2^4\right)^{1/2} = \eta_\beta(V). \end{aligned} \tag{25}$$

Eq. 25 proves Eq. 4. The argument uses the pseudoinverse identities for an arbitrary nonzero matrix, so it also applies to redundant and rank-deficient frames.

Two-sided infinitesimal sensitivity. The reverse triangle inequality for the operator norm and Eq. 25 yield

$$\left|\|\widehat{J}\|_{\mathrm{op}} - \|JP_{E_V}\|_{\mathrm{op}}\right| \le \|\widehat{J} - JP_{E_V}\|_{\mathrm{op}} \le \eta_\beta(V). \tag{26}$$

Since $\mathsf{DSens}(F; x, E_V) = \|JP_{E_V}\|_{\mathrm{op}} \ge 0$, Eq. 26 implies

$$\left[\|\widehat{J}\|_{\mathrm{op}} - \eta_\beta(V)\right]_+ \le \mathsf{DSens}(F; x, E_V) \le \|\widehat{J}\|_{\mathrm{op}} + \eta_\beta(V). \tag{27}$$

This proves Eq. 5.

Finite-scale displacement. Fix any $\delta \in E_V$ with $\|\delta\|_2 \le \rho$. For $\delta = 0$, the displacement bound holds because both sides are zero. Otherwise, since $E_V \subseteq E$, the same Lemma D.3 applies to δ, regardless of whether it was queried. Writing $e_x(\delta) := F(x+\delta) - F(x) - J\delta$, we have $\|e_x(\delta)\|_2 \le (\beta/2)\|\delta\|_2^2$. Moreover, $P_{E_V}\delta = \delta$. Using the upper bound in Eq. 27, we conclude

$$\begin{aligned}\|F(x+\delta) - F(x)\|_2 &\le \|J\delta\|_2 + \|e_x(\delta)\|_2 \\ &\le \|JP_{E_V}\|_{\mathrm{op}}\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2 \\ &\le \left(\|\widehat{J}\|_{\mathrm{op}} + \eta_\beta(V)\right)\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2. \end{aligned} \tag{28}$$

Eq. 28 is Eq. 6, and it holds for every admissible $\delta \in E_V$, not only the columns of V.

Anchored finite-scale sensitivity. Fix $0 < \rho' \le \rho$. The triple (x, E_V, ρ') is admissible because $\mathbb{B}_{E_V}(x, \rho') \subseteq \mathbb{B}_E(x, \rho) \subseteq \Omega$. For every $\delta \in E_V$ with $0 < \|\delta\|_2 \le \rho'$, dividing Eq. 28 by $\|\delta\|_2$ gives $\|F(x+\delta) - F(x)\|_2/\|\delta\|_2 \le \|\widehat{J}\|_{\mathrm{op}} + \eta_\beta(V) + (\beta/2)\|\delta\|_2$. The right-hand side is at most $\|\widehat{J}\|_{\mathrm{op}} + \eta_\beta(V) + (\beta/2)\rho'$ because $\beta \ge 0$. Taking the supremum over these displacements in Eq. 1 therefore yields $\mathsf{ASens}_{\rho'}(F; x, E_V) \le \|\widehat{J}\|_{\mathrm{op}} + \eta_\beta(V) + (\beta/2)\rho'$, which is the final assertion. The proof never divides by β and therefore includes $\beta = 0$. □

D.5 Proof of Proposition 4.4

We compare two smooth scalar maps with identical endpoint observations but opposite derivatives at the anchor. This yields an estimator-independent lower bound, which the empirical secant operator attains for both maps.

Proof. Fix $h > 0$ and $\beta > 0$, and let $F_\pm(t) := \pm(\beta/2)t(h - t)$ for $t \in \mathbb{R}$, as in the proposition. We identify 1×1 matrices with their scalar entries when computing pseudoinverses and operator norms.

Regularity and indistinguishable observations. Both maps are polynomials and hence belong to $C^\infty(\mathbb{R})$. Differentiating gives

$$F'_+(t) = \frac{\beta h}{2} - \beta t, \qquad F'_-(t) = -\frac{\beta h}{2} + \beta t, \qquad t \in \mathbb{R}. \tag{29}$$

Eq. 29 implies $|F'_\nu(s) - F'_\nu(t)| = \beta|s-t|$ for every $s, t \in \mathbb{R}$ and $\nu \in \{+,-\}$. Thus both derivatives are globally β-Lipschitz, and $F'_+(0) = \beta h/2$ while $F'_-(0) = -\beta h/2$. Direct substitution gives $F_\nu(0) = F_\nu(h) = 0$ for both signs. By Definition 3.5, their query records satisfy

$$\mathscr{Q}_{0,[h]}(F_+) = (0,0) = \mathscr{Q}_{0,[h]}(F_-). \tag{30}$$

The frame $V = [h]$ spans $E_V = \mathbb{R}$. Taking $\Omega = E = \mathbb{R}$ and any $\rho \geq h$ gives an admissible triple $(0, \mathbb{R}, \rho)$ for which both maps satisfy Assumption 3.6. Hence they also meet the hypotheses of the recovery theorem.

Lower bound for every endpoint-based estimator. Fix any deterministic real-valued estimator $\mathfrak{A}$ whose dependence on the unknown function is only through $\mathscr{Q}_{0,[h]}(F)$. The known parameters h and β may be available to the estimator; they are fixed and identical for the two candidates. Eq. 30 forces the same estimate $a_{\mathfrak{A}} := \mathfrak{A}((0,0))$ for both maps. The triangle inequality and Eq. 29 at the anchor give

$$\begin{aligned} \beta h &= |F'_+(0) - F'_-(0)| \\ &\leq |F'_+(0) - a_{\mathfrak{A}}| + |a_{\mathfrak{A}} - F'_-(0)| \\ &\leq 2 \max_{\nu \in \{+,-\}} \left|\mathfrak{A}\big(\mathscr{Q}_{0,[h]}(F_\nu)\big) - F'_\nu(0)\right|. \end{aligned} \tag{31}$$

Dividing Eq. 31 by two yields $\max_{\nu\in\{+,-\}} |\mathfrak{A}(\mathscr{Q}_{0,[h]}(F_\nu)) - F'_\nu(0)| \geq \beta h/2$, which proves the estimator-independent claim.

Attainment by the empirical secant operator. Since $h > 0$, the pseudoinverse is $[h]^\dagger = [h^{-1}]$, $\sigma^+_{\min}([h]) = h$, and $P_{E_V} = 1$. The common endpoint record gives $Y_{F_\nu}(0;[h]) = 0$ and therefore $\widehat{J}_{F_\nu}(0;[h]) = 0$ for both signs. Moreover, the definition of the curvature–conditioning term gives $\eta_\beta([h]) = (\beta/(2h))(h^4)^{1/2} = \beta h/2$. As the scalar Jacobian is $J_{F_\nu}(0) = F'_\nu(0)$, Eq. 29 yields

$$\begin{aligned} \left\| \widehat{J}_{F_\nu}(0;[h]) - J_{F_\nu}(0) P_{E_V} \right\|_{\mathrm{op}} &= |0 - F'_\nu(0)| \\ &= \frac{\beta h}{2} = \eta_\beta([h]), \qquad \nu \in \{+,-\}. \end{aligned} \tag{32}$$

Eq. 32 establishes equality in Eq. 4 for both maps, completing the proof. □

D.6 Proof of Corollary 4.5

Theorem 4.3 controls representation displacement on the covered subspace. We propagate this bound through the readout, then compare each perturbed class gap with the margin at the anchor to obtain pointwise and uniform prediction invariance.

Proof. Fix F, g, x, E, ρ, V, $\beta \geq 0$, and $L_g \geq 0$ satisfying the hypotheses of the corollary, and write $f = g \circ F$. Set $B := B_F(x;V) = \|\widehat{J}_F(x;V)\|_{\mathrm{op}} + \eta_\beta(V)$, so B is finite and nonnegative. By the coverage assumption, $E_V = E$.

Logit displacement. Fix $\delta \in E$ with $\|\delta\|_2 \leq \rho$. Admissibility gives $x, x+\delta \in \mathbb{B}_E(x,\rho) \subseteq \Omega$. Thus $F(x)$ and $F(x+\delta)$ both belong to $\mathcal{Z}_{x,E,\rho} = F(\mathbb{B}_E(x,\rho))$, where the assumed readout Lipschitz bound applies. Moreover, $E_V = E$ permits us to apply Eq. 6 to this δ. Combining the two bounds yields

$$\begin{aligned} \|f(x+\delta) - f(x)\|_\infty &= \|g(F(x+\delta)) - g(F(x))\|_\infty \\ &\leq L_g \|F(x+\delta) - F(x)\|_2 \\ &\leq L_g \left(B\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2 \right). \end{aligned} \tag{33}$$

Eq. 33 proves the first assertion for every admissible δ, including $\delta = 0$. Only the readout bound on $\mathcal{Z}_{x,E,\rho}$ is used; no extension of this bound to other representation points is required.

Pointwise prediction invariance. Now suppose that the maximizing class $c^\star := c_f(x)$ is unique, and let $m := \mathsf{mar}_f(x) > 0$ be its margin as in Definition 3.3. For every competing class $c \neq c^\star$, $f_{c^\star}(x) - f_c(x) \geq m$. For the fixed admissible δ, write $\Delta(\delta) := f(x+\delta) - f(x) \in \mathbb{R}^C$, with $\Delta_c(\delta)$ denoting its cth component. For each competing class,

$$\begin{aligned} f_{c^\star}(x+\delta) - f_c(x+\delta) &= f_{c^\star}(x) - f_c(x) + \Delta_{c^\star}(\delta) - \Delta_c(\delta) \\ &\geq m - |\Delta_{c^\star}(\delta)| - |\Delta_c(\delta)| \\ &\geq m - 2\|\Delta(\delta)\|_\infty. \end{aligned} \tag{34}$$

If $L_g(B\|\delta\|_2 + (\beta/2)\|\delta\|_2^2) < m/2$, Eq. 33 implies $\|\Delta(\delta)\|_\infty < m/2$. The right-hand side of Eq. 34 is therefore strictly positive for every $c \neq c^\star$. Hence $c^\star$ remains the unique maximizing class at $x+\delta$, proving $c_f(x+\delta) = c_f(x)$.

Uniform prediction invariance on the closed directional ball. Fix $\varepsilon \in (0, \rho]$ satisfying $L_g(B\varepsilon + (\beta/2)\varepsilon^2) < m/2$. For every $\delta \in E$ with $\|\delta\|_2 \leq \varepsilon$, Eq. 33 and the nonnegativity of L_g, B, and β give

$$\begin{aligned} \|\Delta(\delta)\|_\infty &\leq L_g\left(B\|\delta\|_2 + \frac{\beta}{2}\|\delta\|_2^2\right) \\ &\leq L_g\left(B\varepsilon + \frac{\beta}{2}\varepsilon^2\right) < \frac{m}{2}. \end{aligned} \tag{35}$$

Eq. 35, together with Eq. 34, proves $c_f(x+\delta) = c_f(x)$ simultaneously for all $\delta \in E$ with $\|\delta\|_2 \leq \varepsilon$. The boundary $\|\delta\|_2 = \varepsilon$ is included because the condition at ε is strict. No step divides by L_g, β, or B, so their allowed zero values require no separate argument. This proves all assertions of the corollary. □

D.7 Proof of Proposition 4.6

For a common-scale orthonormal frame, the pairwise score is the normalized squared Frobenius norm of the reconstructed operator. We compare this quantity with its largest singular value and construct affine maps realizing both equality cases.

Proof. Fix $x \in \Omega$, $U \in \mathbb{R}^{d\times r}$ with $1 \leq r \leq d$ and $U^\top U = I_r$, and $h > 0$ as in the proposition. Set $V := hU$ and assume that every queried point $x + hu_j$ belongs to Ω. Write $Y := Y_F(x; V) \in \mathbb{R}^{p\times r}$ and $\widehat{J} := \widehat{J}_F(x; V) = YV^\dagger$. Only the stated finite evaluations of F are needed in the argument; no differentiability assumption is imposed.

The pairwise–Frobenius identity. Since $U^\top U = I_r$ and $h > 0$, the factorization $V = U(hI_r)I_r^\top$ is a compact singular value decomposition. Thus $V^\dagger = h^{-1}U^\top$. Consequently,

$$\begin{aligned} \widehat{J} = h^{-1}YU^\top, \qquad \widehat{J}U = h^{-1}Y, \\ \widehat{J}\widehat{J}^\top = h^{-2}YU^\top UY^\top = h^{-2}YY^\top. \end{aligned} \tag{36}$$

Taking traces in Eq. 36 and using the column definition of Y gives

$$\begin{aligned} \|\widehat{J}\|_{\mathrm{F}}^2 &= \mathrm{tr}(\widehat{J}\widehat{J}^\top) = h^{-2}\,\mathrm{tr}(YY^\top) \\ &= \frac{1}{h^2}\sum_{j=1}^{r}\|F(x+hu_j) - F(x)\|_2^2 = r\,\mathcal{R}_{\mathrm{pair}}(F; x, V). \end{aligned} \tag{37}$$

Here tr denotes the matrix trace, and the final equality uses Eq. 7. Dividing Eq. 37 by $r > 0$ proves the asserted identity.

Operator-norm inequalities and equality conditions. Eq. 36 implies $\mathrm{rank}(\widehat{J}) \leq \min\{p, r\}$. If $\widehat{J} = 0$, Eq. 37 gives $\mathcal{R}_{\mathrm{pair}}(F; x, V) = 0$, so both claimed inequalities hold with equality. Otherwise, let $k := \mathrm{rank}(\widehat{J}) \geq 1$ and denote its nonzero singular values by $\sigma_1 \geq \cdots \geq \sigma_k > 0$. Then $\|\widehat{J}\|_{\mathrm{op}} = \sigma_1$,

$\|\widehat{J}\|_{\mathrm{F}}^2 = \sum_{j=1}^k \sigma_j^2$, and

$$\sigma_1^2 \le \sum_{j=1}^k \sigma_j^2 \le k\sigma_1^2 \le r\sigma_1^2. \tag{38}$$

Combining Eq. 38 with Eq. 37 and taking nonnegative square roots yields

$$\sqrt{\mathcal{R}_{\mathrm{pair}}(F;x,V)} \le \|\widehat{J}\|_{\mathrm{op}} \le \sqrt{r\,\mathcal{R}_{\mathrm{pair}}(F;x,V)}. \tag{39}$$

For a nonzero operator, the right equality in Eq. 39 holds exactly when $\sum_{j=1}^k \sigma_j^2 = \sigma_1^2$, equivalently $k = 1$. The left equality holds exactly when $\sum_{j=1}^k \sigma_j^2 = r\sigma_1^2$. By Eq. 38, this requires $k = r$ and $\sigma_1 = \cdots = \sigma_r > 0$, and these conditions also suffice. In particular, this nonzero equality case requires $p \ge r$. The previously treated zero operator attains both equalities for every allowed p and r.

Affine realizations with a common pairwise score. For any $B \in \mathbb{R}^{p\times r}$ and a fixed $b \in \mathbb{R}^p$, define $F_B(z) := b + BU^\top(z-x)$ for $z \in \Omega$. This is the restriction of an affine map on $\mathbb{R}^d$. Writing e_j for the jth standard basis vector of $\mathbb{R}^r$, we have $U^\top u_j = e_j$ and hence $F_B(x+hu_j) - F_B(x) = hBe_j$. Thus $Y_{F_B}(x;V) = hB$, $\widehat{J}_{F_B}(x;V) = BU^\top$, and $J_{F_B}(x) = BU^\top$. Because $(BU^\top)(BU^\top)^\top = BB^\top$, the matrices $BU^\top$ and B have the same nonzero singular values. It follows that

$$\mathcal{R}_{\mathrm{pair}}(F_B;x,V) = \frac{1}{r}\|B\|_{\mathrm{F}}^2, \qquad \|\widehat{J}_{F_B}(x;V)\|_{\mathrm{op}} = \|B\|_{\mathrm{op}}. \tag{40}$$

Fix any $\gamma > 0$ and choose unit vectors $a \in \mathbb{R}^p$ and $w \in \mathbb{R}^r$. Set $B_{\mathrm{one}} := \sqrt{r}\,\gamma a w^\top$. This matrix has rank one with its sole nonzero singular value $\sqrt{r}\,\gamma$. Eq. 40 therefore gives $\mathcal{R}_{\mathrm{pair}}(F_{B_{\mathrm{one}}};x,V) = \gamma^2$ and $\|\widehat{J}_{F_{B_{\mathrm{one}}}}(x;V)\|_{\mathrm{op}} = \sqrt{r}\,\gamma$. This realizes the right equality in Eq. 39 for every $p \ge 1$.

When $p \ge r$, choose $Q \in \mathbb{R}^{p\times r}$ with $Q^\top Q = I_r$ and set $B_{\mathrm{iso}} := \gamma Q$. Its r nonzero singular values all equal γ, so Eq. 40 gives

$$\begin{aligned} &\mathcal{R}_{\mathrm{pair}}(F_{B_{\mathrm{one}}};x,V) = \mathcal{R}_{\mathrm{pair}}(F_{B_{\mathrm{iso}}};x,V) = \gamma^2, \\ &\|\widehat{J}_{F_{B_{\mathrm{one}}}}(x;V)\|_{\mathrm{op}} = \sqrt{r}\,\gamma, \qquad \|\widehat{J}_{F_{B_{\mathrm{iso}}}}(x;V)\|_{\mathrm{op}} = \gamma. \end{aligned} \tag{41}$$

Eq. 41 realizes both equalities in Eq. 39 at the same positive pairwise score and with the same x, U, and h. The ratio of the two operator norms is $\sqrt{r}$; for $r = 1$, the two bounds coincide. This proves all assertions of the proposition. □

D.8 Analytic Constants Used in the Certificate Experiments

We derive the analytic regularity and readout constants used in the prediction-certificate experiments: the quadratic representation with a linear readout in Experiment 6, and the trained Softplus logit map in Experiment 7. The quadratic-map calculation also justifies the curvature constant used in the secant-recovery audit in Experiment 3. We distinguish exact global constants from sufficient upper bounds and derive each for the specified real-arithmetic model. The quadratic calculation is explicit. We retain the symmetric part of its matrix so that the statement also covers a quadratic form whose matrix has not been assumed symmetric.

Lemma D.4 (Jacobian-Lipschitz Constant of the Quadratic Map)**.** *Let $A \in \mathbb{R}^{p\times d}$, $b_0, b \in \mathbb{R}^p$, and $Q \in \mathbb{R}^{d\times d}$, where $d, p \ge 1$. Define $F : \mathbb{R}^d \to \mathbb{R}^p$ by $F(z) := Az + b_0 + \frac{1}{2}(z^\top Qz)b$, and set $Q_{\mathrm{s}} := (Q + Q^\top)/2$. Then $F \in C^\infty(\mathbb{R}^d;\mathbb{R}^p)$ and*

$$J_F(z) = A + b(Q_{\mathrm{s}}z)^\top, \qquad z \in \mathbb{R}^d, \tag{42}$$

$$\|J_F(z) - J_F(z')\|_{\mathrm{op}} = \|b\|_2\|Q_{\mathrm{s}}(z-z')\|_2, \qquad z, z' \in \mathbb{R}^d. \tag{43}$$

Its smallest global Jacobian-Lipschitz constant is

$$\beta_{\mathrm{quad}}^\star := \sup_{z\ne z'} \frac{\|J_F(z) - J_F(z')\|_{\mathrm{op}}}{\|z - z'\|_2} = \|b\|_2\|Q_{\mathrm{s}}\|_{\mathrm{op}}. \tag{44}$$

In particular, $\overline{\beta}_{\mathrm{quad}} := \|b\|_2\|Q\|_{\mathrm{op}}$ is a valid global upper bound and equals $\beta_{\mathrm{quad}}^\star$ when $Q = Q^\top$. Either constant can be used in Assumption 3.6 on any directional neighborhood in $\mathbb{R}^d$.

Proof. Since $Q - Q^\top$ is skew-symmetric, $z^\top(Q - Q^\top)z = 0$ for every z. Thus $z^\top Qz = z^\top Q_{\mathrm{s}} z$. The gradient of $\frac{1}{2}z^\top Q_{\mathrm{s}} z$ is $Q_{\mathrm{s}} z$, so differentiation proves Eq. 42. Subtracting that identity at two points yields $J_F(z) - J_F(z') = b(Q_{\mathrm{s}}(z - z'))^\top$. For any vectors b and w, the rank-one matrix $bw^\top$ has operator norm $\|b\|_2\|w\|_2$: the upper bound follows from Cauchy–Schwarz, and, when both vectors are nonzero, equality is attained on the unit vector $w/\|w\|_2$. If either vector is zero, the identity is immediate. This proves Eq. 43.

By Eq. 43, every nonzero displacement gives a Jacobian-difference quotient at most $\|b\|_2\|Q_{\mathrm{s}}\|_{\mathrm{op}}$. If this product is positive, choose a unit vector u satisfying $\|Q_{\mathrm{s}}u\|_2 = \|Q_{\mathrm{s}}\|_{\mathrm{op}}$, which exists by compactness of the unit sphere. The pair $z = u$, $z' = 0$ attains that quotient. If the product is zero, all Jacobian differences vanish and the smallest nonnegative constant is zero. This establishes Eq. 44. Moreover, $\|Q_{\mathrm{s}}\|_{\mathrm{op}} \le (\|Q\|_{\mathrm{op}} + \|Q^\top\|_{\mathrm{op}})/2 = \|Q\|_{\mathrm{op}}$, so $\overline{\beta}_{\mathrm{quad}}$ is valid for arbitrary Q and is exact when Q is symmetric.

For any nonzero subspace $E \subseteq \mathbb{R}^d$, $\|P_E\|_{\mathrm{op}} = 1$ and submultiplicativity gives

$$\begin{aligned}\|(J_F(z) - J_F(z'))P_E\|_{\mathrm{op}} &\le \|J_F(z) - J_F(z')\|_{\mathrm{op}} \\ &\le \beta^\star_{\mathrm{quad}}\|z - z'\|_2, \qquad z, z' \in \mathbb{R}^d. \end{aligned} \tag{45}$$

Together with global smoothness, Eq. 45 verifies Assumption 3.6 on every directional neighborhood. Replacing $\beta^\star_{\mathrm{quad}}$ by the larger $\overline{\beta}_{\mathrm{quad}}$ preserves the inequality. □

For symmetric Q, Eq. 44 justifies the experimental constant $\beta = \|b\|_2\|Q\|_{\mathrm{op}}$. When $\|b\|_2 = 1$, this reduces to $\beta = \|Q\|_{\mathrm{op}}$. Without symmetry, the same experimental expression is still a valid upper bound but need not be the smallest one. Even an exact global constant can be conservative for a restricted perturbation subspace. The affine term changes the anchor Jacobian but not its variation across inputs. The next lemma supplies the readout amplification and the mixed matrix norm appearing in the Softplus bound. For a matrix $M \in \mathbb{R}^{k\times n}$ with positive dimensions, write $\|M\|_{2\to\infty} := \sup_{\|u\|_2=1}\|Mu\|_\infty$.

Lemma D.5 (Exact Induced Norm of an Affine Readout)**.** *Let $g : \mathbb{R}^p \to \mathbb{R}^C$ be given by $g(z) := Wz + c$, with $W \in \mathbb{R}^{C\times p}$ and $c \in \mathbb{R}^C$, where $p, C \ge 1$. Its smallest global Lipschitz constant from the Euclidean norm to the infinity norm is*

$$\begin{aligned} L^\star_g &:= \sup_{z\neq z'} \frac{\|g(z) - g(z')\|_\infty}{\|z - z'\|_2} \\ &= \|W\|_{2\to\infty} = \max_{1\le k\le C}\|W_{k,:}\|_2. \end{aligned} \tag{46}$$

Consequently, this constant is valid on any attained representation set $\mathcal{Z}_{x,E,\rho}$ in Corollary 4.5. For the identity readout on $\mathbb{R}^C$, it gives $L_g = 1$.

Proof. The bias cancels, giving $g(z) - g(z') = W(z - z')$. Every nonzero vector can occur as $z - z'$, so homogeneity implies $L^\star_g = \|W\|_{2\to\infty}$. Let $L := \max_k \|W_{k,:}\|_2$. For any $u \in \mathbb{R}^p$, Cauchy–Schwarz gives

$$\|Wu\|_\infty = \max_{1\le k\le C}|W_{k,:}u| \le L\|u\|_2. \tag{47}$$

Eq. 47 proves the upper bound $\|W\|_{2\to\infty} \le L$. If $L = 0$, every row of W is zero, so g is constant and the asserted equality holds. Otherwise, choose $k_\star$ with $\|W_{k_\star,:}\|_2 = L$ and set $u_\star := W^\top_{k_\star,:}/L$. Then $\|u_\star\|_2 = 1$ and $\|Wu_\star\|_\infty \ge |W_{k_\star,:}u_\star| = L$. This proves Eq. 46. A bound valid for every pair in $\mathbb{R}^p$ remains valid when both points are restricted to any subset, including $\mathcal{Z}_{x,E,\rho}$. For $W = I_C$, every row has Euclidean norm one, so Eq. 46 yields $L_g = 1$. □

The maximum row norm is the exact global readout constant for these input and output norms. It need not be the smallest constant on the attained representation set. Applying Corollary 4.5 with the complete logit map f in place of the intermediate representation F and an identity readout allows the valid choice $L_g = 1$, since $\|z - z'\|_\infty \le \|z - z'\|_2$. In this specialization, Assumption 3.6 required by Theorem 4.3 must hold for f itself on the claimed directional neighborhood. This avoids a separate readout-Lipschitz factor. It does not eliminate the readout's contribution to the sensitivity

of the complete predictor. We now derive the sufficient Jacobian-Lipschitz bound reported for the one-hidden-layer smooth classifier. The Softplus parameterization is explicit because it determines the factor multiplying the matrix norms.

Lemma D.6 (Global Curvature Bound for a Softplus Logit Map)**.** *Let $d_a, m, C \geq 1$, $\tau > 0$, $W_1 \in \mathbb{R}^{m\times d_a}$, $W_2 \in \mathbb{R}^{C\times m}$, $b_1 \in \mathbb{R}^m$, and $b_2 \in \mathbb{R}^C$. Define the unthresholded scalar activation $s_\tau(t) := \tau^{-1}\log(1+\exp(\tau t))$, applied componentwise to vectors, and set $f(a) := W_2 s_\tau(W_1 a + b_1) + b_2$ for $a \in \mathbb{R}^{d_a}$. Then $f \in C^\infty(\mathbb{R}^{d_a};\mathbb{R}^C)$ and*

$$\overline{\beta}_f := \frac{\tau}{4}\|W_2\|_{\mathrm{op}}\|W_1\|_{\mathrm{op}}\|W_1\|_{2\to\infty}, \tag{48}$$

$$\|J_f(a) - J_f(a')\|_{\mathrm{op}} \leq \overline{\beta}_f\|a - a'\|_2, \qquad a, a' \in \mathbb{R}^{d_a}. \tag{49}$$

Here, $\|W_1\|_{2\to\infty} = \max_{1\leq j\leq m}\|(W_1)_{j,:}\|_2$. In particular, $\overline{\beta}_f$ is a valid, not necessarily minimal, constant in Assumption 3.6 for $F = f$ on every directional neighborhood in $\mathbb{R}^{d_a}$.

Proof. Differentiation gives $s'_\tau(t) = (1+\exp(-\tau t))^{-1}$ and $s''_\tau(t) = \tau s'_\tau(t)(1 - s'_\tau(t))$. Since $0 < s'_\tau(t) < 1$ and $u(1-u) \leq 1/4$ for $u \in [0,1]$, we have $0 \leq s''_\tau(t) \leq \tau/4$ for every $t \in \mathbb{R}$. The mean value theorem therefore gives

$$|s'_\tau(t) - s'_\tau(t')| \leq \frac{\tau}{4}|t - t'|, \qquad t, t' \in \mathbb{R}. \tag{50}$$

The activation is smooth on all of $\mathbb{R}$, so its composition with the affine layers is smooth on $\mathbb{R}^{d_a}$. Define the diagonal matrix $D(a) := \mathrm{diag}\big(s'_\tau((W_1a+b_1)_1), \ldots, s'_\tau((W_1a+b_1)_m)\big)$. The chain rule yields

$$J_f(a) = W_2 D(a) W_1, \qquad a \in \mathbb{R}^{d_a}. \tag{51}$$

For any $a, a' \in \mathbb{R}^{d_a}$, the Euclidean operator norm of the diagonal matrix $D(a) - D(a')$ is the largest absolute value of its diagonal entries. Applying Eq. 50 entrywise gives

$$\begin{aligned}\|D(a) - D(a')\|_{\mathrm{op}} &= \max_{1\leq j\leq m}|s'_\tau((W_1a+b_1)_j) - s'_\tau((W_1a'+b_1)_j)| \\ &\leq \frac{\tau}{4}\|W_1(a-a')\|_\infty \\ &\leq \frac{\tau}{4}\|W_1\|_{2\to\infty}\|a-a'\|_2.\end{aligned} \tag{52}$$

The last inequality is the definition of the induced norm; its maximum-row-norm expression follows from Lemma D.5 applied to the linear map with matrix W_1. Subtracting Eq. 51 at a and a', and using Eq. 52, yields

$$\begin{aligned}\|J_f(a) - J_f(a')\|_{\mathrm{op}} &= \|W_2(D(a) - D(a'))W_1\|_{\mathrm{op}} \\ &\leq \|W_2\|_{\mathrm{op}}\|D(a) - D(a')\|_{\mathrm{op}}\|W_1\|_{\mathrm{op}} \\ &\leq \frac{\tau}{4}\|W_2\|_{\mathrm{op}}\|W_1\|_{\mathrm{op}}\|W_1\|_{2\to\infty}\|a-a'\|_2 \\ &= \overline{\beta}_f\|a-a'\|_2.\end{aligned} \tag{53}$$

By Eq. 48, Eq. 53 proves Eq. 49. For a nonzero subspace $E \subseteq \mathbb{R}^{d_a}$, multiplication by P_E cannot increase the operator norm. Consequently, the same bound controls $\|(J_f(a) - J_f(a'))P_E\|_{\mathrm{op}}$, in particular for all pairs in any specified directional neighborhood. This verifies Assumption 3.6 with $F = f$ and $\beta = \overline{\beta}_f$. No step divides by a weight norm, so the argument includes zero weight matrices and a zero value of $\overline{\beta}_f$. □

Eq. 48 provides a global upper bound from the fixed network parameters, not an estimate from sampled input pairs. The factor $\tau/4$ bounds scalar activation curvature. $\|W_1\|_{2\to\infty}$ controls the largest change of an individual preactivation, and the remaining operator norms arise from the Jacobian's linear factors. The resulting product need not equal the true Jacobian-Lipschitz constant and can be conservative. Sampled checks of Jacobian variation are implementation diagnostics. Eq. 48 applies to the stated unthresholded parameterization, not automatically to a differently scaled or piecewise-modified activation.

E Additional Experimental Details

E.1 Setup and Metrics

The exact analytic experiments (Experiments 1-6) are deterministic audits. Experiment 7 is a neural certificate bridge on prepared MNIST PCA artifacts. Experiments 8-9 are CIFAR-10 GPU experiments run with frozen seed-specific prepared data.

Common metric definitions. Activation-only statistics are functions of finitely observed representation values only. In the experiments, these include finite-sample distances, covariance spectra, effective rank, participation ratio, linear CKA, alignment, and uniformity. A local secant frame is a finite set of perturbation vectors $V = [v_1, \ldots, v_q]$ with covered span $E_V = \mathrm{range}(V)$. The empirical secant operator is $\widehat{J}_F(x; V) = Y_F(x; V)V^\dagger$, where Y_F contains endpoint representation differences. The recovery uncertainty $\eta_\beta(V)$ is the curvature-scale-conditioning radius from Theorem 4.3. Unless stated otherwise, radii and $\eta_\beta(V)$ are reported in the native coordinate units of the perturbation model being audited. In Experiment 7, these units are standardized train-fitted PCA coefficient units: for example, radius 0.05 means Euclidean radius 0.05 in that coefficient space, and median certified radius 0.336 is a raw coefficient-space radius. The pairwise score is the average squared endpoint response over the queried directions, whereas the secant operator score is $\|\widehat{J}_F(x; V)\|_{\mathrm{op}}$ and measures the largest reconstructed gain over unit directions in the covered span.

Neural and CIFAR metrics. Clean accuracy is standard test accuracy. Prediction invariance measures the fraction of anchors whose predicted label remains unchanged under the covered perturbation evaluation. Joint robust correctness requires both clean correctness and prediction invariance. The secant operator norm is the finite query estimate of worst covered response. The autodiff local-Jacobian norm is the operator norm of the autodiff-selected local derivative of the logits with respect to the chart coefficients on the sampled audit points. For nonsmooth networks and interpolation charts, it is a sampled local sensitivity diagnostic. Finite-radius worst gain is an empirical maximum response over the covered finite-radius perturbation set. The exhaustive 2D-grid diagnostic enumerates a 41-by-41 grid restricted to the radius-1 ball in two transformation coordinates and reports robust correctness on a fixed 32-example diagnostic subset. CIFAR-C mean accuracy is a secondary corruption-transfer diagnostic. Full-pixel PGD and AutoAttack (Croce & Hein, 2020) endpoints are recorded as stress tests.

Statistical reporting. For exact audits in Experiments 1-6, we report absolute residuals and violations under the configured tolerances. For Experiment 8 and Experiment 9 factorial, the independent unit is the seed $\{42, 128, 999, 1022, 3407\}$. We report percentile confidence intervals over seeds and paired seed effects where applicable.

E.2 Experiment 1: Activation-Only Non-identifiability

Implementation. Experiment 1 audits Theorem 4.1. For each geometry the base representation is a smooth quadratic map with zero anchor Jacobian in the tested direction. The perturbed family is $F_\alpha(z) = F_0(z) + \alpha\langle u, z - x_i\rangle\psi(z)a$, where ψ is one near the anchor and zero on every protected finite observation. The protected set includes the original finite data and the paired-view inputs used by alignment and uniformity diagnostics. The grid uses dimensions $d \in \{2, 8, 32\}$, output dimensions $p \in \{1, 8, 64\}$, sample sizes $N \in \{32, 256, 2048\}$, five geometric configurations, and $\alpha \in \{0, 0.1, 1, 10, 100, 1000\}$. The run used seed 42 and tolerance 10^{-10}.

Results. Table 4 summarizes the results. All activation-only diagnostics remain fixed while the anchor directional derivative grows with α. Experiment 1 supports the activation-only non-identifiability claim by constructing pairs of representations that are indistinguishable from any fixed finite activation record but have arbitrarily different local directional sensitivities at an anchor. Across 135 geometries and 810 measurements, the maximum invariance or analytic error is 4.547×10^{-13}, and the recorded finite-observation mismatch is exactly zero at the largest perturbation level. Nevertheless, when $\alpha = 1000$, the measured local directional sensitivity is 1000. Thus, the same finite activation observations can correspond to representations with very different local derivatives. This directly instantiates the theorem's information-level statement: activation-only statistics may describe the

observed embedding geometry, but they cannot, by themselves, certify a deterministic upper bound on local sensitivity over the stated smooth function class.

Table 4: Experiment 1 results.

Quantity	Value
Geometric instances	135
Alpha-level measurements	810
Maximum static finite-observation mismatch	0
Maximum invariance/analytic error	4.547×10^{-13}
Maximum sensitivity identity error	4.547×10^{-13}
Maximum Jacobian-difference change	1.128×10^{-13}
Largest α tested	1000
Sensitivity at largest α	1000
Status	Pass

E.3 Experiment 2: Uncovered-Direction Non-identifiability

Implementation. Experiment 2 tests Proposition 4.2. The target space is $E = \mathbb{R}^d$, but the query frame spans only the first $r < d$ coordinate directions. The omitted direction is $u = e_{r+1}$. The affine family is $F_\alpha(z) = Az + b + \alpha\langle u, z - x\rangle a$, with $Au = 0$ and $\|a\|_2 = 1$. Static observations and query endpoints lie in the covered affine plane, so the finite record is independent of α. The grid uses $d \in \{8, 32, 128\}$, ranks $r \in \{2, 4, 8, 16\}$, frame types {`orthonormal`, `redundant_tight`, `duplicated_imbalanced`}, redundancies $q/r \in \{1, 2, 4\}$, 64 static observations, step size $h = 0.2$, and $\alpha \in \{0, 1, 10, 100\}$. The run used seed 42 and tolerance 10^{-8}.

Results. Table 5 reports the negative-control and positive-control checks. The original under-covered query record remains unchanged even while omitted sensitivity reaches 100. Appending the missing direction reveals that response.

Table 5: Experiment 2 results.

Quantity	Value
Maximum exact error	2.205×10^{-13}
Maximum finite-activation mismatch at $\alpha = 100$	0
Maximum secant-matrix mismatch at $\alpha = 100$	0
Maximum original empirical-operator mismatch at $\alpha = 100$	0
Mean omitted-direction sensitivity at $\alpha = 100$	100
Mean recovered response after appending missing direction	100
Most demanding frame group	duplicated_imbalanced, $q/r = 1$
Status	Pass

This experiment supports the coverage-necessity proposition. Exact endpoint secants are informative, but only over E_V. They cannot rule out arbitrary sensitivity in $E \cap E_V^\perp$ without additional assumptions. The positive control shows that the failure is not the secant algebra itself. The missing direction becomes visible once it is queried.

E.4 Experiment 3: Coverage-Aware Secant Recovery

Implementation. Experiment 3 examines Theorem 4.3 and its two-sided interval. The analytic map is $F(z) = Az + \frac{1}{2}b(z^\top Qz)$ with $\|b\|_2 = 1$, symmetric Q, and $\beta = \|Q\|_{\mathrm{op}}$. The anchor Jacobian is $J_F(x) = A + b(Qx)^\top$. Frames are generated as $V = Q_E \,\mathrm{diag}(s) R^\top$ with controlled probe scale, rank, redundancy, and condition number. The grid uses $d \in \{8, 32, 128\}$, output dimensions $\{4, 16, 64\}$, ranks $\{1, 2, 4, 8, 16\}$, redundancies $q/r \in \{1, 2, 4\}$, condition numbers $\{1, 3, 10, 30, 100, 300\}$, probe scales $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 0.1, 0.3\}$, curvatures $\{0, 0.1, 1, 10\}$, and $\rho'/\rho \in \{0.25, 0.5, 1\}$.

Results. Tables 6 and 7 report the recovery audit and the scaling checks. The run evaluates 24,192 valid frame settings with no recovery violations and no interval misses. The zero-violation recovery audit supports the theorem's covered-subspace claim: the finite endpoint operator recovers $J_F(x)P_{E_V}$ only up to the explicit uncertainty $\eta_\beta(V)$. The scaling audit supports the form of that uncertainty,

especially the linear dependence on probe scale and curvature. The experiment does not claim recovery of the full Jacobian unless E_V covers the claimed space.

Table 6: Experiment 3 core results.

Check	Expected	Observed	Maximum Margin
Recovery bound	0 violations	0	2.575×10^{-13}
Two-sided interval	0 misses	0	1.075×10^{-13}
Finite-scale consequence	0 certified violations	0	1.075×10^{-13}
Frame radius	max column norm $\leq \rho$	0	2.220×10^{-16}

Table 7: Scaling audit for $\eta_\beta(V)$.

Scaling	Target	Median Slope	Mean Slope	Groups
η vs. probe scale h	1.000	1.000	1.000	2268
η vs. curvature β	1.000	1.000	1.000	6048
η vs. $1/\sigma^+_{\min}(V)$	positive, near 1	0.824	0.814	2376
η formula identity	exact	n/a	n/a	18144

E.5 Experiment 4: Sharpness of the Curvature-Scale Term

Implementation. Experiment 4 audits the one-dimensional sharpness construction for the $\beta h/2$ term. The maps are $F_+(t) = +(\beta/2)t(h-t)$ and $F_-(t) = -(\beta/2)t(h-t)$. Both satisfy $F(0) = F(h) = 0$, so the endpoint observation and one-sided secant are identical, but their anchor derivatives are $\pm\beta h/2$. The grid has five curvature values $\{0.01, 0.1, 1, 10, 100\}$ and 25 logarithmic probe scales from 10^{-6} to 1, for 125 exact cases.

Results. Tables 8 and 9 summarize the results. The maximum exact error is 2.220×10^{-16}, below tolerance 10^{-10}. The result supports Proposition 4.4: one-sided endpoint agreement does not determine the anchor derivative more accurately than $\beta h/2$ over the stated function class. The symmetric-probe comparison clarifies the boundary of the claim: additional query structure can improve this particular quadratic case, but that is a stronger observation model.

Table 8: Experiment 4 core results.

Check	Audit Statistic	Value
Endpoint observations agree	max endpoint mismatch	0
One-sided secant estimate is zero	max $\|\mathrm{secant}\|$	0
Anchor derivative magnitude	max formula residual	0
Recovery error equals $\eta_\beta([h])$	max $\|\eta - \mathrm{error}\|$	0

Table 9: Scaling and probe-design comparison.

Audit	Target	Observed	Comment
η vs h	slope 1	median 1.000, mean 1.000	exact scaling
η vs β	slope 1	median 1.000, mean 1.000	exact scaling
One-sided endpoint $[0, h]$	error $\beta h/2$	max error 50.0	attains bound
Symmetric endpoints $[-h, h]$	exact for this quadratic	max error 2.220×10^{-16}	stronger query design

E.6 Experiment 5: Pairwise Consistency vs. Worst-Direction Operator Control

Implementation. Experiment 5 has an exact affine audit and a secondary learned stress test. The exact audit constructs two affine maps with identical pairwise score γ^2: an isotropic construction with r equal singular values γ, and a rank-one construction with one singular value $\sqrt{r}\,\gamma$, where $\gamma > 0$. It uses ranks $\{2, 4, 8, 16, 32\}$, output dimensions $p \in \{r, 2r\}$, amplitudes $\{0.1, 1, 10\}$, eight random orthogonal draws, and $h = 0.2$, for 480 exact cases. The learned stress test uses three seeds $\{101, 202, 303\}$, 120 epochs, batch size 128, nuisance rank 4, and methods $\{$`pairwise`, `sfr`, `exact_jacobian`$\}$.

Results. Tables 10, 11, and 12 report the exact identity audit, the rank-dependent gap, and the learned stress test. In Table 10, the superscripts iso and $\mathrm{rank1}$ identify the pairwise scores of the corresponding affine constructions, and $\widehat{J}$ abbreviates $\widehat{J}_F(x;V)$. The exact audit passes with maximum exact error 1.137×10^{-13}. The results of the exact identity audit support Proposition 4.6: a pairwise average is a normalized Frobenius quantity, while the operator norm controls the largest covered direction. The learned stress test shows that finite operator control can change the learned local spectrum in the predicted direction.

Table 10: Exact operator-vs-pairwise.

Check	Target	Audit Residual
Pairwise/Frobenius identity	$\mathcal{R}_{\mathrm{pair}} = \|\widehat{J}\|_{\mathrm{F}}^2/r$	5.684×10^{-14}
Pairwise equality across constructions	$\mathcal{R}_{\mathrm{pair}}^{\mathrm{iso}} = \mathcal{R}_{\mathrm{pair}}^{\mathrm{rank1}}$	8.527×10^{-14}
Equality-case operator norms	γ and $\sqrt{r}\,\gamma$	3.553×10^{-14}
Inequality band	$\sqrt{\mathcal{R}_{\mathrm{pair}}} \le \|\widehat{J}\|_{\mathrm{op}} \le \sqrt{r\,\mathcal{R}_{\mathrm{pair}}}$	2.132×10^{-14}
Rank of equality cases	rank r vs rank one	0

Table 11: Dimension gap for the rank-one construction.

Rank r	Observed Gap	Expected Gap	Max Residual
2	1.414	1.414	6.661×10^{-16}
4	2.000	2.000	8.882×10^{-16}
8	2.828	2.828	1.776×10^{-15}
16	4.000	4.000	2.220×10^{-15}
32	5.657	5.657	2.665×10^{-15}

Table 12: Learned stress-test summary over three seeds.

Method	Test Acc.	Finite $\mathcal{R}_{\mathrm{pair}}$	Finite op.	Finite Gap	Lead Conc.	Jac. op.
Pairwise finite endpoints	0.906 ± 0.008	5.596	1.725	1.963	0.964	1.717
SFR finite-operator	0.900 ± 0.008	4.809	1.478	1.936	0.940	1.455
Exact restricted Jacobian	0.898 ± 0.007	5.090	1.528	1.940	0.943	1.502

E.7 Experiment 6: Analytic Prediction-Certificate Audit

Implementation. Experiment 6 tests Corollary 4.5. The representation is $F(x) = Ax + \mathrm{bias} + \frac{1}{2}(x^\top Qx)b$, where Q is symmetric, so the exact Jacobian Lipschitz constant is $\beta = \|b\|_2\|Q\|_{\mathrm{op}}$. The readout is linear, $g(z) = Wz + c$, with exact ℓ_2-to-ℓ_∞ Lipschitz constant $L_g = \max_c \|W_{c,:}\|_2$. Frames are fully covered and orthonormal in the claimed subspace. The audit uses ranks $\{1,2,3,4\}$, class counts {2,5}, $\beta \in \{0, 0.1, 0.5, 1\}$, 30 trials per setting, radius cap $\rho = 0.5$, and 960 anchors. Certified roots are shrunk with numerical safety factors and rechecked with strict inequalities in float64.

Results. Tables 13, 1, and 14 report the certificate audit, target-radius coverage, and rank/class diagnostics. The mean certified radius across all anchors is 0.261, and the median is 0.237. This experiment supports the intended use of Corollary 4.5: the coverage-aware secant bound supplies a local covered-subspace movement estimate, and a prediction certificate is issued only after accounting for margin and readout amplification.

Table 13: Certificate validity results.

Check	Target	Value
Strict certificate inequality	no failures	0
Observed attack violations	zero found attacks	0
Secant recovery bound	$\|\widehat{J}_F(x;V) - J_F(x)P_{E_V}\|_{\mathrm{op}} \le \eta_\beta(V)$	4.392×10^{-14}
Nonvacuous certificates	radius above reporting floor	1.000
Correct-anchor nonvacuous coverage	correct unique-margin anchors	1.000

Table 14: Radius diagnostics by rank and class count.

Rank	Classes	Anchors	Mean Radius	Median Radius	Boundary Status
1	2	120	0.354	0.408	exact 1D
1	5	120	0.238	0.215	exact 1D
2	2	120	0.303	0.313	attack upper bound
2	5	120	0.220	0.185	attack upper bound
3	2	120	0.304	0.301	attack upper bound
3	5	120	0.185	0.145	attack upper bound
4	2	120	0.295	0.292	attack upper bound
4	5	120	0.185	0.148	attack upper bound

E.8 Experiment 7: Certified Neural Bridge on Real Data

Implementation. Experiment 7 tests whether the certificate pipeline can be instantiated on trained neural classifiers. The data is MNIST. It stores a 55,000-example optimization-training split, 5,000-example validation split, and untouched 10,000-example official test split. PCA and coefficient standardization are fit only on the optimization-training split. All models are one-hidden-layer unthresholded Softplus classifiers. The certificate uses the logit map directly and computes the analytic curvature upper bound $\beta = (\tau/4)\|W_2\|_{\mathrm{op}}\|W_1\|_{\mathrm{op}}\|W_1\|_{2\to\infty}$ in float64 with an upward safety factor. The finite-query scale h is selected on validation data only. The run uses seed 424242 and completes 11 training runs. The formal certificate scope is standardized train-fitted PCA coefficient space, not pixel space. For the main MNIST PCA-16 comparison, validation selects $h = 0.025$, the full-test certificate table uses all 10,000 official test examples, and the attack audit uses a fixed 256-example attack subset.

Results. Tables 15, 16, and 17 report the main PCA-16 comparison, rank sweep, and attack audit. The results show that the maximum observed sampled Jacobian-difference quotient divided by the analytic Softplus bound is 0.3240, the maximum sampled curvature-bound violation is -9.884×10^{-2}, and the strict-radius float64 recheck passes. The apparent difference between the MNIST PCA-16 $r = 0.05$ certificate rates in Table 16 and Table 17 is a denominator difference, not a conflict. Table 16 reports full-test certificate correctness on all 10,000 official test examples. For ERM/SFR, the rates are $84.0\%/86.0\%$. Table 17 reports the same selected PCA-16 checkpoints and the same raw radius on the fixed 256-example attack subset. For ERM/SFR, the subset rates are $84.4\%/85.5\%$. Experiment 7 supports the Corollary 4.5 in a trained neural setting whose perturbation space is fully covered and whose curvature bound is analytic. The result also shows the practical limitation: global curvature bounds can be conservative, and the certificate is for PCA coefficient space, not pixel space.

Table 15: MNIST PCA-16 method comparison. Certified radii are raw standardized PCA coefficient-space quantities.

Method	Clean	β	Secant op.	Exact J op.	Median Cert. r	False Certs
ERM	89.4%	5.770	6.690	6.692	0.231	0
Pairwise	89.5%	5.158	5.701	5.701	0.253	0
Exact-Jacobian	89.2%	4.115	3.699	3.704	0.336	0
SFR	89.2%	4.112	3.698	3.702	0.336	0

Table 16: MNIST rank sweep. η_β and certified radii are raw standardized PCA coefficient-space quantities.

Rank	Method	Clean	η_β	Median Cert. r	Cert Correct @ $r = 0.05$
8	ERM	83.3%	0.220	0.175	73.6%
8	SFR	82.8%	0.153	0.255	76.6%
16	ERM	89.4%	0.289	0.231	84.0%
16	SFR	89.2%	0.206	0.336	86.0%
32	ERM	93.5%	0.414	0.281	90.2%
32	SFR	93.5%	0.262	0.422	91.4%

Table 17: Attack audit on MNIST PCA-16. Radii are raw standardized PCA coefficient-space radii, and rows are computed on the fixed 256-example attack subset.

Radius r	Method	Cert Correct	Attack Robust	Attack Success Clean	False Certs
0.05	ERM	84.4%	88.3%	1.7%	0
0.05	Pairwise	84.8%	88.7%	1.3%	0
0.05	Exact-Jacobian	85.5%	87.9%	1.7%	0
0.05	SFR	85.5%	87.9%	1.7%	0
0.10	ERM	76.6%	85.9%	4.3%	0
0.10	Pairwise	77.3%	85.9%	4.3%	0
0.10	Exact-Jacobian	80.5%	85.9%	3.9%	0
0.10	SFR	80.5%	85.9%	3.9%	0
0.25	ERM	41.8%	81.6%	9.1%	0
0.25	Pairwise	49.2%	82.4%	8.3%	0
0.25	Exact-Jacobian	63.3%	81.2%	9.2%	0
0.25	SFR	63.3%	81.2%	9.2%	0
0.50	ERM	2.3%	68.4%	23.9%	0
0.50	Pairwise	3.5%	67.2%	25.2%	0
0.50	Exact-Jacobian	15.2%	71.1%	20.5%	0
0.50	SFR	15.6%	71.1%	20.5%	0

E.9 Experiment 8: CIFAR-10 Covered Transformation Training

Implementation. Experiment 8 is the primary CIFAR-10 training experiment. It uses frozen seeds $\{42, 128, 999, 1022, 3407\}$, seed-specific prepared CIFAR-10 splits, a seed-specific frozen protocol, and four methods: ERM, Pairwise, Subspace TRADES, and SFR. Training and evaluation were run on a supercomputer, using 8 CPU cores, 64GB memory, and one GPU. CIFAR-C was prepared as a secondary corruption-transfer endpoint. Confidence intervals are percentile intervals over the five seeds, and paired effects are seed-paired. Table 18 reports the frozen protocol parameters used in this experiment.

Table 18: Experiment 8 frozen protocol parameters used.

Component	Frozen setting
Data and split	CIFAR-10 from prepared seed-specific splits; augmentation enabled; batch size 128; evaluation batch size 128; metric subset size 256.
Model and precision	ResNet-18, base width 64, ReLU activation, automatic mixed precision.
Covered transformation chart	Five-dimensional `transform5` chart: translation radius 2 pixels, rotation radius 5°, brightness delta 0.1, contrast delta 0.2, reflection padding, `align_corners=false`, and clamping enabled.
Frame and radius	Rank-5 frame with redundancy 1 and scale 0.25; frozen coefficient-space attack/evaluation radius $r = 1.0$.
Training	SGD for 200 epochs with learning rate 0.1, minimum learning rate 0, momentum 0.9, and weight decay 5×10^{-4}; archived epochs {1,50,100,150,200}.
Method parameters	ERM, Pairwise, Subspace TRADES, and SFR are run for every seed. Regularizer parameters are $\lambda = 0.01$, $\zeta = 0$, with $\lambda_{\mathrm{SF}} = \lambda$ for SFR. The TRADES loss hyperparameter is $\beta_{\mathrm{TRADES}} = 6.0$. Training uses 5 adversarial steps where applicable. The matched forward-query target is 12 effective conceptual forwards per example: ERM uses 1 conceptual query plus padding, Pairwise and SFR use 7, and Subspace TRADES uses 8.
Checkpoint selection	Validation only. ERM is selected by validation clean accuracy; non-ERM methods are selected by validation primary constrained robustness subject to the seed-matched ERM clean-accuracy floor. Final test metrics are opened only after checkpoint selection.
Primary evaluation	Covered-coordinate joint robustness is evaluated at radius $r = 1.0$ with 50 attack steps, 10 restarts, losses {cross-entropy, margin, DLR}, and step-size multipliers {0.5,1.0,2.0}.
Diagnostics and transfer endpoints	2D-grid diagnostic: coefficient indices {0,1}, 41 points per axis, fixed 32-example test diagnostic subset, radius 1.0. Out-of-frame DCT diagnostic: rank 8, coefficient scale 0.05, radius 1.0, 50 steps, 10 restarts, 128-example subset. Full-pixel PGD and AutoAttack use ℓ_∞ radius $8/255$, 50 PGD steps and 10 restarts where applicable; CIFAR-C uses all 15 standard corruptions and severities 1-5.

Results. Tables 19, 20, 21, and 22 report the main results, diagnostic summaries, paired effects, and SFR-vs-Pairwise per-seed effects. AutoAttack is omitted from the summary tables because all four methods obtain 0 under that full-pixel stress endpoint. The 2D-grid column in Table 20 is the 32-example diagnostic subset statistic defined above. Experiment 8 provides empirical evidence that directly controlling the covered secant operator is useful in the covered transformation-coordinate threat model. SFR improves the matched-query pairwise baseline on joint robust correctness and prediction invariance with a small clean-accuracy cost, and it reduces the local sensitivity diagnostics targeted by the objective.

Table 19: Primary accuracy and robustness metrics. The results are means [95% confidence interval] over the five frozen seeds; all values are percentages.

Method	Clean ↑	Joint Robust ↑	Pred. Inv. ↑	CIFAR-C ↑
ERM	95.1 [95.0, 95.2]	48.1 [46.5, 50.2]	50.8 [49.1, 52.9]	74.5 [73.9, 75.0]
Pairwise	95.5 [95.4, 95.6]	82.6 [82.3, 82.9]	86.3 [85.9, 86.6]	83.2 [83.1, 83.4]
Subspace TRADES	95.3 [95.1, 95.4]	85.8 [85.5, 86.1]	89.7 [89.4, 90.0]	79.1 [78.2, 80.0]
SFR	95.2 [95.1, 95.3]	85.1 [84.9, 85.3]	89.1 [88.9, 89.3]	85.1 [84.9, 85.3]

Table 20: Covered-sensitivity diagnostics. Lower is better for the four raw sensitivity diagnostics; the 2D-grid column is a percentage on the fixed 32-example diagnostic subset.

Method	Secant op. ↓	Pairwise ↓	Autodiff Local J ↓	Worst Gain ↓	2D Grid ↑
ERM	8.958 [8.630, 9.195]	32.848 [30.078, 35.331]	7.553 [7.239, 7.867]	5.685 [5.495, 5.886]	80.0 [76.2, 83.8]
Pairwise	1.360 [1.281, 1.438]	1.859 [1.637, 2.056]	1.681 [1.539, 1.828]	1.293 [1.213, 1.398]	98.8 [97.5, 100.0]
Subspace TRADES	4.097 [4.023, 4.186]	5.557 [5.391, 5.766]	5.233 [5.012, 5.409]	2.045 [2.006, 2.087]	95.0 [93.8, 97.5]
SFR	1.126 [0.984, 1.276]	1.766 [1.316, 2.319]	1.394 [1.210, 1.553]	1.045 [0.979, 1.148]	98.8 [97.5, 100.0]

Table 21: Seed-paired effects favoring SFR. Sign-flip values are descriptive.

Comparison	Metric	Effect	Unit	Seeds Favoring SFR	Sign-Flip
SFR vs Pairwise	Clean acc.	$-0.304[-0.498, -0.128]$	percentage points	0/5	0.0625
SFR vs Pairwise	Joint robust	$2.46[2.26, 2.64]$	percentage points	5/5	0.0625
SFR vs Pairwise	Prediction inv.	$2.85[2.63, 3.07]$	percentage points	5/5	0.0625
SFR vs Pairwise	Secant op.	$0.234[0.105, 0.363]$	raw units	5/5	0.0625
SFR vs Pairwise	Autodiff local J	$0.287[0.178, 0.401]$	raw units	5/5	0.0625
SFR vs Pairwise	Worst gain	$0.247[0.205, 0.282]$	raw units	5/5	0.0625
SFR vs Subspace TRADES	Joint robust	$-0.734[-1.22, -0.296]$	percentage points	1/5	0.1250
SFR vs Subspace TRADES	Secant op.	$2.97[2.86, 3.08]$	raw units	5/5	0.0625
SFR vs Subspace TRADES	Worst gain	$1.00[0.926, 1.06]$	raw units	5/5	0.0625

Table 22: SFR-vs-Pairwise per-seed effects.

Seed	Clean pp	Joint Robust pp	Pred. Inv. pp	Worst Gain Reduction
42	-0.340	2.060	2.530	0.301
128	-0.270	2.700	3.010	0.271
999	-0.650	2.440	3.230	0.171
1022	-0.010	2.460	2.610	0.257
3407	-0.250	2.660	2.890	0.238

E.10 Experiment 9: Checkpoint-Zoo Diagnostic Evidence

Implementation. Experiment 9 asks whether perturbation-aware diagnostics align with covered robustness across many trained checkpoints. The full-zoo study uses CIFAR-10, two architectures (ResNet-18 and CompactViT), four methods (ERM, Pairwise, SFR, Subspace TRADES), five seeds, final-training runs, archived early/late checkpoints, and SFR strength sweeps. The full-zoo output has 266 eligible checkpoints, 50 unique training runs, 40 primary final-training runs, and matched forward queries. All required artifact-integrity checks pass, including complete architecture-method-seed factorial, frozen view pairs, fixed CKA references, required datasets/methods/seeds, and zero metric failures. The run seed is 2027, and the recorded backward-evaluation count is 77,456,128.

Results. Table 23 reports the representative diagnostic correlations. Experiment 9 supports the information hierarchy at the diagnostic level. Perturbation-aware diagnostics, especially finite-radius worst gain and secant-related scores, align more strongly with covered robust accuracy than static activation-only statistics in this checkpoint zoo. This is observational evidence, not a causal claim.

E.11 Summary of Experimental Evidence

The experiments provide a staged validation of the paper's central message: finite activation-only summaries are insufficient for certifying local sensitivity, while covered perturbation queries expose the directional information needed for recovery and prediction guarantees (Table 24). Experiments

Table 23: Checkpoint-zoo correlations with covered joint robust accuracy.

Metric	Type	Spearman ρ_S	95% CI	Partial ρ_S	All-checkpoint ρ_S
Finite-radius worst gain	perturbation-aware	-0.893	[-0.932, -0.844]	-0.776	-0.584
Autodiff local-J norm	perturbation-aware	-0.832	[-0.878, -0.788]	-0.629	-0.511
Secant score	perturbation-aware	-0.815	[-0.836, -0.779]	-0.535	-0.516
Pairwise score	perturbation-aware	-0.803	[-0.830, -0.767]	-0.445	-0.486
Curvature proxy	perturbation-aware	-0.799	[-0.846, -0.770]	-0.427	-0.357
Cov. condition #	static	0.671	[0.582, 0.738]	0.282	-0.215
Between/within scatter	static	0.637	[0.593, 0.711]	0.112	0.791
Uniformity	static	0.603	[0.505, 0.681]	-0.164	0.010
Effective rank	static	-0.593	[-0.655, -0.471]	0.032	0.017
Participation ratio	static	-0.541	[-0.621, -0.401]	0.302	0.183

1-5 first show that activation-only statistics can remain unchanged even when local sensitivity changes, and that secant recovery succeeds only within the span actually covered by the queried directions. The certificate experiments then connect this recovery view to prediction stability. Finally, the CIFAR-10 experiments show that this theory leads to useful empirical diagnostics and training behavior: SFR improves over the matched-query pairwise baseline on covered robustness metrics while reducing covered sensitivity measures, and the checkpoint-zoo study shows that perturbation-aware diagnostics align more strongly with covered robust accuracy than static activation-only statistics.

Table 24: Summary of experimental evidence.

Question	Evidence	Main Experimental Takeaway	Scope
Can activation-only statistics certify local sensitivity?	E1	No. Identical finite activation observations can coexist with arbitrarily different local directional sensitivities, so static summaries alone cannot certify local behavior.	Worst-case finite-observation construction.
What does coverage add?	E2-E4	Covered secant queries recover the restricted local response with explicit uncertainty, while under-covered frames miss sensitivity in unqueried directions.	Guarantees apply only inside the covered span.
Why control an operator norm rather than only pairwise averages?	E5-E7	Pairwise averages can hide worst-case directions, whereas operator control matches the local quantity needed by the prediction certificate.	Local certificate; PCA-coordinate and covered-subspace scope.
Does SFR help in a practical training setting?	E8	On CIFAR-10 covered transformations, SFR improves over the matched-query pairwise baseline on covered joint robust correctness and prediction invariance, while lowering covered sensitivity diagnostics.	Empirical robustness evidence.
Do perturbation-aware diagnostics track covered robustness?	E9	Across the checkpoint zoo, perturbation-aware quantities such as finite-radius worst gain and secant-related scores align more strongly with covered robust accuracy than static activation-only statistics.	Observational diagnostic evidence; not causal identification.

E.12 Computing Hardware

Experiments 1 through 7 were conducted using only CPUs: the AMD EPYC 9655P and the AMD EPYC 75F3. Each of these experiments required 64 GB of memory. Experiments 8 and 9 were run on a supercomputer, each requesting 8 CPU cores, 64 GB of memory, and 1 NVIDIA L4 GPU.

F LIMITATION AND BROADER IMPACT

Limitation The framework differentiates between information derived from finite observations, guarantees of conditional stability, and an objective for empirical training. Theorem 4.3 bounds the Jacobian restricted to $E_V = \mathrm{range}(V)$ and anchored representation displacements within the stated radius. This anchored sensitivity is distinct from a two-point Lipschitz constant, and the stated result is not a global guarantee over arbitrary inputs. Corollary 4.5 applies to the prescribed subspace E only when $E_V = E$. A lower bound on $\sigma^{+}_{\min}(V)$ limits pseudoinverse amplification within the nonzero singular subspace but does not establish this coverage. In the absence of additional functional restrictions, covering E requires at least $\dim(E)$ linearly independent perturbations. Consequently, a small number of probes in a high-dimensional ambient space cannot support an unrestricted certificate through this framework. Additional structural assumptions or separate certification arguments are needed to extend the claim beyond the observed span.

Broader Impact This paper offers a framework for using finite representation evaluations to examine how a learned model reacts to specific input variations. Its potential impact lies in enhancing model development and assessment by linking representation analysis, perturbation-test design, and sensitivity-aware training. The distinction between descriptive geometry, measured responses, and conditional guarantees provides researchers and practitioners with a shared foundation for interpreting evidence of stability.

Secant frames complement static embedding statistics by measuring local responses to specified perturbations. Their use in training can improve evaluated robustness in selected settings, potentially at a cost in clean accuracy. Using a fixed perturbation model, evaluators can create a query frame to examine the largest gain from a fitted local operator, enabling a more comprehensive analysis than inspecting endpoint differences in isolation. This approach facilitates comparing checkpoints or training objectives using a common input metric, output metric, and query protocol. For instance, this formulation can accommodate a specified low-frequency subspace or the parameter coordinates of a differentiable transformation family. These analyses help focus testing on task-relevant variations and identify instances with significant observed responses for further investigation. The value of secant frames complements representation geometry by adding insights into local functional behavior, thereby enhancing the understanding of where observed samples are embedded.

The coverage and conditioning analysis offers clear guidance for developing evaluation protocols. A practitioner can define the target subspace, verify whether the query directions span it, and examine how their scales and conditioning affect reconstruction uncertainty. This approach distinguishes the number of evaluations from the directional information they provide. To support a stability claim about a specified subspace, additional queries can focus on missing directions rather than repeatedly testing directions already included. This framework enables more transparent comparisons between perturbation-testing strategies and clarifies the relationship between the tested directions and the claims being made.

SFR provides a training objective aimed at applications where specific input variations should lead to limited changes in predictions. Developers can encode these variations using a chosen perturbation frame and combine the empirical operator penalty with the task loss. The thresholds and regularization weights enable the examination of the desired level of response control alongside predictive performance. This approach utilizes finite forward evaluations and standard parameter backpropagation without explicitly forming the input Jacobian, making it a practical component for studying sensitivity control within existing gradient-based training pipelines. Proposition 4.6 offers a foundation for comparing average squared consistency with the control of the largest reconstructed response. These methods support a systematic investigation of which local invariances enhance a task, rather than viewing all input variation only as something to suppress.