WHY ROBUST CLASSIFIERS CAN GENERATE: HIDDEN DENOISING STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarially robust neural networks, while designed for classification, exhibit surprising generative capabilities when appropriately probed. We provide a theoretical framework explaining this phenomenon by connecting adversarial robustness to implicit denoising structure. Building on prior work that links Jacobian spectral structure to robustness, we show that the Gram operator $\mathbf{J}^\top\mathbf{J}$ functions as an implicit denoiser, selectively preserving dominant directions while suppressing noise in orthogonal directions. We propose a mechanistic explanation for emergent generative capabilities in robust classifiers, validated through a diagnostic probing procedure that makes the $\mathbf{J}^\top\mathbf{J}$ denoising operator empirically observable. Our results support a connection between discriminative robustness training and generative inference, showing that robust classifiers encode statistical priors that enable structured pattern generation without explicit generative objectives.

1 INTRODUCTION

Adversarially robust classifiers exhibit a striking property: when probed with gradient-based methods, they produce structured, semantically meaningful images rather than the high-frequency noise typical of standard networks (Santurkar et al., 2019; Engstrom et al., 2019). This emergent generative capability has proven remarkably useful beyond machine learning. In neuroscience, robust classifiers now serve as tools for synthesizing stimuli that precisely control neural activation in primate visual cortex (Gaziv et al., 2023; 2025) and generating maximally exciting images for individual neurons in mice and monkeys Willeke et al. (2025). The practical utility of these models for both interpretability in ML and experimental neuroscience makes understanding *why* they exhibit generative properties a pressing open question.

The prevailing explanation attributes this behavior to *perceptually-aligned gradients* (PAGs): robust classifiers have input gradients that align with human-interpretable features, so gradient ascent toward a class produces recognizable images (Tsipras et al., 2018; Kaur et al., 2019; Engstrom et al., 2019). Subsequent work formalized this via off-manifold robustness, proposing that gradients become perceptually aligned when constrained to lie near the data manifold (Srinivas et al., 2024). Moreover, this connection appears to be partly bidirectional: Ganz et al. (2023) showed that directly encouraging perceptually aligned gradients can itself induce adversarial robustness, using score-based targets to construct denoiser-like classifier gradients. While PAGs accurately *describe* the phenomenon, they do not *explain* it mechanistically. Why does adversarial training, an objective concerned only with classification robustness, induce gradients aligned with semantic structure? What mathematical property of robust networks enables generation?

Recent theoretical advances have begun to illuminate the mathematical structure underlying adversarial robustness. Studies have demonstrated that the spectral properties of neural networks, particularly input-output Jacobians, are directly linked to generalization and robustness (Oymak et al., 2019; Wu & Li, 2024). These spectral properties force networks to suppress sensitivity along most input directions while preserving discriminative power along a small subspace (Hoffman et al., 2019; Jakubovitz & Giryes, 2018). Jacobian regularization techniques have formalized this connection, showing that controlling gradient norms directly improves robustness by constraining the Jacobian spectrum (Ross & Doshi-Velez, 2017; Rodríguez-Muñoz et al., 2025). However, despite these insights into the *discriminative* implications of spectral structure, the potential *generative* consequences of low-rank Jacobians remain largely unexplored. This question is also related to recent efforts to unify

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

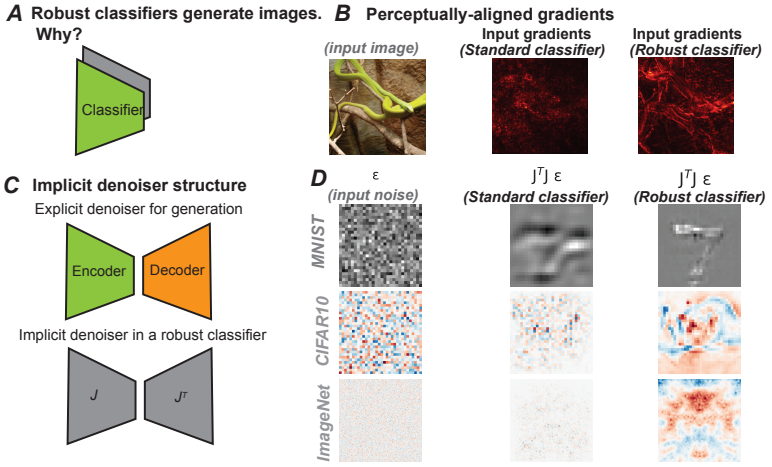

Figure 1: **From perceptually-aligned gradients to implicit denoiser structure.** **(A)** Adversarially robust classifiers exhibit unexpected (weak) generative capabilities when probed with gradient-based methods (Santurkar et al., 2019; Engstrom et al., 2019). Why does adversarial training induce this property? **(B)** The prevailing explanation: robust classifiers have *perceptually-aligned gradients* (PAGs), where input gradients align with human-interpretable features (Kaur et al., 2019). PAGs describe the phenomenon but do not explain its mechanistic origin. **(C)** This work: robust classifiers contain implicit denoiser structure. Explicit denoisers use encoder-decoder architectures for generation (Vincent et al., 2008; Vincent, 2011). We show that robust classifiers develop mathematically equivalent structure where the Jacobian $\mathbf{J}$ and its transpose $\mathbf{J}^\top$ form an implicit denoising operator $\mathbf{J}^\top\mathbf{J}$. Schematic analogy meant to highlight the structural correspondence between an encoder–decoder denoiser and the Jacobian Gram operator $\mathbf{J}^\top\mathbf{J}$. **(D)** Empirical validation across MNIST, CIFAR-10, and ImageNet: applying $\mathbf{J}^\top\mathbf{J}$, evaluated at a test image $\mathbf{x}$, to random noise $\boldsymbol{\epsilon}$, standard classifiers preserve random structure while robust classifiers extract class-consistent patterns, demonstrating the denoising capability predicted by Theorem 1. Images individually normalized to reveal structure (details in Appendix C.1).

classification and denoising more directly, including classification-denoising formulations derived from Tweedie-style identities Thiry & Guth (2024).

We address this gap by connecting spectral structure to implicit denoising. Our argument rests on a three-part chain: (i) adversarial training induces spectral concentration in the Jacobian, which we measure across datasets and architectures (Section 5); prior work connects related structure to explicit Jacobian regularization (Hoffman et al., 2019; Wu & Li, 2024; Oymak et al., 2019); (ii) under spectral concentration, $\mathbf{J}^\top\mathbf{J}$ acts as an *implicit denoiser*, shrinking components orthogonal to a low-dimensional discriminative subspace much more than components aligned with it (Theorem 1); (iii) this denoising structure enables generation from noise through our probing procedure. The resulting mechanism is mathematically analogous to the denoising operators in diffusion models (Ho et al., 2020; Song et al., 2021), except here it emerges implicitly from the robustness objective alone, rather than from an explicit generative model or a hybrid generative-discriminative training setup (Grathwohl et al., 2020; Golan et al., 2024). Our primary contribution is the connection (ii), with (iii) and empirical validation for (i).

In this account, spectral concentration explains why noise is suppressed. The semantic content of the directions that survive comes from the class gradients, which we check empirically. Together these give a candidate mechanism linking adversarial training to generation.

To validate this theory, we develop a probing procedure that iteratively applies the implicit $\mathbf{J}^\top\mathbf{J}$ operator through gradient-based updates. This procedure requires no generative training or architectural modifications, and confirms that robust classifiers produce class-consistent structure from noise, directly demonstrating the denoising behavior predicted by our theory. We validate our predictions across MNIST, CIFAR-10, and ImageNet, showing that spectral concentration correlates with generation quality across architectures (CNNs, ViTs) and robustness types (ℓ_2, ℓ_∞).

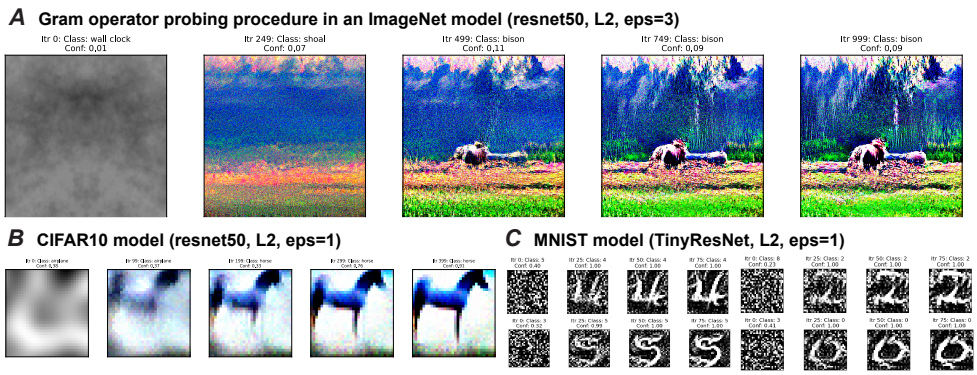

Figure 2: **Probing procedure generates coherent images from input noise using robust classifiers.** Probing procedure applied to **A** a robust ResNet-50 (ImageNet, ℓ_2 adversarially trained with $\epsilon = 3$ and $\epsilon = 5$) **B** a robust ResNet-50 (CIFAR10, ℓ_2 adversarially trained with $\epsilon = 1$ and $\epsilon = 5$) **C** a robust TinyResNet (MNIST, ℓ_2 adversarially trained with $\epsilon = 1$ and $\epsilon = 5$) demonstrates progressive refinement from random noise (same for both shown trajectories) to semantically coherent images. Starting from noise (Itr 0), the algorithm iteratively moves away from a noisy representation of the original input (see Algorithm 1). No explicit generative training was used, only probing procedure on pretrained adversarially robust classifiers (Table 6).

Our contributions are:

1. **Mechanism**: under spectral concentration, $\mathbf{J}^\top\mathbf{J}$ acts as an implicit denoiser (Theorem 1), which gives a candidate explanation for generation from robust classifiers.
2. **Empirical validation**: spectral analysis on MNIST, CIFAR-10, and ImageNet shows the predicted eigenvalue concentration and energy suppression in robust classifiers.
3. **Probing method**: a simple loss whose gradient implicitly applies $\mathbf{J}^\top\mathbf{J}$, usable across architectures and layers without computing the Jacobian.

2 Related Work

Generative capabilities of robust classifiers. The observation that adversarially robust classifiers can synthesize structured images was first demonstrated by Santurkar et al. (2019), with subsequent work attributing this to perceptually aligned gradients (Kaur et al., 2019; Engstrom et al., 2019). These studies document the phenomenon but do not explain its mechanistic origin. Our work provides this missing explanation through the spectral structure of $\mathbf{J}^\top\mathbf{J}$.

Classifiers as generative models. Joint Energy-based Models train classifiers with explicit generative objectives (Grathwohl et al., 2020), and gradient alignment methods optimize for interpretable density models (Singla et al., 2021). Unlike these approaches, we show that generative structure emerges implicitly from adversarial training without any generative loss during training.

Spectral properties and robustness. Jacobian regularization improves robustness by constraining spectral properties (Hoffman et al., 2019; Jakubovitz & Giryes, 2018; Ross & Doshi-Velez, 2017; Rodríguez-Muñoz et al., 2025). Prior work links low-rank Jacobians to generalization and to explicit Jacobian regularization (Oymak et al., 2019; Wu & Li, 2024). We test whether standard adversarial training produces the same structure. We extend this line by showing that low-rank structure enables implicit denoising, connecting spectral properties to generative capability for the first time.

Denoising and generation. The link between denoising and score estimation is well established (Vincent et al., 2008; Vincent, 2011), and underlies modern diffusion models (Ho et al., 2020; Song et al., 2021). We show that robust classifiers contain analogous structure without explicit denoising training.

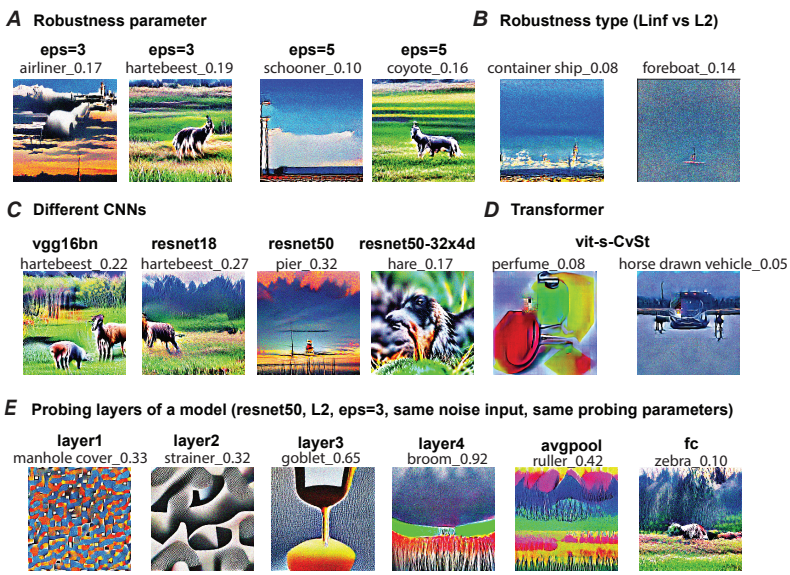

Figure 3: **Probing procedure across architectures and applied at different internal layers.** Probing procedure applied to robust ImageNet classifiers demonstrates consistent semantic generation across: (A) Different robustness parameters ($\varepsilon = 3$ and $\varepsilon = 5$) showing class-consistent outputs for various ImageNet categories; (B) Different adversarial training norms (L_∞ vs L_2) with comparable generation quality; (C) Different CNN architectures (VGG16-BN, ResNet-18, ResNet-50, ResNeXt-50-32x4d) all producing coherent semantic content; (D) Vision Transformer architecture (ViT-s-CvSt (Singh et al., 2023)) demonstrating that implicit denoising structure generalizes beyond CNNs; (E) Different internal layers of ResNet-50 (L_2, $\varepsilon = 3$) showing that the probing procedure can be applied at various representational levels to generate semantically meaningful patterns. All examples use identical noise initialization and the probing procedure parameters (Table 6).

3 Implicit Denoiser Structure in Robust Classifiers

We now formalize how spectral concentration, when present, gives rise to implicit denoising structure in $\mathbf{J}^\top\mathbf{J}$. We take as a premise that adversarial training induces such concentration, supported by our measurements (Section 5, Tables 1–2) and consistent with prior work on Jacobian regularization (Hoffman et al., 2019; Wu & Li, 2024).

Structure of J. We begin by characterizing the Jacobian itself. The rows of $\mathbf{J}(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x})$ are the gradients $\nabla_{\mathbf{x}} f_i(\mathbf{x})$ for each class i. Under the standard energy-based interpretation of classifier logits (Grathwohl et al., 2020; Singla et al., 2021), these correspond to class-conditional score functions: $\nabla_{\mathbf{x}} f_i(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} \mid y{=}i)$. We treat this as an interpretive assumption. Cross-entropy training leaves the logits free up to an input-dependent shift, so they need not be exact density scores (Singla et al., 2021). This gives the Gram operator a concrete meaning:

Proposition 1 (Score-Space Mapping). *Let $f : \mathbb{R}^P \to \mathbb{R}^K$ be a classifier whose logits are interpreted as unnormalized class-conditional log-densities, i.e., $\nabla_{\mathbf{x}} f_i(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} \mid y{=}i)$ for each class i (Grathwohl et al., 2020; Singla et al., 2021). Let $\mathbf{s}_i(\mathbf{x}) = \nabla_{\mathbf{x}} f_i(\mathbf{x})$ denote the i-th class-conditional score. Then for any perturbation $\boldsymbol{\epsilon} \in \mathbb{R}^P$,*

$$\mathbf{J}^\top\mathbf{J}\,\boldsymbol{\epsilon} \;=\; \sum_{i=1}^{K} \langle \mathbf{s}_i(\mathbf{x}),\, \boldsymbol{\epsilon} \rangle\, \mathbf{s}_i(\mathbf{x}), \tag{1}$$

i.e., $\mathbf{J}^\top\mathbf{J}$ maps $\boldsymbol{\epsilon}$ into the subspace spanned by the K class-conditional score functions, weighted by alignment with each score direction. It is an orthogonal projection only if the scores are orthonormal. Consequently, the dominant eigenspace of $\mathbf{J}^\top\mathbf{J}$ corresponds to directions of maximal variation in the learned score field.

Proof. The i-th row of $\mathbf{J}$ is $\nabla_{\mathbf{x}} f_i(\mathbf{x})^\top = \mathbf{s}_i(\mathbf{x})^\top$, so $\mathbf{J} = [\mathbf{s}_1 \ \cdots \ \mathbf{s}_K]^\top$. Then $\mathbf{J}\boldsymbol{\epsilon}$ has i-th entry $\langle \mathbf{s}_i, \boldsymbol{\epsilon} \rangle$, and $\mathbf{J}^\top(\mathbf{J}\boldsymbol{\epsilon}) = \sum_{i=1}^K \langle \mathbf{s}_i, \boldsymbol{\epsilon} \rangle \, \mathbf{s}_i$. □

Thus $\mathbf{J}^\top\mathbf{J}$ maps any input perturbation into the subspace spanned by learned class-conditional score functions.

From projection to denoising. Proposition 1 establishes *what* $\mathbf{J}^\top\mathbf{J}$ maps into. Theorem 1 establishes *how strongly*: under spectral concentration, components outside this subspace shrink much more than those inside it. The combination yields implicit denoising: perturbations aligned with the learned score field are preserved, while orthogonal noise is removed.

3.1 Theoretical Framework: Spectral Decomposition in $\mathbf{J}^\top\mathbf{J}$

Theorem 1 (Spectral Decomposition in $\mathbf{J}^\top\mathbf{J}$). *Let $f : \mathbb{R}^P \to \mathbb{R}^K$ be a neural classifier with Jacobian $\mathbf{J}(\mathbf{x})$ at input $\mathbf{x}$, and let the Gram operator admit the eigendecomposition $\mathbf{J}^\top\mathbf{J} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ with eigenvalues $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_P)$ and orthonormal eigenvectors $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_P]$. For any perturbation $\boldsymbol{\epsilon} = \sum_{i=1}^P c_i \mathbf{v}_i$, we have*

$$\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon} = \sum_{i=1}^P \lambda_i c_i \mathbf{v}_i, \tag{2}$$

$$\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|_2^2 = \sum_{i=1}^P \lambda_i^2 c_i^2. \tag{3}$$

Suppose spectral concentration holds: there exists a small set of directions $\mathcal{S} = \{1, \ldots, k\}$ with $k \ll P$ such that

$$\lambda_1 \geq \cdots \geq \lambda_k \gg \lambda_{k+1} \approx \cdots \approx \lambda_P \approx 0.$$

Only the gap between λ_k and λ_{k+1} is needed; the leading eigenvalues need not be well separated from each other. The set $\mathcal{S}$ corresponds to the input directions along which the classifier's output is most sensitive; whether these directions carry class-relevant structure is verified empirically in Section 5. Define the decomposition $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_\parallel + \boldsymbol{\epsilon}_\perp$ where $\boldsymbol{\epsilon}_\parallel = \sum_{i\in\mathcal{S}} c_i \mathbf{v}_i$ and $\boldsymbol{\epsilon}_\perp = \sum_{i\notin\mathcal{S}} c_i \mathbf{v}_i$. Then for isotropic noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_P)$,

$$\begin{aligned} \mathbb{E}\Big[\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\parallel\|_2^2\Big] &= \sigma^2 \sum_{i\in\mathcal{S}} \lambda_i^2, \\ \mathbb{E}\Big[\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\perp\|_2^2\Big] &= \sigma^2 \sum_{i\notin\mathcal{S}} \lambda_i^2 \approx 0. \end{aligned} \tag{4}$$

In particular, under spectral concentration $\sum_{i\notin\mathcal{S}} \lambda_i^2 \ll \sum_{i\in\mathcal{S}} \lambda_i^2$, we obtain

$$\mathbb{E}\Big[\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\parallel\|_2^2\Big] \gg \mathbb{E}\Big[\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\perp\|_2^2\Big], \tag{5}$$

i.e., $\mathbf{J}^\top\mathbf{J}$ shrinks components orthogonal to $\mathcal{S}$ much more than components aligned with $\mathcal{S}$. Thus low-rank spectral structure gives $\mathbf{J}^\top\mathbf{J}$ an implicit denoising effect.

Proof. By expanding $\boldsymbol{\epsilon}$ in the eigenbasis $\{\mathbf{v}_i\}$, we have $\boldsymbol{\epsilon} = \sum_{i=1}^P c_i \mathbf{v}_i$ and, since $\mathbf{J}^\top\mathbf{J}\mathbf{v}_i = \lambda_i \mathbf{v}_i$,

$$\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon} = \sum_{i=1}^P \lambda_i c_i \mathbf{v}_i, \quad \|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|_2^2 = \sum_{i=1}^P \lambda_i^2 c_i^2.$$

For $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_P)$, the coefficients c_i are independent with $\mathbb{E}[c_i^2] = \sigma^2$. Therefore,

$$\mathbb{E}\Big[\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\parallel\|_2^2\Big] = \mathbb{E}\Big[\sum_{i\in\mathcal{S}} \lambda_i^2 c_i^2\Big] = \sigma^2 \sum_{i\in\mathcal{S}} \lambda_i^2,$$

and similarly

$$\mathbb{E}\Big[\|\mathbf{J}^\top \mathbf{J}\boldsymbol{\epsilon}_\perp\|_2^2\Big] = \sigma^2 \sum_{i \notin \mathcal{S}} \lambda_i^2.$$

Under spectral concentration, the latter sum is negligible compared to the former, yielding

$$\mathbb{E}\Big[\|\mathbf{J}^\top \mathbf{J}\boldsymbol{\epsilon}_\parallel\|_2^2\Big] \gg \mathbb{E}\Big[\|\mathbf{J}^\top \mathbf{J}\boldsymbol{\epsilon}_\perp\|_2^2\Big].$$

This shows that $\mathbf{J}^\top\mathbf{J}$ concentrates energy in the discriminative subspace $\mathcal{S}$ and suppresses orthogonal components, providing the claimed energy concentration and implicit denoising behavior. □

Remark 1. *Theorem 1 takes spectral concentration as a premise. We verify empirically that it holds in adversarially trained networks (Section 5); prior work connects related structure to explicit Jacobian regularization (Hoffman et al., 2019; Wu & Li, 2024). Our contribution is the implication: given this spectral structure,* $\mathbf{J}^\top\mathbf{J}$ *acts as a denoiser.*

When $\mathbf{J}$ is the Jacobian of the logits, $\mathbf{J} \in \mathbb{R}^{K\times P}$ with $K \ll P$, so $\mathrm{rank}(\mathbf{J}^\top\mathbf{J}) \leq K$. For MNIST ($K = 10$, $P = 784$), the operator has at most 10 non-zero eigenvalues, providing a natural upper bound for $|\mathcal{S}|$.

Combining Proposition 1 and Theorem 1: $\mathbf{J}^\top\mathbf{J}$ maps into the span of class-conditional scores (Proposition 1), and spectral concentration ensures this mapping strongly suppresses orthogonal components (Theorem 1). Whether the preserved directions carry class-relevant structure is an empirical question, which we verify in Section 5. This framework makes several testable predictions: (1) the energy ratio $\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|/\|\boldsymbol{\epsilon}\|$ should be much smaller for robust than standard models, (2) applying $\mathbf{J}^\top\mathbf{J}$ to random noise should reveal class-consistent structure in robust classifiers, and (3) robust models should have much smaller tail eigenvalues and overall operator magnitude ($\mathrm{tr}(\mathbf{J}^\top\mathbf{J})$) than standard models.

4 Empirical Validation via $\mathbf{J}^\top\mathbf{J}$ Probing

Theorem 1 predicts that $\mathbf{J}^\top\mathbf{J}$ in robust classifiers acts as an implicit denoiser. We now describe a probing procedure that tests this prediction by iteratively applying the operator and observing whether class-consistent structure emerges from noise.

Directly computing $\mathbf{J}^\top\mathbf{J}$ is intractable for high-dimensional inputs. For an image with P pixels, storing the full Jacobian requires $O(PK)$ memory and explicit matrix multiplication requires $O(P^2K)$ operations. We need a procedure that applies $\mathbf{J}^\top\mathbf{J}$ implicitly.

Probing objective. We construct a simple objective whose gradient implicitly applies $\mathbf{J}^\top\mathbf{J}$. The key insight is that moving an input away from a noisy version of itself, in representation space, naturally invokes the Gram operator.

For noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and network layer $r(\cdot)$, define:

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\epsilon}) = \big\| \, r(\mathbf{x}) \; - \; \mathrm{sg}[r(\mathbf{x} + \boldsymbol{\epsilon})] \, \big\|_2^2, \tag{6}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Taking the gradient with respect to $\mathbf{x}$:

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}; \boldsymbol{\epsilon}) = 2\,\mathbf{J}_r(\mathbf{x})^\top\big(r(\mathbf{x}) - \mathrm{sg}[r(\mathbf{x} + \boldsymbol{\epsilon})]\big). \tag{7}$$

Using the first-order Taylor approximation $r(\mathbf{x} + \boldsymbol{\epsilon}) \approx r(\mathbf{x}) + \mathbf{J}_r(\mathbf{x})\boldsymbol{\epsilon}$:

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}; \boldsymbol{\epsilon}) \approx -2\,\mathbf{J}_r(\mathbf{x})^\top\mathbf{J}_r(\mathbf{x})\,\boldsymbol{\epsilon}. \tag{8}$$

Thus, gradient ascent on $\mathcal{L}$ implicitly applies $-\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}$, the denoising operation predicted by Theorem 1.

Probing procedure. Starting from noise $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, we iteratively update:

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta\,\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_t; \boldsymbol{\epsilon}) \approx \mathbf{x}_t - 2\eta\,\mathbf{J}_r(\mathbf{x}_t)^\top\mathbf{J}_r(\mathbf{x}_t)\,\boldsymbol{\epsilon}, \tag{9}$$

for T iterations. Thus, $\mathbf{J}^\top\mathbf{J}$ emerges naturally from the simple objective of moving away from noisy representations. Note that the first-order approximation (Eq. 8) explains why the probing objective

invokes $\mathbf{J}^\top\mathbf{J}$; in practice, each update uses the exact gradient of Eq. 6 computed via automatic differentiation. In practice (Algorithm 1), the noisy reference $r(\mathbf{x}_0 + \boldsymbol{\epsilon})$ is computed once and held fixed, so the first step applies $\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}$ exactly, and later steps apply $\mathbf{J}^\top\mathbf{J}$ to the current offset from this reference, which holds approximately while the iterate stays near it. Each update is also normalized and projected onto an ℓ_2 ball around $\mathbf{x}_0$ to bound the step size and trajectory; we do not analyze the effect of these two steps separately. A single application of $\mathbf{J}^\top\mathbf{J}$ maps noise into the dominant subspace; the iterative procedure progressively refines this mapping as the input moves toward structured regions of input space, analogous to the multi-step denoising process in diffusion models. The full algorithm (Algorithm 1) includes an optional stochastic diffusion term η_t scaled by a *diffusion noise ratio* parameter, which adds small random perturbations at each step to prevent convergence to sharp local optima.

The probing procedure provides a direct test of our theoretical framework. For robust classifiers, Theorem 1 predicts that under strong spectral concentration, $\mathbf{J}^\top\mathbf{J}$ maps noise into the discriminative subspace, producing structured, class-consistent outputs from noise. For standard classifiers, weak spectral concentration means $\mathbf{J}^\top\mathbf{J}$ largely preserves the noise, yielding unstructured outputs. We test these predictions in Section 5.

5 Results

We evaluate our proposed theoretical framework through two complementary experimental approaches. First, we validate the spectral properties predicted by our theory using MNIST classifiers, where computational tractability allows detailed Jacobian analysis. Second, we demonstrate the practical generative capabilities of probing procedure across datasets and architectures, showing that implicit denoising structure enables coherent image generation from robust classifiers. Unless otherwise stated, $r(\cdot)$ in equation (6) refers to the logits (pre-softmax) layer of the classifier throughout all experiments.

5.1 Spectral Analysis: Validating the Implicit Denoiser Hypothesis

Notation. Let $\mathbf{J}(\mathbf{x})$ denote the Jacobian of the classifier at input $\mathbf{x}$, and $\mathbf{G}(\mathbf{x}) = \mathbf{J}(\mathbf{x})^\top\mathbf{J}(\mathbf{x})$ the Gram operator. Let $\lambda_1 \geq \lambda_2 \geq \cdots$ denote its eigenvalues. We define the *energy ratio* as $\|\mathbf{G}(\mathbf{x})\epsilon\|_2/\|\epsilon\|_2$ for $\epsilon \sim \mathcal{N}(0, I)$, and *tail@k* as λ_k, the k-th largest eigenvalue. Since $\mathbf{J} \in \mathbb{R}^{K\times P}$, the operator $\mathbf{J}^\top\mathbf{J}$ has at most K non-zero eigenvalues. In practice, the 90% cumulative energy threshold $\sum_{i\leq k}\lambda_i^2/\sum_i\lambda_i^2 \geq 0.9$ is crossed at $k = 2$–3 consistently across all robust checkpoints (see Appendix, Figure S7). We refer to the energy ratio as *Gain* in Table 2. The *soft-rank* is the entropy-based effective rank $\exp(H(\hat{\boldsymbol{\lambda}}))$, where $\hat{\lambda}_i = \lambda_i/\sum_j\lambda_j$ and H is the Shannon entropy; it measures how many eigenvalues contribute meaningfully to the spectrum. R^2 denotes the coefficient of determination for the quadratic fit $\Delta_{\text{robust}}(\sigma) \approx \alpha\,\sigma^2$.

Eigenvalue decay and energy ratios. Eigenvalue analysis shows strong tail suppression in robust models, a $49\times$ reduction (tail@k: $8.251 \rightarrow 0.167$). While the leading eigenvalue ratios show modest improvement (λ_1/λ_2: $1.47 \rightarrow 1.96$), the critical denoising mechanism operates through suppression of orthogonal noise directions, as evidenced by the dramatic tail eigenvalue reduction. Our quantitative analysis (Table 1) shows robust classifiers achieve dramatically stronger denoising: energy ratio statistics ($\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|/\|\boldsymbol{\epsilon}\|$) decrease from 6.19 in standard models to 0.16 in robust models, a $40\times$ improvement. The robust classifier also exhibits a near-perfect fit to the theoretical robustness relationship ($R^2 = 0.992$ vs 0.717), confirming our framework's predictive accuracy.

Visual residuals reveal class priors. Applying $\mathbf{J}^\top\mathbf{J}$ (evaluated at a test image) to random noise produces dramatically different outcomes in robust versus standard classifiers (Figure 1D). Standard models preserve the random structure of input noise, while robust models extract digit-like patterns aligned with learned class priors. When images are individually normalized to reveal structure, robust models consistently produce recognizable features that correspond to the network's predictions, demonstrating that $\mathbf{J}^\top\mathbf{J}$ acts as a denoiser that keeps class-relevant structure.

Spectral concentration scales to ImageNet. Table 2 extends our spectral analysis to ImageNet-scale models, confirming that the same patterns observed on MNIST hold at scale. Robust models show much smaller trace, energy ratios, and tail eigenvalues than the standard model. Soft-rank is low

Table 1: **Quantitative verification of the implicit denoiser hypothesis.** Robust models show (1) dramatically smaller Jacobian spectrum (reduced λ_1, λ_2, and tail@8), (2) lower energy amplification $\|\mathbf{G}\epsilon\|/\|\epsilon\|$, and (3) stronger quadratic noise response (R^2).

Model	λ_1	λ_2	λ_1/λ_2	tail@k	$\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|/\|\boldsymbol{\epsilon}\|$	R^2
Standard	116.8	79.57	1.47	8.251	6.19	0.717
Robust L2	3.166	1.612	1.96	0.167	0.16	0.992
Ratio	0.027×	0.020×	1.33×	0.020×	0.026×	1.38×

Table 2: **Spectral statistics for models trained on ImageNet.** Spectral properties of ImageNet ResNet-50 models under varying ℓ_2 adversarial robustness. Columns report noise suppression (Gain), Hutchinson trace $\mathrm{tr}(J^\top J)$, leading eigenvalues, normalized eigenvalue ratio, spectral tail, entropy-based soft-rank, and Probing procedure denoising performance (ΔPSNR).

Model	**Gain**	**tr($\mathbf{J}^\top\mathbf{J}$)**	λ_1	λ_2	λ_1/λ_2	**tail@k**	**soft-rank**	**ΔPSNR (dB)**
$\epsilon = 0.00$ (Std.)	116.28	152,042.0	34,815.5	18,709.4	1.86	4,538.0	4.77	-2.80
$\epsilon = 1.00$	0.0044	7.37	1.82	0.82	2.22	0.11	4.04	-0.80
$\epsilon = 2.00$	0.0012	1.28	0.45	0.18	2.51	0.03	4.13	1.48
$\epsilon = 3.00$	0.0001	0.24	0.14	0.06	2.26	0.01	4.33	1.21
$\epsilon = 5.00$	0.0002	0.16	0.08	0.03	2.34	0.01	4.42	0.83

($\sim$4–5) for all models, including the standard one, so robust training mainly shrinks the operator's overall magnitude and leaves the shape of the normalized spectrum largely unchanged.

5.2 Probing Reveals Class-Consistent Structure in Robust Classifiers

Robust classifiers produce coherent semantic structure from noise. To test whether $\mathbf{J}^\top\mathbf{J}$ functions as an implicit denoiser as predicted by Theorem 1, we apply the probing procedure to ResNet-50 models (He et al., 2016) trained on ImageNet (Deng et al., 2009) with different ℓ_2 perturbation budgets ($\epsilon = 3$ vs $\epsilon = 5$) using PGD adversarial training (Madry et al., 2018). Starting from identical noise patterns, the procedure converges to semantically distinct but equally coherent patterns: parachute/landscape scenes versus container ship/seascape scenes (Figure 3). This matches the theoretical prediction: the $\mathbf{J}^\top\mathbf{J}$ operator in robust classifiers maps noise toward class-consistent structure. Different robustness constraints create distinct semantic attractors, reflecting how the choice of ϵ shapes the spectral properties of the learned Jacobian.

The implicit denoiser generalizes across architectures and layers. If spectral concentration is a fundamental property of adversarial training rather than an architecture-specific artifact, the probing procedure should reveal class-consistent structure across diverse networks. Figure 3 confirms this prediction across: (A) different robustness parameters, (B) ℓ_∞ vs ℓ_2 adversarial training norms (Madry et al., 2018), (C) multiple CNN architectures including VGG16-BN (Simonyan & Zisserman, 2014), ResNet-18, ResNet-50, and ResNeXt-50-32x4d (Xie et al., 2017), (D) Vision Transformer architecture (ViT-S-CvSt (Singh et al., 2023)), and (E) different internal layers of ResNet-50. Proposition 1 assumes $r(\cdot)$ is the logits layer, so its score interpretation does not cover the intermediate-layer results in (E), which we report as empirical observations. Starting from identical noise, the procedure consistently produces semantically coherent, class-consistent patterns across all conditions. The diversity of semantic categories, from animals and vehicles to natural scenes, demonstrates that the implicit denoising structure of $\mathbf{J}^\top\mathbf{J}$ scales across ImageNet's 1000 categories. The consistency across architectures (CNNs and ViTs) supports our claim that spectral concentration is induced by adversarial training itself, not by specific architectural choices.

ℓ_2 training produces stronger denoising than ℓ_∞. Comparing models trained with ℓ_2 vs ℓ_∞ adversarial training at comparable robustness levels, we find ℓ_2 models produce more semantically coherent outputs (Figure 3B). This is consistent with the geometric alignment between ℓ_2 robustness and Gaussian noise: diffusion models universally use Gaussian noise, and ℓ_2 training creates structure matched to this geometry, while ℓ_∞ training optimizes for hypercube perturbations.

Generation vs. interpolation. A nearest-neighbor analysis against the full CIFAR-10 training set, in both pixel and logit space, found no direct reproduction or interpolation of individual training examples in the probing procedure outputs (Appendix E, Figure S9).

6 LIMITATIONS

Our detailed spectral analysis of $\mathbf{J}^\top\mathbf{J}$ eigenstructure is limited to MNIST due to computational cost; for CIFAR-10 and ImageNet we rely on efficient approximations (trace estimation, top-k eigenvalues). While these confirm the same qualitative trends, full spectral characterization at scale remains future work. Our spectral measurements are also taken at test images, while probing starts from noise far from the data; we do not measure whether the same spectral structure holds early in the probing trajectory.

Although we provide correlational evidence linking robustness, spectral properties, and denoising capability—increasing the adversarial training budget ϵ systematically shrinks the tail eigenvalues and operator magnitude as predicted by our framework, and this correlates with improved quality of patterns produced by the probing procedure—we do not establish direct causation. Perturbing gradients post-hoc (e.g., adding noise or truncating eigenvalues during inference) would be trivial but uninformative, as such interventions conflate the learned structure with the perturbation itself. A more rigorous test would involve training networks under explicit Jacobian regularization (Jakubovitz & Giryes, 2018; Hoffman et al., 2019) with varying regularization strengths, then measuring whether the resulting spectral concentration predicts denoising capability independently of adversarial training. Such experiments could reveal whether spectral concentration is necessary, sufficient, or merely correlated with implicit denoising structure, and whether the same generative properties emerge from Jacobian regularization alone without adversarial examples.

7 CONCLUSION

This work proposes a mechanistic explanation for the generative capabilities of adversarially robust classifiers. We show that spectral concentration, which we observe in adversarially trained models from MNIST to ImageNet, is a condition under which the Gram operator $\mathbf{J}^\top\mathbf{J}$ acts as an implicit denoiser that suppresses orthogonal noise much more than class-aligned components (Theorem 1).

Our empirical validation confirms the theoretical predictions across MNIST, CIFAR-10, and ImageNet: robust classifiers exhibit lower energy ratios, much smaller tail eigenvalues, and produce class-consistent structure when probed with iterative $\mathbf{J}^\top\mathbf{J}$ application. The consistency across architectures (multiple CNNs, ViTs) suggests that spectral concentration is a general consequence of adversarial training.

By revealing the mathematical structure underlying emergent generative properties, this work provides a foundation for understanding the dual discriminative-generative nature of robust representations, with implications for interpretability, neuroscience, robustness and the design of models that jointly optimize for robustness and generation. In particular, our results suggest that the interpretability of input gradients, inception loops and neural activation interventions in neuroscience may reflect querying the implicit learned prior over natural images, yielding image-space updates that resemble denoising dynamics and, potentially, score-like behavior.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks with required disclosure under the ICLR 2027 AI Policy for Authors, including developing the theoretical framework, formulating or proving mathematical claims, designing experiments, implementing methods, or interpreting results. Additionally, we used generative AI tools to assist with drafting and editing text for readability, formatting the manuscript and references, and suggesting revisions based on LLM-generated feedback on earlier drafts. We have reviewed all AI-assisted work: every suggested edit was read by the authors, checked against our results, and either accepted, rewritten, or discarded. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work advances understanding of why adversarially robust classifiers exhibit generative capabilities, with applications in ML interpretability and computational neuroscience. The probing procedure operates on pretrained classifiers and does not introduce new generative models. We do not foresee specific negative societal consequences from this theoretical contribution.

REFERENCES

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? In *International Conference on Machine Learning*, pp. 10628–10648. PMLR, 2023.

Guy Gaziv, Michael J. Lee, and James J. DiCarlo. Robustified anns reveal wormholes between human category percepts, 2023. URL https://arxiv.org/abs/2308.06887.

Guy Gaziv, Sarah Goulding, Ani Ayvazian-Hancock, Yoon Bai, and James J. DiCarlo. Noninvasive precision modulation of high-level neural population activity via natural vision perturbations, 2025. URL https://arxiv.org/abs/2506.05633.

Shelly Golan, Roy Ganz, and Michael Elad. Enhancing consistency-based image generation via adversarialy-trained classification and energy-based discrimination. *Advances in Neural Information Processing Systems*, 37:47966–47986, 2024.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 514–529, 2018.

Simran Kaur, Jeremy Cohen, and Zachary C Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *arXiv preprint arXiv:1910.08640*, 2019.

Zico Kolter and Aleksander Madry. Adversarial robustness: Theory and practice (neurips 2018 tutorial). https://adversarial-ml-tutorial.org/, 2018. Tutorial notes and code; accessed 2025.

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

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Adrián Rodríguez-Muñoz, Tongzhou Wang, and Antonio Torralba. Characterizing model robustness via natural input gradients. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 161–178, Cham, 2025. Springer Nature Switzerland.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. 2017.

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems*, pp. 1260–1271, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Naman D Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models, 2023. URL https://arxiv.org/abs/2303.01870.

Suraj Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Suraj Srinivas, Sebastian Bordt, and Hima Lakkaraju. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness, 2024. URL https://arxiv.org/abs/2305.19101.

Louis Thiry and Florentin Guth. Classification-denoising networks. *arXiv preprint arXiv:2410.03505*, 2024.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, July 2011. ISSN 0899-7667. doi: 10.1162/NECO_a_00142. URL https://doi.org/10.1162/NECO_a_00142.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 supports topological organization. *bioRxiv*, 2025. doi: 10.1101/2023.05.12.540591. URL https://www.biorxiv.org/content/early/2025/11/05/2023.05.12.540591.

Dongya Wu and Xin Li. Adversarially robust generalization theory via jacobian regularization for deep neural networks. *arXiv preprint arXiv:2412.12449*, 2024.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

A Appendix

B Spectral Experiments setup

All MNIST and spectral experiments in this paper were conducted using the reference CNN and training code from the NeurIPS 2018 tutorial *Adversarial Robustness: Theory and Practice* by Zico Kolter and Aleksander Madry Kolter & Madry (2018), unless otherwise explicitly mentioned (e.g. TinyResNet architecture).

ℓ_2-PGD attack (training and evaluation). All adversarial training and evaluation use ℓ_2-PGD with: $\epsilon = 1.5$, step size $\alpha = 0.2$, 20 iterations, projection onto the ℓ_2-ball per example, and clamping to $[0, 1]$. Random start was disabled unless otherwise noted.

Spectral metrics. For each checkpoint we compute Jacobian-based spectral quantities on held-out data, including the energy ratio derived from $Q = \mathbf{J}^\top \mathbf{J}$ (expectations over multiple random ξ), and power-law exponents fit to the spectrum. These metrics are used to map $(\varepsilon, \text{epoch})$ phase diagrams.

Accuracies. Final clean and robust accuracies will be reported in Table 3.

Table 3: MNIST CNNs for Figure 1 and Table 1. Robust accuracy measured under ℓ_2-PGD ($\epsilon = 1.5$, 20 steps).

Model	Clean Acc.	Robust Acc.
Clean CNN	0.98	0.02
Robust CNN	0.98	0.47

C Supplementary: Implicit Denoiser Probes

To produce the quantitative results reported in Table 1 of the main text, we ran three complementary probes designed to test the hypothesis that $\mathbf{J}^\top \mathbf{J}$ acts as an implicit denoiser in robust classifiers. All code is adapted from our MNIST experimental framework (see Section S1). Below, we summarize the key implementation details.

Energy ratios. For a held-out test input x, we repeatedly sample isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and compute

$$r = \frac{\|\mathbf{J}^\top \mathbf{J}\epsilon\|}{\|\epsilon\|}.$$

We report the mean of r across $n = 128$ draws, with $\sigma = 0.25$. Lower values indicate stronger denoising by suppression of noise directions.

Eigenvalue decay. We estimate the leading eigenvalues of $\mathbf{J}^\top \mathbf{J}$ at test points using subspace iteration with Gram–Schmidt re-orthogonalization. Starting from k random vectors (we use $k = 8$), we repeatedly apply $\mathbf{J}^\top \mathbf{J}$ and orthogonalize, then compute Rayleigh quotients $\lambda_i = \langle v_i, \mathbf{J}^\top \mathbf{J} v_i \rangle$. We report the sorted eigenvalues and the λ_1/λ_2 ratio to capture spectral dominance.

Visual residuals. For qualitative inspection, we sample a random ϵ at fixed scale $\sigma = 0.35$ and display both the raw input noise and its transformation $\mathbf{J}^\top \mathbf{J}\epsilon$. This illustrates that robust models suppress noise and emphasize structured residuals, consistent with implicit denoising.

C.1 Spectral properties for increasing robustness across MNIST, CIFAR10 and ImageNet

We analyze the spectral properties of TinyResNet models trained on MNIST with varying levels of adversarial robustness. The TinyResNet architecture is a lightweight ResNet variant implemented

using the Hugging Face transformers library, consisting of two stages with basic residual blocks. Specifically, the architecture uses depths $[2, 2]$ (two blocks per stage) and hidden channel sizes $[32, 64]$ (32 channels in the first stage, 64 in the second), with skip connections enabling gradient flow through the network. Models are trained using L2 PGD adversarial training with training perturbation budgets $\varepsilon_{\text{train}} \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$.

Figure S1 visualizes how the Jacobian transpose-Jacobian operator $\mathbf{J}^\top\mathbf{J}$ transforms random input noise ε for models with increasing robustness. The computation proceeds as follows: (1) a single random noise pattern ε is generated and normalized to $[0, 1]$; (2) for each model, we compute $\mathbf{J}^\top\mathbf{J}\varepsilon$ at a test image x using automatic differentiation, where $\mathbf{J}$ is the Jacobian of the model's logit output with respect to the input; (3) the transformation $\mathbf{J}^\top\mathbf{J}\varepsilon$ is computed implicitly via Jacobian-vector products (JVP) followed by vector-Jacobian products (VJP) to avoid explicitly constructing the full Jacobian matrix.

The figure demonstrates a key spectral property: as the training perturbation budget $\varepsilon_{\text{train}}$ increases, the operator $\mathbf{J}^\top\mathbf{J}$ becomes more selective, extracting clearer patterns from random noise. For the standard model ($\varepsilon_{\text{train}} = 0.00$), the transformed residual remains largely noisy, while robust models ($\varepsilon_{\text{train}} \geq 0.50$) reveal distinct digit-like structures, indicating that $\mathbf{J}^\top\mathbf{J}$ acts as a pattern extraction operator whose selectivity increases with adversarial training. This visualization provides an intuitive understanding of how robust models develop more structured input-output relationships, with the Jacobian transpose-Jacobian operator effectively filtering noise to reveal the underlying data manifold structure.

Table 4 summarizes the quantitative spectral properties across all models. As robustness increases, we observe several key trends: (1) the trace $\text{tr}(\mathbf{J}^\top\mathbf{J})$ decreases monotonically, indicating reduced overall sensitivity; (2) the energy ratio $\|\mathbf{J}^\top\mathbf{J}\varepsilon\|/\|\varepsilon\|$ decreases, showing that the operator becomes more selective; (3) the top eigenvalues λ_1 and λ_2 decrease, with their ratio λ_1/λ_2 initially increasing then stabilizing, suggesting a shift in the eigenspectrum structure; (4) the R^2 values for the $\Delta_{\text{robust}} \approx \alpha \cdot \sigma^2$ relationship improve with robustness, indicating better predictability of robustness gaps. These quantitative measures complement the visual analysis, confirming that adversarial training fundamentally alters the spectral properties of the model's input-output mapping.

Table 4: Spectral properties summary for TinyResNet models trained with varying adversarial perturbation budgets $\varepsilon_{\text{train}}$ on MNIST. Clean Acc and Robust Acc denote clean and robust test accuracies, respectively. The trace $\text{tr}(\mathbf{J}^\top\mathbf{J})$ measures overall operator magnitude, while $\|\mathbf{J}^\top\mathbf{J}\varepsilon\|/\|\varepsilon\|$ is the median energy ratio. The eigenvalues λ_1 and λ_2 are the top two eigenvalues of $\mathbf{J}^\top\mathbf{J}$.

Model	Clean Acc	Robust Acc	$\text{tr}(\mathbf{J}^\top\mathbf{J})$	$\frac{\|\mathbf{J}^\top\mathbf{J}\varepsilon\|}{\|\varepsilon\|}$	λ_1	λ_2	$\frac{\lambda_1}{\lambda_2}$
$\varepsilon = 0.00$	0.9844	0.9844	127.1337	1.1854	33.5095	19.6171	1.71
$\varepsilon = 0.25$	0.9882	0.9770	58.8217	0.6695	19.0332	10.2712	1.85
$\varepsilon = 0.50$	0.9898	0.9678	41.3736	0.3697	10.8113	5.6020	1.93
$\varepsilon = 0.75$	0.9891	0.9533	31.7437	0.3725	11.6364	3.5081	3.32
$\varepsilon = 1.00$	0.9911	0.9410	23.8401	0.2457	7.2376	3.4450	2.10

We extend this spectral analysis to more complex datasets and architectures to validate the generality of our findings. Figure S2 demonstrates the same spectral concentration phenomenon on CIFAR-10 using ResNet-50 models trained with varying adversarial perturbation budgets $\varepsilon_{\text{train}} \in \{0.00, 0.25, 0.50, 1.00\}$. The top row shows five different random noise samples ε, while subsequent rows display the visual residuals $\mathbf{J}^\top\mathbf{J}\varepsilon$ for each model. Consistent with our MNIST findings, standard CIFAR-10 classifiers ($\varepsilon = 0.00$) preserve much of the random structure in the noise, producing residuals that closely resemble the input patterns. In contrast, robust classifiers show progressive noise suppression and structure extraction as training robustness increases. For $\varepsilon = 1.00$, the operator $\mathbf{J}^\top\mathbf{J}$ transforms random noise into coherent spatial patterns with clear edge structures and color organization, demonstrating that spectral concentration enables semantic pattern extraction even in complex natural image datasets.

Figure S3 provides validation at the largest scale, showing spectral residuals for ResNet-50 models trained on ImageNet with perturbation budgets $\varepsilon_{\text{train}} \in \{0.00, 3.00, 5.00\}$. Despite the significantly higher complexity of ImageNet (1000 classes, 224×224 resolution), the fundamental spectral properties remain consistent. Standard ImageNet classifiers produce residuals that largely preserve the

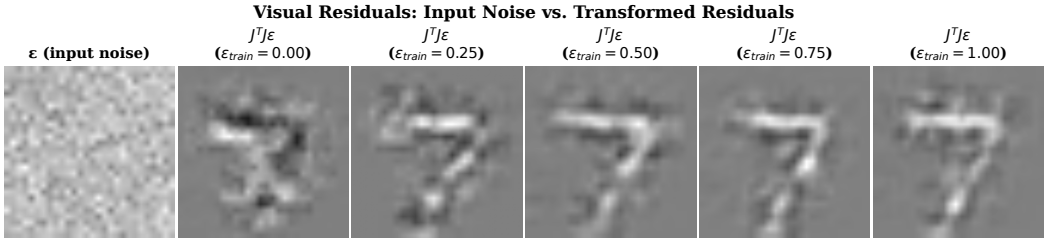

Figure S1: Visual residuals showing the transformation of random input noise ε by the operator $\mathbf{J}^\top\mathbf{J}$ for TinyResNet models trained with increasing adversarial perturbation budgets $\varepsilon_{\text{train}}$. The leftmost panel shows the input noise pattern, while subsequent panels display $\mathbf{J}^\top\mathbf{J}\varepsilon$ for models trained with $\varepsilon_{\text{train}} \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$. As robustness increases, the operator becomes more selective, extracting clearer digit-like patterns from the noise, demonstrating how adversarial training shapes the spectral properties of the model's input-output mapping.

Table 5: **ImageNet spectral statistics.** Spectral properties of ImageNet ResNet-50 models under varying ℓ_2 adversarial robustness. Columns report noise suppression (Gain), Hutchinson trace $\mathrm{tr}(J^\top J)$, leading eigenvalues, normalized eigenvalue ratio, spectral tail, entropy-based soft-rank, and Probing procedure denoising performance (ΔPSNR).

Model	**Gain**	**tr($\mathbf{J}^\top\mathbf{J}$)**	λ_1	λ_2	λ_1/λ_2	**tail@k**	**soft-rank**	**ΔPSNR (dB)**
$\epsilon = 0.00$ (Std.)	116.28	152,042.0	34,815.5	18,709.4	1.86	4,538.0	4.77	-2.80
$\epsilon = 1.00$	0.0044	7.37	1.82	0.82	2.22	0.11	4.04	-0.80
$\epsilon = 2.00$	0.0012	1.28	0.45	0.18	2.51	0.03	4.13	1.48
$\epsilon = 3.00$	0.0001	0.24	0.14	0.06	2.26	0.01	4.33	1.21
$\epsilon = 5.00$	0.0002	0.16	0.08	0.03	2.34	0.01	4.42	0.83

random structure of input noise, while robust models ($\varepsilon = 3.00, 5.00$) generate structured patterns with recognizable spatial organization and semantic coherence. The residuals reveal edge-like structures, texture patterns, and spatial arrangements that suggest the network has learned to extract meaningful visual features from random inputs. This cross-dataset consistency—from simple MNIST digits to complex natural images—provides strong empirical support for our theoretical framework.

The quantitative trends observed in MNIST (Table 4) are replicated across all three datasets, with energy ratios $\|\mathbf{J}^\top\mathbf{J}\varepsilon\|/\|\varepsilon\|$ decreasing monotonically with increasing robustness and trace values $\mathrm{tr}(\mathbf{J}^\top\mathbf{J})$ showing systematic reduction. This systematic behavior across datasets of varying complexity and different network architectures suggests that spectral concentration is a fundamental consequence of adversarial training rather than a dataset-specific artifact. The visual evidence from Figures S1, S2, and S3 collectively demonstrate that the Jacobian transpose-Jacobian operator $\mathbf{J}^\top\mathbf{J}$ transitions from a noise-preserving transformation in standard networks to a semantic pattern extractor in robust networks, providing the mechanistic foundation for the generative capabilities we observe through the probing procedure.

ImageNet spectral summary. Finally, our quantitative analysis on ImageNet (Table 2), together with the normalized eigenvalue decay shown in Figure S4, confirms that the same spectral trends observed on MNIST and CIFAR-10 persist at large scale. The standard ImageNet model exhibits extremely high energy ratios and large trace values, indicating strong sensitivity to random perturbations and a broadly distributed spectrum. In contrast, robust ImageNet models show two hallmarks of implicit denoiser structure: (1) *aggressive noise suppression*, with gains reduced by more than five orders of magnitude relative to the standard model, and (2) *tail suppression*, with tail eigenvalues and trace several orders of magnitude smaller. Soft-rank ($\sim$4–5) and the shape of the normalized eigenvalue decay are similar for standard and robust models, so the difference lies in the overall magnitude of the operator. These results indicate that the same pattern of noise suppression seen on MNIST holds for high-resolution, 1000-class ImageNet models.

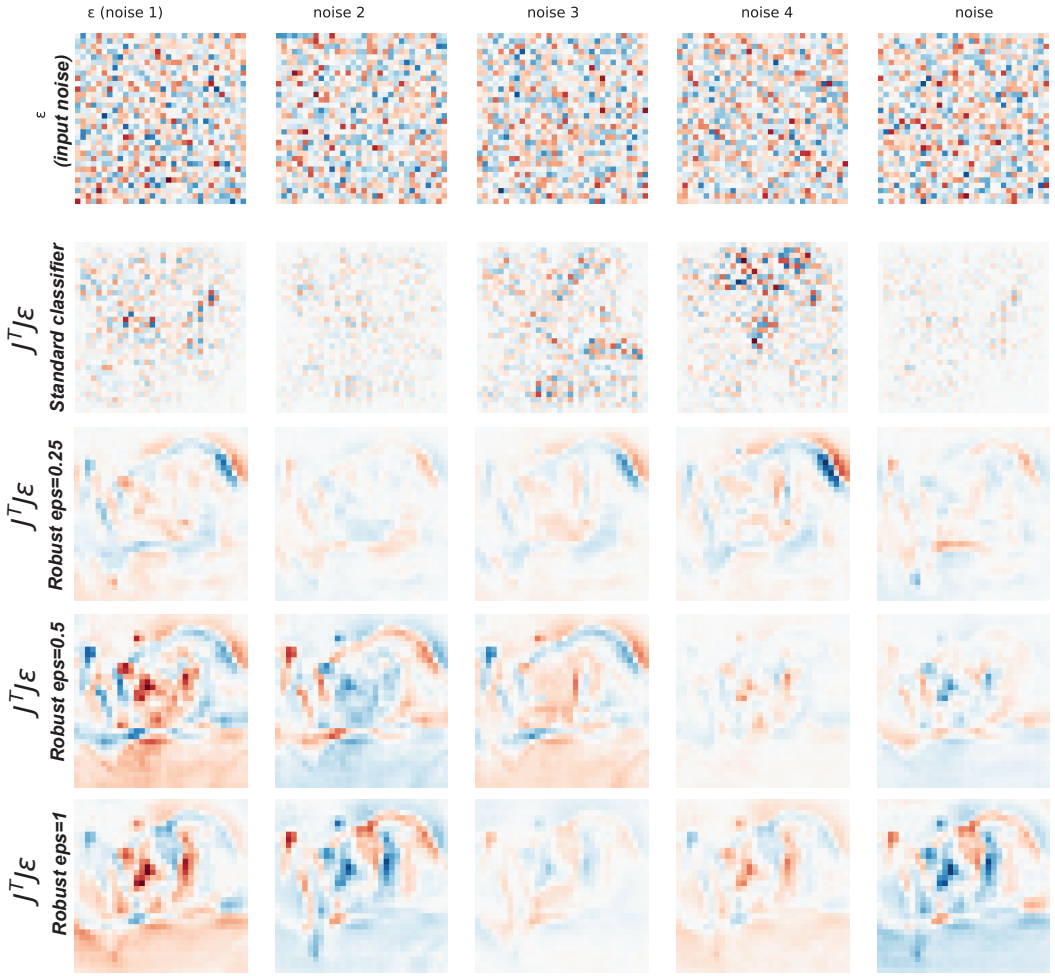

Figure S2: Visual residuals showing the action of $\mathbf{J}^\top\mathbf{J}$ on random noise for CIFAR10 ResNet50 models with varying robustness. **Top row:** Five different input noise samples (ε). **Subsequent rows:** Visual residuals ($\mathbf{J}^\top\mathbf{J}\varepsilon$) for a standard classifier ($\varepsilon = 0.00$) and robust classifiers with increasing adversarial training strength ($\varepsilon = 0.25, 0.50, 1.00$). All visualizations use symmetric normalization (red-blue diverging colormap) with per-model scaling.

C.2 Additional Analyses on Spectral Properties

C.3 Probing Procedure: Algorithm and Implementation Details

We present the detailed algorithm for the $\mathbf{J}^\top\mathbf{J}$ probing procedure used to validate the implicit denoiser hypothesis. The procedure is designed to iteratively apply the Gram operator through gradient-based updates, testing whether robust classifiers produce class-consistent structure from noise as predicted by Theorem 1. Probing procedure parameters for main text figures are provided in Table 6. The MSE objective (Eq. 6) is preferred over alternatives such as class-score maximization because: (i) its gradient recovers the full Gram operator $\mathbf{J}^\top\mathbf{J}$ (Eq. 8), whereas maximizing a single class score yields only one row of $\mathbf{J}^\top$; (ii) it applies to any intermediate layer $r(\cdot)$, not only the output; and (iii) it requires no class labels or prototypes.

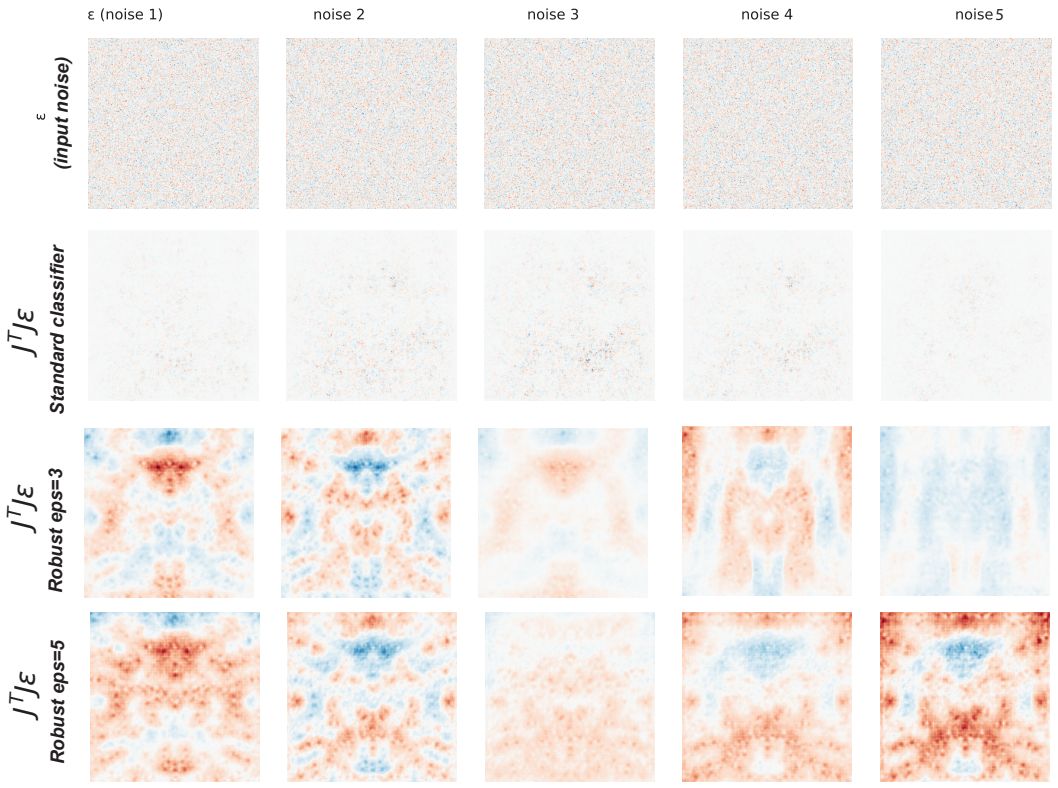

Figure S3: Visual residuals showing the action of $\mathbf{J}^\top\mathbf{J}$ on random noise for ImageNet ResNet50 models with varying robustness. **Top row:** Five different input noise samples (ε). **Subsequent rows:** Visual residuals ($\mathbf{J}^\top\mathbf{J}\varepsilon$) for a standard classifier ($\varepsilon = 0.00$) and robust classifiers with increasing adversarial training strength ($\varepsilon = 3.00, 5.00$). All visualizations use symmetric normalization (red-blue diverging colormap) with per-model scaling.

D FID Score for evaluation of synthesized images.

While Fréchet Inception Distance (FID) (Heusel et al., 2017) has become the standard metric for evaluating generative model quality, its application to MNIST presents fundamental limitations. The Inception network used in standard FID computation is trained on ImageNet, which contains natural images with fundamentally different statistical properties than handwritten digits. This domain mismatch makes Inception-based FID scores unreliable for MNIST evaluation, as the feature representations may not capture the relevant structure for digit classification. Additionally, for ImageNet-scale generation, computing FID scores would require generating thousands of samples to obtain statistically meaningful comparisons, which is computationally prohibitive for our large-scale trajectory analysis.

To address these limitations, we introduce a FID-like score that uses domain-appropriate feature extractors. For MNIST, we employ TinyResNet models trained on MNIST as feature extractors, computing the Fréchet distance between feature distributions of real test images and the probing procedure-generated samples. This approach provides a principled evaluation metric that aligns with the domain of interest while maintaining the statistical rigor of the original FID framework.

Our FID-like score computation follows the standard FID procedure but replaces Inception with TinyResNet feature extractors. Specifically, we:

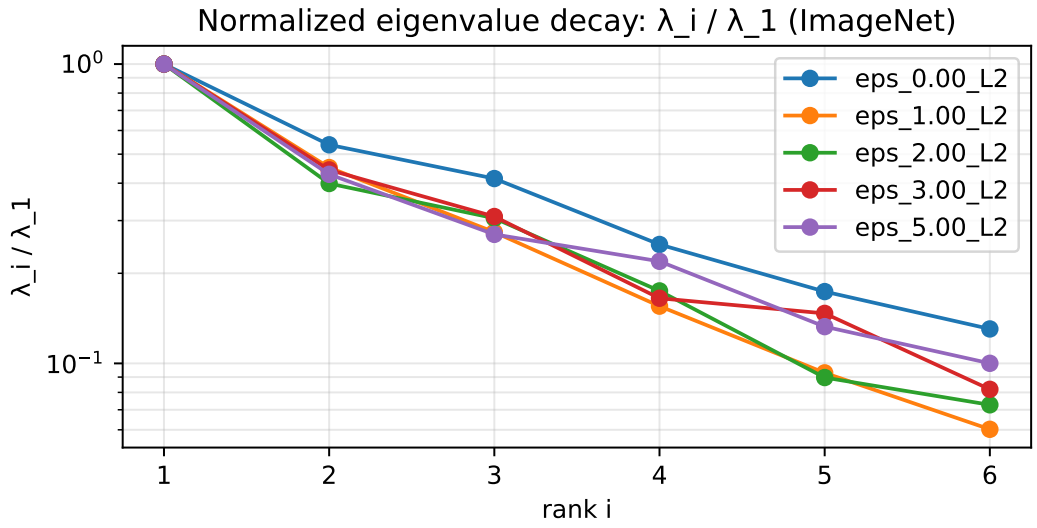

Figure S4: Normalized eigenvalue decay. Top-k normalized eigenvalues λ_i/λ_1 of $\mathbf{J}^\top\mathbf{J}$ for standard and robust ImageNet classifiers. (ϵ =0 shows a standard resnet50, all trained on ImageNet)

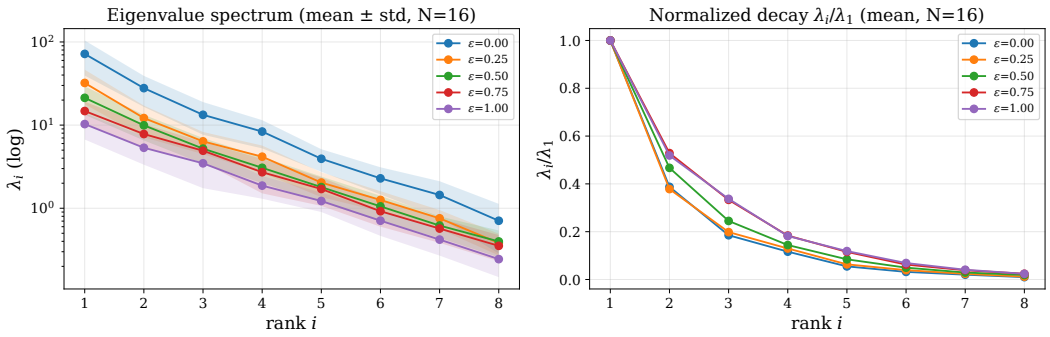

Figure S5: Eigenvalue spectrum of $\mathbf{J}^\top\mathbf{J}$ across adversarial training budgets (MNIST, $N = 16$ images). **Left:** Top-8 eigenvalues of $\mathbf{J}^\top\mathbf{J}$ (mean $\pm$ std over 16 test images) for models trained with increasing adversarial perturbation budgets $\varepsilon \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$. The entire spectrum shifts down monotonically with ε; λ_1 drops 7$\times$ from 71.8 to 10.3. **Right:** Eigenvalues normalized by λ_1 show similar decay shape across all ε, indicating uniform spectral compression rather than increased dominance of individual directions. Computed via subspace iteration with implicit JVP/VJP ($k = 8$, 30 iterations).

Algorithm 1 Gram operator probing procedure. Gradient normalization and the ℓ_2 projection are practical choices that bound step size and trajectory length; they are not part of the analysis in Section 4.

Input: Image $\mathbf{x}_{\text{input}}$, model f, target layer ℓ, constraint ϵ, step size α, noise ratio σ, diffusion noise ratio β, iterations T
Output: Refined images $\{\mathbf{x}_t\}_{t=0}^{T}$
// Step 1: Feedforward pass
$\mathbf{x}_0 \leftarrow \text{normalize}(\mathbf{x}_{\text{input}})$
$f_\ell \leftarrow \text{extract_layers}(f, \ell)$ {Extract model up to layer ℓ}
$\mathbf{x}_{\text{noisy}} \leftarrow \mathbf{x}_0 + \sigma \cdot \mathcal{N}(0, I)$
$r_{\text{anti-target}} \leftarrow f_\ell(\mathbf{x}_{\text{noisy}})$ {Generate noisy reference representation}
for $t = 1$ to T **do**
  // Step 2: Feedback error propagation
  $h_t \leftarrow f_\ell(\mathbf{x}_{t-1})$ {Forward pass through target layers}
  $L_t \leftarrow \|h_t - r_{\text{anti-target}}\|^2$ {MSE loss in representation space}
  $g_t \leftarrow \nabla_{\mathbf{x}_{t-1}} L_t$ {Gradient via feedback pathways}
  // Step 3: Constrained activation update
  $\tilde{g}_t \leftarrow \alpha \cdot g_t/(\|g_t\| + 10^{-10})$ {Normalize gradient}
  $\eta_t \leftarrow \beta \cdot \mathcal{N}(0, I)$ {Add stochastic noise}
  $\mathbf{x}'_t \leftarrow \mathbf{x}_{t-1} + \tilde{g}_t + \eta_t$ {Move away from representation of noisy input}
  $\mathbf{x}_t \leftarrow \text{project}(\mathbf{x}'_t, \mathbf{x}_0, \epsilon)$ {Enforce $\|\mathbf{x}_t - \mathbf{x}_0\|_2 \leq \epsilon$}

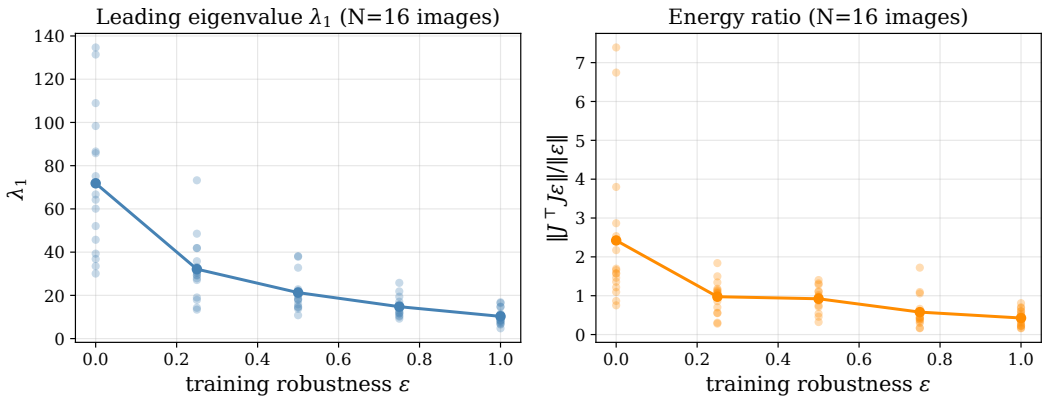

Figure S6: Spectral suppression scales with adversarial training budget (MNIST, $N = 16$ images). Per-image scatter (faded dots) and mean (solid line) for two spectral metrics across $\varepsilon \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$. **Left:** Leading eigenvalue λ_1 of $\mathbf{J}^\top\mathbf{J}$ drops monotonically from 71.8 to 10.3 (7× reduction). **Right:** Energy ratio $\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|/\|\boldsymbol{\epsilon}\|$ drops from 2.07 to 0.39, crossing below 1.0 at $\varepsilon = 0.25$, meaning $\mathbf{J}^\top\mathbf{J}$ actively suppresses rather than amplifies random noise.

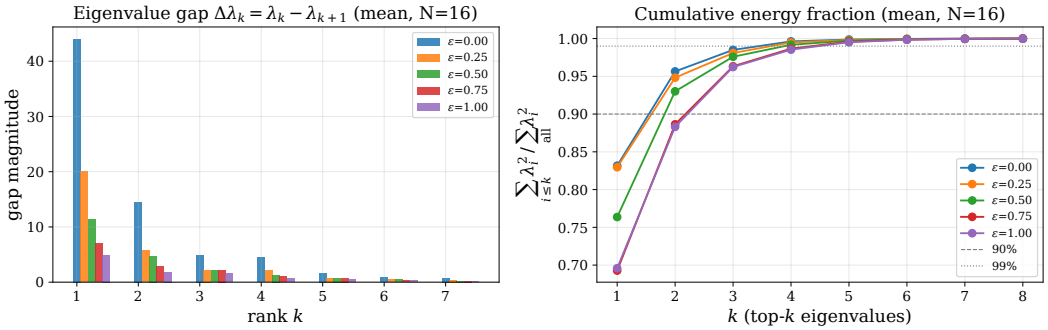

Figure S7: Principled k selection via spectral gap and cumulative energy (MNIST, $N = 16$ images). **Left:** Eigenvalue gap $\Delta\lambda_k = \lambda_k - \lambda_{k+1}$ averaged over 16 test images. **Right:** Cumulative energy fraction $\sum_{i\le k}\lambda_i^2/\sum_i\lambda_i^2$. For all models, $k = 2$ captures ∼95% of spectral energy and $k = 4$ reaches 99%. The 90% threshold (dashed) is crossed at $k = 2$–3 for all ε.

1. Extract penultimate layer features from both real MNIST test images and Probing procedure-generated samples using a reference TinyResNet model
2. Compute the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the feature distributions for both sets
3. Calculate the Fréchet distance between the two multivariate Gaussian distributions:

$$\text{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|^2 + \text{tr}(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}), \tag{10}$$

where subscripts r and g denote real and generated distributions, respectively.

The generated patterns used for FID-like score computation were obtained through systematic parameter sweeps across multiple dimensions. For each robust TinyResNet model (trained with different adversarial perturbation budgets $\varepsilon_{\text{train}} \in \{0.25, 1.00\}$), we generated samples by:

- **Input noise seed sweep:** 100 different input noise seeds (ranging from 1000-1100) to vary the initial starting conditions
- **Inference seed sweep:** 10 different inference seeds (ranging from 2000-2009) to vary the stochastic components of Probing procedure
- **Probing procedure parameter configuration:** Fixed Probing procedure parameters (step size=0.5, diffusion noise ratio=0.005, initial inference noise ratio=0.5, $\epsilon_{\text{infer}} = 2.0$, iterations=50) consistent with the main experimental setup

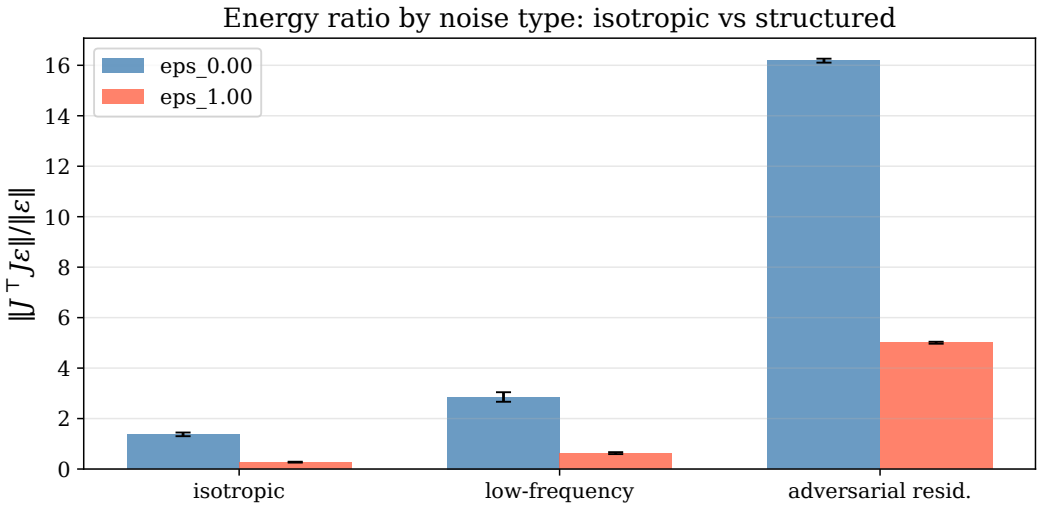

Figure S8: Energy ratio by noise type: isotropic vs structured (MNIST). Energy ratio $\|\mathbf{J}^\top \mathbf{J}\boldsymbol{\epsilon}\|/\|\boldsymbol{\epsilon}\|$ for standard ($\varepsilon = 0.00$) and robust ($\varepsilon = 1.00$) TinyResNet models across three noise types. Isotropic: Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Low-frequency: low-pass filtered Gaussian (cutoff=4). Adversarial residual: PGD gradient direction with random start (5 steps, $\sigma = 0.25$). 64 draws per condition.

Table 6: Probing procedure parameter settings for main test figures (ImageNet ResNet-50, L_2 adversarially trained).

Figure	Model	Drift Noise Ratio	Diffusion Noise Ratio	n_{itr}	ϵ_{infer}	Step Size (α)
2	ResNet-50, L_2, $\epsilon = 3$	0.2	0.01	1001	40	1
2	ResNet-50, L_2, $\epsilon = 5$	0.5	0.03	1001	40	1
3	vgg16-bn, L_2, $\epsilon = 3$	0.2	0.01	1001	40	1
3	resnext50-4dx32, L_2, $\epsilon = 3$	0.2	0.01	1001	40	1

This systematic sweep produced 1000 trajectories per model configuration, from which we extracted the final iteration images (iteration 49) for FID-like score computation. The diversity of initialization conditions ensures that our evaluation captures the full range of generative patterns accessible through the probing procedure rather than cherry-picking specific examples.

We evaluate FID-like scores using three different reference models to understand how the choice of feature extractor affects the assessment of generative quality:

- **Clean model ($\varepsilon = 0.00$):** Standard TinyResNet trained without adversarial perturbations
- **Moderately robust model ($\varepsilon = 0.25$):** TinyResNet trained with ℓ_2 adversarial training at $\varepsilon = 0.25$
- **Highly robust model ($\varepsilon = 1.00$):** TinyResNet trained with ℓ_2 adversarial training at $\varepsilon = 1.00$

This multi-reference evaluation reveals how the feature space geometry varies with robustness training and provides insights into which models' priors are best captured by the probing procedure.

Table 7 presents FID-like scores computed across different combinations of reference models and generated image sources. Several key observations emerge:

The choice of reference model significantly affects FID-like scores, reflecting fundamental differences in feature space geometry. When using the clean model ($\varepsilon = 0.00$) as reference, images generated from the highly robust model ($\varepsilon = 1.00$) achieve lower FID scores (98.73) than those from the moderately robust model ($\varepsilon = 0.25$, score 119.98). This suggests that highly robust models produce patterns that are more aligned with the clean model's feature space, despite being trained with stronger adversarial constraints. Interestingly, when using the same model as both generator and reference, the results are not uniformly favorable. For the $\varepsilon = 0.25$ reference model, its own generated images

Table 7: FID-like scores for Probing procedure-generated MNIST samples using different TinyResNet reference models. Lower scores indicate greater similarity between generated and real image feature distributions. Scores computed using 10,000 real MNIST test images and 995-1000 generated samples per configuration.

Reference Model	**Generated from** $\varepsilon = 0.25$	**Generated from** $\varepsilon = 1.00$
$\varepsilon = 0.00$ (clean)	119.98	98.73
$\varepsilon = 0.25$	101.96	97.86
$\varepsilon = 1.00$	88.69	133.11

achieve a score of 101.96, while $\varepsilon = 1.00$ generated images score slightly better at 97.86. Conversely, when using the $\varepsilon = 1.00$ model as reference, $\varepsilon = 0.25$ generated images achieve the best score (88.69), while its own generated images score worse (133.11). This asymmetry suggests that the feature representations learned by different robustness levels are not simply nested subspaces but rather exhibit complex geometric relationships.

The FID-like scores demonstrate that the probing procedure successfully generates patterns that occupy meaningful regions of the feature space, with scores in the range of 88-134. While these values are higher than typical FID scores for high-quality generative models (which often achieve scores below 10), this is expected given that: (1) Probing procedure operates without explicit generative training, (2) the feature space dimensionality (10 dimensions for TinyResNet penultimate layer) is much smaller than Inception's 2048-dimensional space, and (3) the generated patterns represent prototypical attractors rather than direct reproductions of training data. The consistent convergence to stable patterns (as evidenced by the trajectory analysis in Figure S11) combined with these FID-like scores provides quantitative evidence that robust classifiers encode meaningful generative priors accessible through the probing procedure.

We note that direct comparison to standard Inception-based FID scores is not meaningful due to the domain mismatch discussed above. However, our FID-like scores serve as a principled evaluation metric that: (1) uses domain-appropriate feature extractors, (2) maintains the statistical rigor of the Fréchet distance framework, and (3) enables quantitative assessment of generative quality for MNIST without requiring thousands of ImageNet-scale generations. This approach provides a practical solution for evaluating implicit generative capabilities in robust classifiers while remaining computationally tractable for systematic parameter sweeps.

E Generation vs. Interpolation: Nearest-Neighbor Analysis

To examine whether probing procedure-generated samples reflect instance-level recall of training images, we conducted a nearest-neighbor (NN) analysis comparing each generated image to the full CIFAR-10 training set in both pixel space and logit space. In pixel space, we measured Euclidean distance between the flattened RGB vectors of the generated sample and each training image. To assess similarity in the model's learned representation, we also computed nearest neighbors in logit space, using the penultimate logits of a standard (non-robust) CIFAR-10 ResNet-50; these distances reflect semantic similarity as encoded by the classifier. The closest training images exhibit large pixel-space distances and are visually distinct, differing in background, pose, and object boundaries. Across all generated samples we evaluated (examples shown in Figure S9), we did not observe direct reproduction or interpolation of individual training examples.

F Analysis of Generated Digit Distribution and Per-Class Accuracy

To understand how the implicit generative priors encoded by robust classifiers relate to their discriminative performance, we conducted a comprehensive analysis comparing the distribution of digits generated by the probing procedure with the per-class classification accuracy of the same models on the test set. This analysis provides insights into whether models generate digits in proportion to their classification performance, revealing potential biases in the learned representations.



Figure S9: Nearest-neighbor analyses for two instances of the patterns synthesized by the probing procedure on a robust CIFAR-10 model ($\epsilon = 1$). Top rows: nearest neighbors in pixel space (Euclidean distance). Bottom rows: nearest neighbors in logit space (Euclidean distance in the model's penultimate logits).

We evaluated multiple ResNetTiny models trained on MNIST with different adversarial perturbation budgets ($\epsilon \in \{0.25, 0.50, 0.75, 1.00\}$) and at various training epochs (excluding epoch 5 to focus on more mature models). For each model checkpoint, we collected all generated images produced by the probing procedure across multiple inference runs with varying input noise seeds and inference parameters. We extracted the final predicted digit class for each generated image (the class predicted by the model at the final iteration of Probing procedure) and computed the distribution of generated digits across all 10 classes. This distribution represents what the model's implicit generative prior considers to be the most natural or prototypical digit representations. For the same model checkpoints, we evaluated classification performance on the MNIST test set, computing both clean accuracy and adversarial robust accuracy for each digit class (0-9). This provides a measure of how well the model discriminates between different digit classes, allowing us to assess whether generation frequency correlates with classification performance.

Our analysis reveals that the distribution of generated digits is highly non-uniform across all models, with certain digits (notably digits 2, 4, and 8) being generated significantly more frequently than others. This non-uniformity persists across different training epochs and perturbation budgets, suggesting that robust training induces consistent biases in the learned generative priors. Interestingly, the frequency with which digits are generated does not directly correlate with per-class classification accuracy. While models achieve relatively uniform clean accuracy across digit classes (typically 96-99%), the generated digit distribution shows substantial variation, with some digits appearing 5-10 times more frequently than others. This suggests that the implicit generative prior encoded by $\mathbf{J}^\top\mathbf{J}$ reflects structural properties of the learned representations that extend beyond simple classification performance.

The generated digit distribution remains relatively stable across different training epochs for models with the same perturbation budget, indicating that the implicit generative structure emerges early in training and remains consistent. However, we observe subtle variations in the distribution as training progresses, potentially reflecting refinement of the learned priors. Models trained with different adversarial perturbation budgets (ϵ) exhibit distinct generation patterns, with higher ϵ values generally producing more diverse digit distributions. This observation aligns with our theoretical framework: stronger adversarial training constraints induce more structured $\mathbf{J}^\top\mathbf{J}$ operators, which may access richer generative priors.

These findings suggest that the implicit generative structure in robust classifiers captures statistical regularities that are not directly reflected in classification metrics. The non-uniform generation patterns may reflect the intrinsic difficulty of representing certain digit classes in the learned feature space, or they may indicate that the implicit denoiser $\mathbf{J}^\top\mathbf{J}$ has learned to prioritize certain prototypical digit representations that serve as attractors in the generative space. The stability of generation patterns across training epochs indicates that the implicit generative structure is a fundamental property of the learned representations, rather than an artifact of specific training configurations. This stability supports our theoretical framework, which posits that adversarial training fundamentally alters the spectral structure of the Jacobian operator. The variation in generation patterns with different perturbation budgets provides empirical support for the connection between robustness constraints and generative capabilities, suggesting that stronger adversarial training may enable access to more diverse and structured implicit priors. These findings contribute to our understanding of how discriminative robustness relates to generative modeling, revealing that robust classifiers encode statistical knowledge that extends beyond their primary classification function. The systematic analysis of generation patterns provides a new lens through which to examine the learned representations of robust models, potentially informing future work on interpretability and explainability.

F.1 Probing procedure trajectories

Probing procedure trajectories seem to converge to stable patterns. To understand Probing procedure's behavior systematically, we conduct comprehensive experiments across multiple robust classifiers trained with different adversarial perturbation budgets (ϵ), training epochs, and initialization seeds. For each model, we run the probing procedure starting from 100 different noise seeds with 10 different inference seeds, generating 1000 trajectories per model. The results reveal consistent convergence properties (Figure S11): despite diverse starting conditions, trajectories converge to stable class predictions with high confidence. Multiple runs frequently arrive at the same predicted class, suggesting that the probing procedure reliably accesses the most prominent modes of the implicit generative model. The generated patterns are distinct from real training digits, appearing more like internal prototypical templates that serve as attractors for Probing procedure trajectories. Rather than reproducing memorized training examples, the implicit denoiser $\mathbf{J}^\top\mathbf{J}$ reveals canonical digit representations that guide the convergence dynamics, although further work is needed to characterize the formation of these stable patterns.

Generated patterns reveal distinct prototypical features. To characterize the generated patterns against an equivalent number of samples from training data, we employ t-SNE visualization (van der Maaten & Hinton, 2008) (Figure S12) to investigate their similarity to real data and examine their distinct characteristics. The generated patterns occupy regions distinct from the real training data. This is consistent with the probing procedure not reproducing memorized examples, and it also shows that the generated patterns lie away from the training data in this feature space. However, when samples are colored by class (predicted class for generated samples), we observe distinct clusters

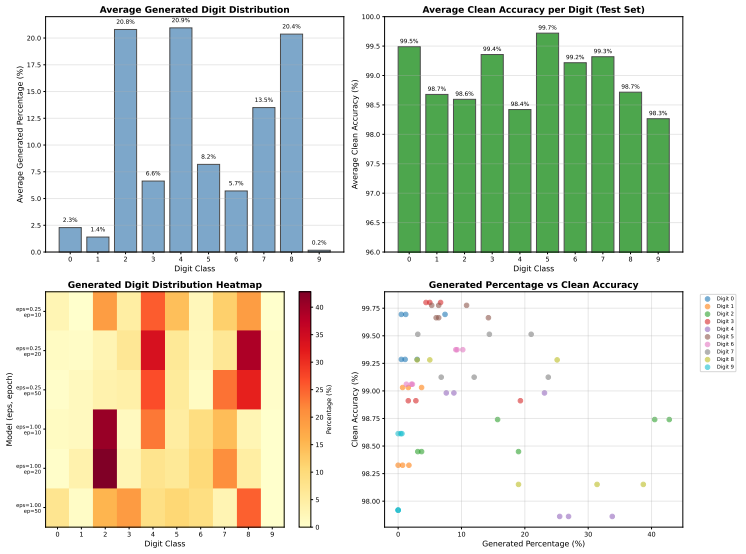

Figure S10: Comparison of generated digit distribution and per-class classification accuracy for ResNetTiny models trained on MNIST with different adversarial perturbation budgets ($\epsilon \in \{0.25, 0.50, 0.75, 1.00\}$) and training epochs. (Top left) Average percentage of generated images classified as each digit class (0-9) across all evaluated models. (Top right) Average clean accuracy per digit class on the test set, with y-axis range 96-100%. (Bottom left) Heatmap of generated digit distribution for each model (rows) and digit class (columns), with color intensity representing the percentage of generated images. (Bottom right) Scatter plot of generated percentage versus clean accuracy for each digit class, with different colors indicating different digits.

of similarly recognized patterns that remain stable across different t-SNE random initialization seeds . This consistent clustering suggests that these generated patterns may represent the stable discriminative features the robust model has learned during training - canonical representations that serve as attractors in the implicit generative space rather than copies of specific training instances. The separation between generated and real data, combined with the coherent and reproducible class-based clustering of generated samples, provides evidence that robust classifiers encode prototypical feature representations that extend beyond the original training distribution.

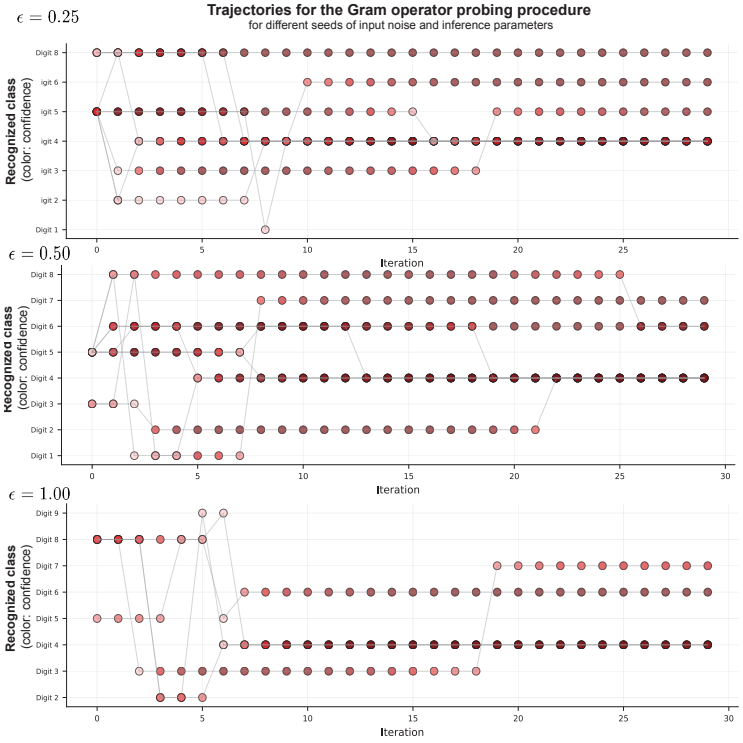

Figure S11: **Probing procedure exhibits convergent class prediction trajectories across different initializations.** Classification trajectories for Probing procedure applied to a robust MNIST classifier, starting from different random noise seeds and inference parameters (L_2 robust models with top to bottom: $\epsilon = 0.25, 0.5, 1$). Each line represents the predicted class and confidence evolution (depicted in color) during generation. (examples shown: final outputs for class 6 with confidence 1.0 and 0.57).

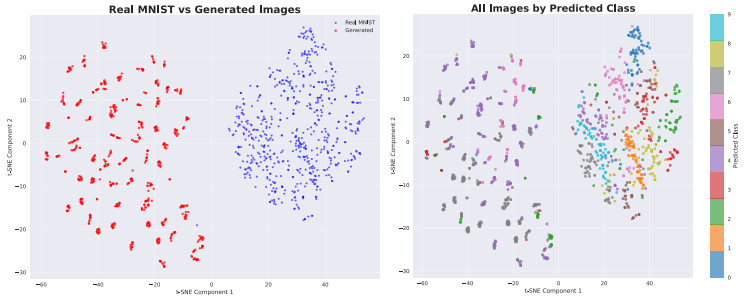

Figure S12: **Probing procedure-generated images occupy distinct regions in representation space from real MNIST data.** t-SNE visualization of penultimate layer representations comparing real MNIST training images (red) with probing procedure-generated samples (blue) from robust classifiers across different initialization seeds for input noise and inference seeds (same probing procedure parameters).