IMPLICIT DENOISER STRUCTURE IN ROBUST CLASSIFIERS EXPLAINS GENERATIVE CAPABILITIES

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 established results that robust training drives Jacobians toward low-rank solutions, we demonstrate that the Gram operator $\mathbf{J}^\top\mathbf{J}$ functions as an implicit denoiser, selectively preserving signal along discriminative subspaces while suppressing noise in orthogonal directions. This insight leads to Prior-Guided Drift Diffusion (PGDD), a simple algorithm that leverages this structure for generation through inference objectives rather than explicit Jacobian computation. PGDD requires no generative training or architectural modifications, yet produces class-consistent samples across different datasets and architectures. We extend our approach to standard networks via sPGDD, demonstrating that implicit generative structure exists beyond adversarially trained models. Our results establish a connection between discriminative robustness and generative modeling, showing that robust classifiers encode statistical priors that enable structured pattern generation without explicit generative objectives.

1 INTRODUCTION

Adversarial training has emerged as a critical defense mechanism for ensuring the safety and reliability of neural networks deployed in high-stakes applications, where robustness to input perturbations is essential for preventing adversarial attacks and maintaining system integrity (Madry et al., 2018; Wong et al., 2020). Originally developed to address security vulnerabilities in machine learning systems (Goodfellow et al., 2015; Carlini & Wagner, 2017), adversarial training has revealed unexpected emergent properties: robust models can function as implicit generative models and produce structured images when appropriately probed (Santurkar et al., 2019; Engstrom et al., 2019). This dual discriminative–generative behavior suggests that the mechanisms underlying adversarial robustness may be more fundamental than previously recognized, yet the theoretical foundations of these emergent capabilities remain largely unexplored.

Recent theoretical advances have begun to illuminate the mathematical structure underlying adversarial robustness. Studies have demonstrated that the spectral properties of neural networks such as input-output Jacobians are directly linked to generalization and robustness (Oymak et al., 2019; Wu & Li, 2024). This 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). However, despite these insights into the *discriminative* implications of spectral structure, the potential *generative* consequences of low-rank Jacobians remain largely unexplored.

In this work, we bridge this gap by drawing inspiration from the success of denoising diffusion probabilistic models (DDPMs), which achieve remarkable generative performance through learned denoising operations (Ho et al., 2020; Song & Ermon, 2019). The core insight from diffusion models is that networks trained to remove noise implicitly learn the score function of the data distribution, enabling iterative generation through gradient-based sampling (Song et al., 2021). This connection between denoising and generation motivates our central hypothesis: the low-rank structure induced by adversarial training makes the Gram operator $\mathbf{J}^\top\mathbf{J}$ function as an implicit denoiser, selectively

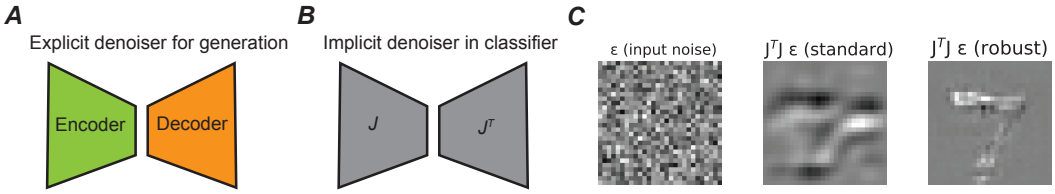

Figure 1: **Adversarial robustness creates implicit denoisers through $\mathbf{J}^\top\mathbf{J}$ structure. (A)** Explicit denoisers use separate encoder-decoder architectures for noise removal and generation (Vincent et al., 2008; Vincent, 2011) . **(B)** Our hypothesis: robust classifiers develop mathematically equivalent structure where the Jacobian $\mathbf{J}$ and its transpose $\mathbf{J}^\top$ naturally form an implicit denoising operator $\mathbf{J}^\top\mathbf{J}$. **(C)** Empirical validation on MNIST: when $\mathbf{J}^\top\mathbf{J}$ is applied to input noise $\boldsymbol{\epsilon}$, standard classifiers fail to reject the random structure, while robust classifiers extract digit-like patterns, demonstrating genuine denoising capability. Images are individually normalized to reveal structure.

preserving discriminative directions while suppressing noise along orthogonal subspaces. Just as DDPMs leverage explicit denoising networks for generation, we demonstrate that robust classifiers contain implicit denoising structure that can be exploited for the same purpose. We introduce *Prior-Guided Drift Diffusion (PGDD)*, an algorithm that harnesses this hidden structure through inference objectives rather than explicit Jacobian computation, enabling practical generation from robust classifiers. We further develop *sPGDD*, a variant that extends our approach to standard networks through gradient smoothing techniques.

This work makes several key contributions to understanding the connection between adversarial robustness and generative modeling:

- This work establishs that the Gram operator $\mathbf{J}^\top\mathbf{J}$ in adversarially robust classifiers functions as an implicit denoiser, connecting prior findings on low-rank Jacobian structure to generative capabilities. This provides the first principled explanation for why robust classifiers exhibit emergent generative properties.
- It also demonstrate through spectral analysis, energy ratio measurements, and visual residuals that robust classifiers suppress noise while amplifying class-consistent structure.
- We introduce Prior-Guided Drift Diffusion (PGDD), a practical algorithm that leverages implicit $\mathbf{J}^\top\mathbf{J}$ structure for generation through inference objectives rather than explicit Jacobian computation. PGDD requires no architectural modifications or generative training.
- Finally, we develop sPGDD (smooth PGDD), which enables generative inference in standard classifiers through gradient smoothing techniques, demonstrating that implicit generative structure exists beyond robust networks.

2 Implicit Denoiser in Robust Classifiers

2.1 Generative Power of Adversarially Robust Classifiers

The observation that adversarially robust classifiers exhibit unexpected generative capabilities has garnered increasing attention across multiple domains. Recent empirical work has demonstrated that robust models can synthesize structured images (Santurkar et al., 2019), produce perceptually aligned gradients Kaur et al. (2019), and exhibit improved correspondence between their internal representations and human-perceivable features (Engstrom et al., 2019). Intriguingly, these generative properties emerge without explicit generative training objectives, suggesting an intrinsic connection between discriminative robustness and generative modeling capacity.

Previous work has explored connections between classifiers and generative models, but these approaches typically require training classifiers with explicit generative objectives. Joint Energy-based Models (JEMs) train networks to simultaneously perform classification and generation by optimizing both discriminative and generative losses (Grathwohl et al., 2020). Similarly, gradient alignment methods improve model interpretability by training the implicit density model to align with ground

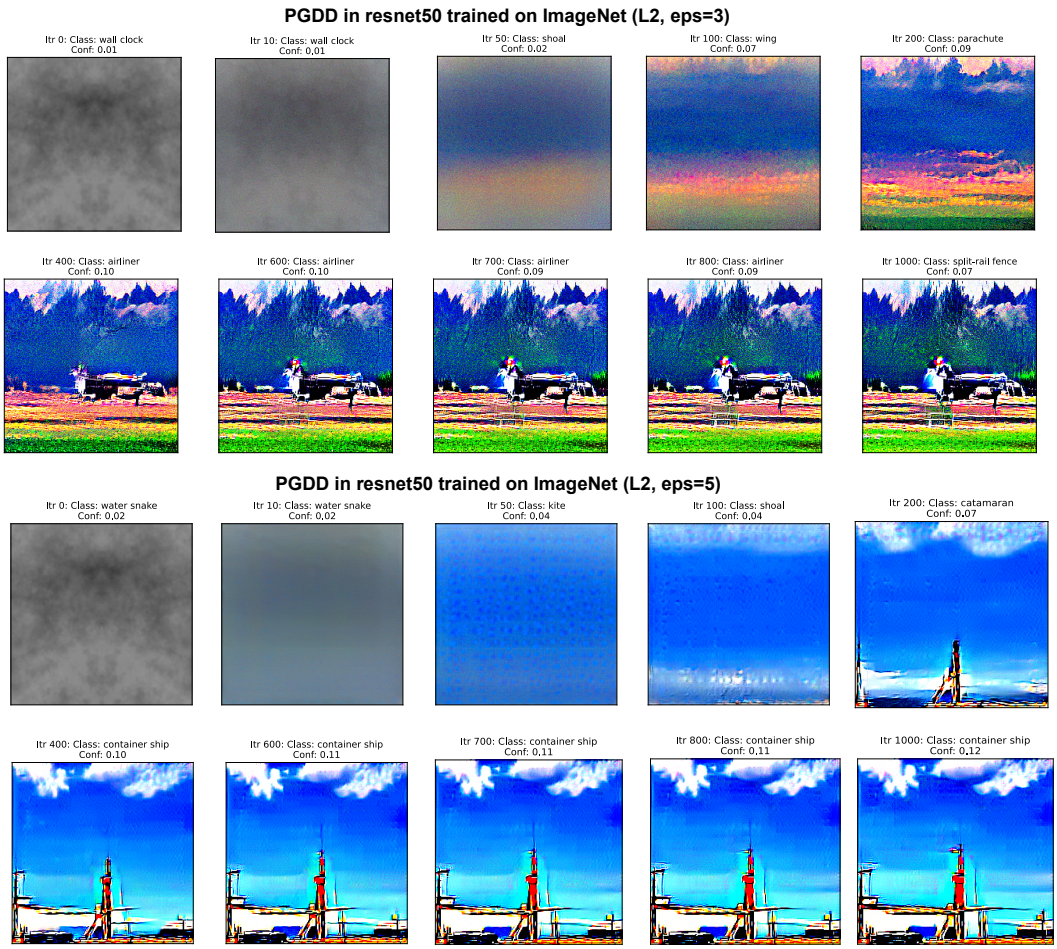

Figure 2: **Prior-Guided Drift Diffusion (PGDD) generates coherent images from noise using robust classifiers.** PGDD applied to a robust ResNet-50 (ImageNet, ℓ_2 adversarially trained with $\epsilon = 3, 5$) demonstrates progressive refinement from random noise (same for both shown trajectories) to semantically coherent images. Starting from noise (Itr 0), the algorithm iteratively moves away from a noisy representation of the original input (see algorithm in 1). No explicit generative training was used, only PGDD on pretrained adversarially robust classifiers (Supplementary C.3).

truth distributions (Singla et al., 2021). However, these methods fundamentally alter the training process to achieve generative capabilities, whereas robust classifiers exhibit these properties as emergent byproducts of adversarial training alone.

Despite these compelling empirical demonstrations, the theoretical foundations underlying the generative capacity of robust classifiers remain largely unexplored. While extensive work has characterized the links between spectral properties and robustness (Hoffman et al., 2019; Jakubovitz & Giryes, 2018; Oymak et al., 2019; Wu & Li, 2024), no principled framework has emerged to explain how these spectral characteristics translate to generative functionality. Several lines of research have provided crucial building blocks for our theoretical development. Studies on Jacobian regularization have established that controlling gradient norms enhances robustness to adversarial perturbations (Hoffman et al., 2019; Jakubovitz & Giryes, 2018). Complementary work has demonstrated that adversarial training fundamentally alters the spectral properties of neural networks (Du et al., 2019; Sinha et al., 2018). However, these findings have been studied in isolation, without connecting spectral structure to generative inference capabilities.

The success of denoising diffusion probabilistic models (DDPMs) provides additional theoretical context. DDPMs achieve remarkable generative performance by learning to reverse a noise process,

with the core insight that score functions can guide iterative denoising Ho et al. (2020); Song & Ermon (2019). This raises a natural question: could robust classifiers similarly encode implicit score functions that enable generative inference?

2.2 Theoretical Framework: Signal-Noise Decomposition in $\mathbf{J}^\top\mathbf{J}$

Building on established results that adversarial training drives neural networks toward low-rank Jacobian solutions, we formalize how this spectral structure enables implicit denoising. Since $\text{rank}(\mathbf{J}^\top\mathbf{J}) \leq \text{rank}(\mathbf{J})$, the Gram operator inherits the low-rank structure induced by robust training.

Let $\mathbf{J}^\top\mathbf{J} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ with eigenvalues $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_P)$ and orthonormal eigenvectors $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_P]$. For any perturbation $\boldsymbol{\epsilon} = \sum_i c_i \mathbf{v}_i$,

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

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

For random $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_P)$, the expected energy becomes:

$$\mathbb{E}\left[\|\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}\|_2^2\right] = \sigma^2 \text{tr}\left((\mathbf{J}^\top\mathbf{J})^2\right) = \sigma^2 \sum_{i=1}^{P} \lambda_i^2, \tag{3}$$

which is small when only few λ_i are appreciable.

The denoising mechanism emerges naturally: If we denote the small set of "signal" directions by $\mathcal{S} = \{i : \lambda_i \text{ large}\}$ with $|\mathcal{S}| = k \ll P$ and decompose $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_\parallel + \boldsymbol{\epsilon}_\perp$ along $\mathcal{S}$ and $\mathcal{S}^c$, then

$$\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\parallel \approx \boldsymbol{\epsilon}_\parallel, \tag{4}$$

$$\mathbf{J}^\top\mathbf{J}\boldsymbol{\epsilon}_\perp \approx \mathbf{0}. \tag{5}$$

This formalizes our central hypothesis: $\mathbf{J}^\top\mathbf{J}$ suppresses random noise components while preserving structured components aligned with the discriminative subspace $\mathcal{S}$. Crucially, this denoising capability requires no additional training beyond the original robustness objective. Just as explicit denoising autoencoders use encoder-decoder architectures for generation (Figure 1**A**), robust classifiers naturally develop mathematically equivalent structure where $\mathbf{J}$ and $\mathbf{J}^\top$ form an implicit denoising operator $\mathbf{J}^\top\mathbf{J}$ (Figure 1**B**). This framework makes several testable predictions that we validate empirically: (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 exhibit stronger spectral concentration with higher λ_1/λ_2 ratios and steeper eigenvalue decay.

3 Prior–Guided Drift Diffusion (PGDD)

Having established that robust classifiers contain implicit denoising structure through $\mathbf{J}^\top\mathbf{J}$, we now address how to leverage this capability for generation. Previous methods for generating images using adversarially robust classifiers have employed procedures essentially equivalent to targeted adversarial attacks (Santurkar et al., 2019). These approaches optimize inputs to maximize specific class predictions, effectively using the classifier's gradients to guide generation. We wish to apply the denoising operator $\mathbf{J}^\top\mathbf{J}$ without forming it explicitly.

Our approach is inspired by adversarial purification methods: we first corrupt the input with noise, then attempt to move away from that noisy representation using gradients. Intuitively, this process should reveal the underlying structure that the network has learned to distinguish from corruption. We show that $\mathbf{J}^\top\mathbf{J}$ emerges naturally from this simple inference objective.

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

$$\mathcal{L}_{\text{PGDD}}(\mathbf{x}; \boldsymbol{\epsilon}) = \left\| r(\mathbf{x}) \; - \; \text{sg}[r(\mathbf{x} + \boldsymbol{\epsilon})] \right\|_2^2, \tag{6}$$

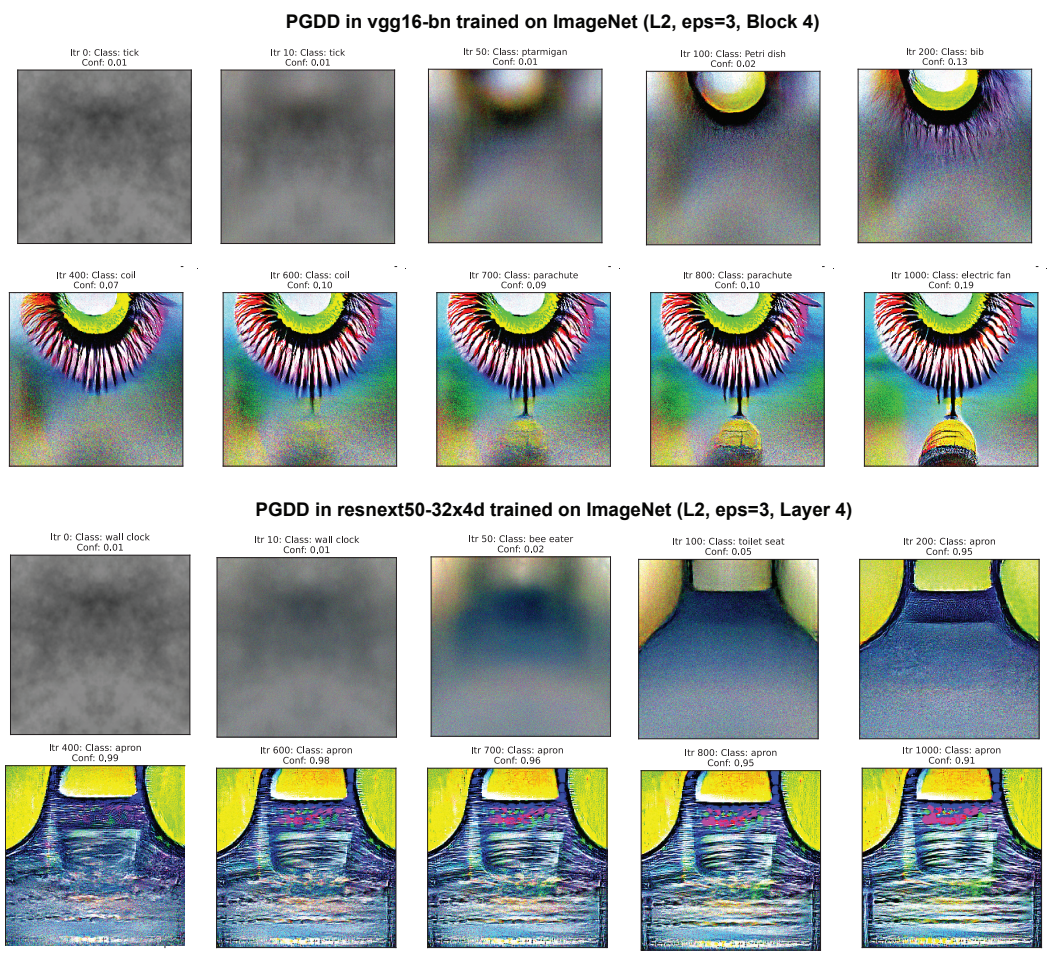

Figure 3: **PGDD across architectures and applied at different internal layers** PGDD applied to robust vgg16-bn and resnext50-32x4d (ImageNet, ℓ_2 adversarially trained with $\epsilon = 3, 5$) demonstrates progressive refinement from random noise (same for both shown trajectories) to semantically coherent images. Starting from noise (Itr 0), the algorithm iteratively moves away from a noisy representation of the original input (see algorithm in 1). No explicit generative training was used, only PGDD on pretrained adversarially robust classifiers (Supplementary C.3).

where $\mathrm{sg}[\cdot]$ stops gradients through its argument.

Let $\mathbf{J}_r(\mathbf{x}) = \nabla_{\mathbf{x}} r(\mathbf{x})$. Then

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

$$\approx 2\,\mathbf{J}_r(\mathbf{x})^\top \Big(r(\mathbf{x}) - \big(r(\mathbf{x}) + \mathbf{J}_r(\mathbf{x})\boldsymbol{\epsilon}\big)\Big) = -2\,\mathbf{J}_r(\mathbf{x})^\top \mathbf{J}_r(\mathbf{x})\,\boldsymbol{\epsilon}. \tag{8}$$

Since the goal is to move away from the noisy representation, we ascend on equation 6:

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

which applies the denoising step $-\mathbf{J}_r^\top \mathbf{J}_r\,\boldsymbol{\epsilon}$. Thus, $\mathbf{J}^\top\mathbf{J}$ emerges naturally from the simple objective of moving away from noisy representations.

3.1 Smoothed PGDD (sPGDD) for Standard Neural Networks

Standard networks do not exhibit the strongly structured $\mathbf{J}^\top\mathbf{J}$ operator induced by adversarial training. To adapt our approach, we introduce sPGDD (smooth PGDD), which reduces gradient variance

through smoothing. Our method draws inspiration from SmoothGrad (Smilkov et al., 2017), which averages gradients over multiple noise perturbations to produce more perceptually aligned saliency maps. However, while SmoothGrad has proven effective for interpretation tasks, it has never been applied to generation. We demonstrate that this smoothing principle can be leveraged to access implicit generative structure in standard networks.

Rather than relying on a single noisy gradient, we fix a reference representation $r(\mathbf{x} + \boldsymbol{\epsilon})$ at initialization and, at each iteration t, average gradients over multiple independent perturbations $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$:

$$\mathbf{g}_t = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{x}_t} \mathcal{L}_{\text{PGDD}}(\mathbf{x}_t; \boldsymbol{\epsilon}_i).$$

This procedure suppresses unstable noise-sensitive components and emphasizes more stable prior information within the network. Empirically, sPGDD produces smoother, more interpretable trajectories in non-robust networks, though with lower fidelity than in robust models.

4 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 PGDD across datasets and architectures, showing that implicit denoising structure enables coherent image generation from robust classifiers.

4.1 SPECTRAL ANALYSIS: VALIDATING THE IMPLICIT DENOISER HYPOTHESIS

Eigenvalue decay and energy ratios. Eigenvalue analysis confirms that robust models exhibit rapidly decaying Jacobian spectra, with tail eigenvalue suppression showing 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.

Table 1: Quantitative verification of implicit denoiser hypothesis

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×

Visual residuals reveal class priors. Applying $\mathbf{J}^\top \mathbf{J}$ to random noise produces dramatically different outcomes in robust versus standard classifiers (Figure 1C). 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 $\mathbf{J}^\top \mathbf{J}$ functions as both a denoiser and an amplifier of implicit statistical knowledge.

4.2 GENERATION WITH PGDD: INSIGHTS INTO CLASSIFIERS' PRIORS

PGDD on ImageNet-trained robust models. Starting from identical noise patterns and using identical inference parameters, PGDD applied 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) converges to semantically distinct but equally coherent patterns—parachute/landscape scenes versus container ship/seascape scenes (Figure 2). These

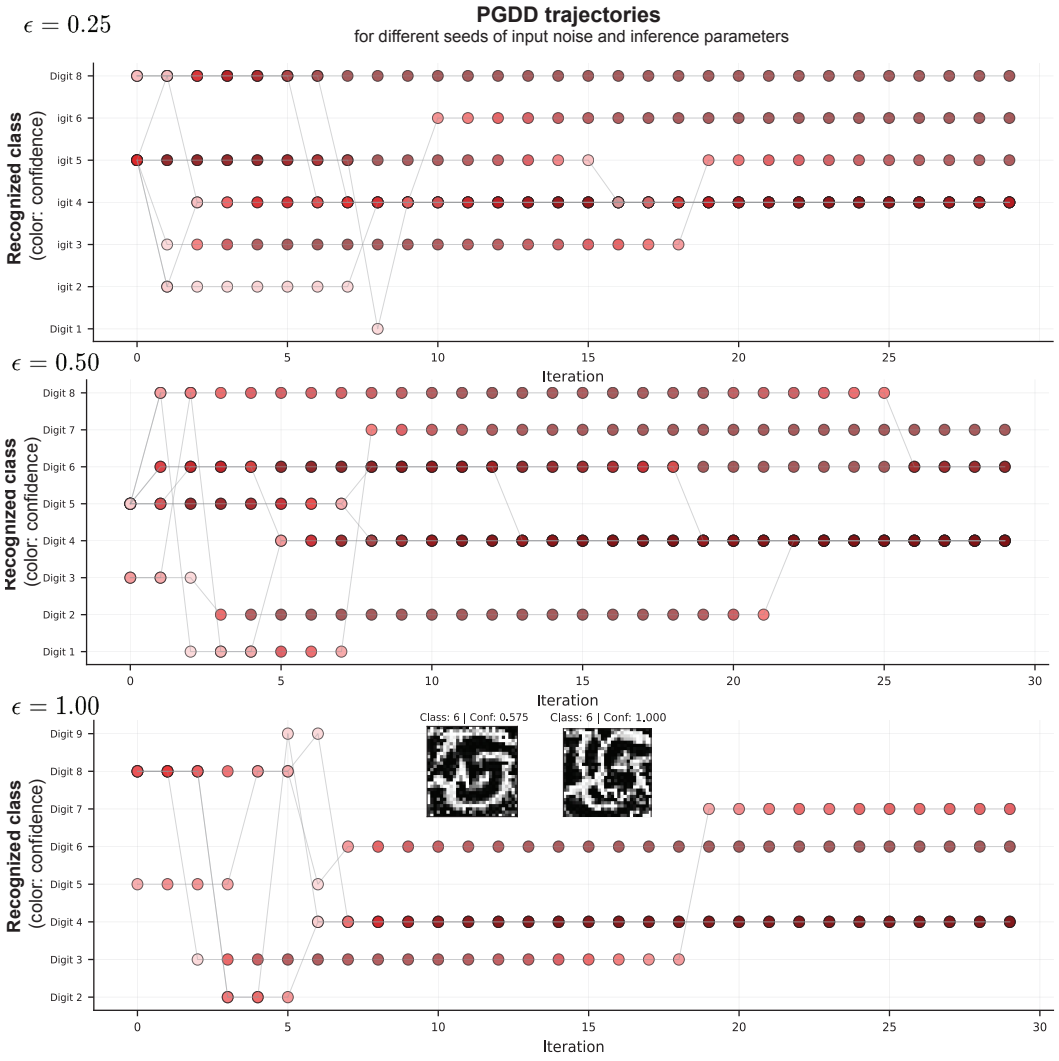

Figure 4: **PGDD exhibits convergent class prediction trajectories across different initializations.** Classification trajectories for PGDD applied to a robust MNIST classifier, starting from different random noise seeds and inference parameters (L_2 robust models with top to bottom: $\epsilon = 0.25, 0.5, 1$). 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).

serve as representative instances of PGDD trajectory convergence on ImageNet-trained priors, demonstrating that different robustness constraints create distinct semantic attractors. We further validated our approach by testing sPGDD on standard ResNet-50 and self-supervised ResNet-50 (MoCo), with trajectory examples provided in the supplementary (C.2.1).

PGDD across architectures and internal layers. PGDD can be applied on the internal layers as well, here we tested vgg16-bn (Simonyan & Zisserman, 2014) a smaller network compared to ResNet50 and ResNeXt-50-32x4d (Xie et al., 2017) is a wider network (Figure 3). Starting from noise, the algorithm progressively refines inputs to reveal class-consistent images across diverse categories in the case of imagenet, 1000 learned categories (Figure 3). The generative process exhibits a similar structure: early iterations establish global color gradients and spatial organization, while later iterations refine object-specific details. Importantly, predicted classes stabilize over iterations and confidence scores increase throughout generation, indicating that PGDD parameters can be tuned for convergence (see limitations).

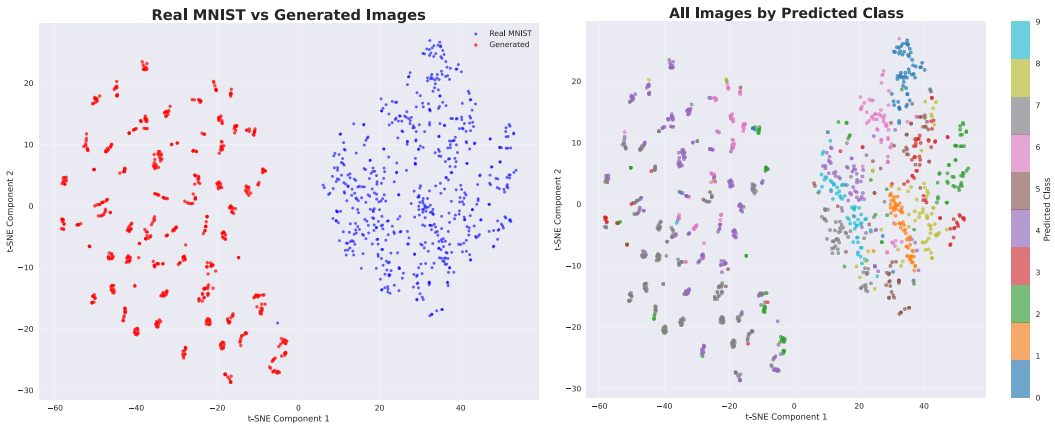

Figure 5: **PGDD-generated images occupy distinct regions in representation space from real MNIST data.** t-SNE visualization of penultimate layer representations comparing real MNIST training images (red) with PGDD-generated samples (blue) from robust classifiers across different initialization seeds for input noise and inference seeds (same PGDD parameters).

PGDD trajectory analysis on MNIST. To understand PGDD'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 PGDD starting from 100 different noise seeds with 10 different inference seeds, generating 1000 trajectories per model. The results reveal consistent convergence properties (Figure 4): despite diverse starting conditions, trajectories converge to stable class predictions with high confidence. Multiple runs frequently arrive at the same predicted class, suggesting that PGDD 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 PGDD 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 learned 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 5) to investigate their similarity to real data and examine their distinct characteristics. As expected, the generated patterns occupy regions very distinct from real training data, confirming that PGDD does not simply reproduce memorized examples. However, when samples are colored by class (predicted class for generated samples), we observe distinct clusters 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.

5 Limitations

While our results provide strong evidence for the implicit denoiser hypothesis, several limitations suggest directions for future work. Our detailed spectral analysis of $\mathbf{J}^\top\mathbf{J}$ eigenstructure is currently limited to relatively simple architectures and small dataset (MNIST) due to computational tractability. Also, the space of possible generative outputs accessible through PGDD is vast, determined by the complex interaction between input noise patterns, algorithm hyperparameters, and the network's learned priors. Our experiments represent only a small fraction of this combinatorial space. The rich diversity of patterns achievable through different parameter configurations suggests that systematic exploration of this landscape could reveal much deeper insights into the structure of implicit generative

models. This limitation also presents an opportunity: PGDD exhibits natural convergence properties, with trajectories consistently arriving at stable attractors in the learned prior distribution. With appropriate hyperparameter tuning strategies, the algorithm could automatically navigate to the nearest meaningful attractor without manual parameter selection. Developing principled methods for adaptive hyperparameter adjustment based on convergence dynamics would significantly improve the method's practical applicability across different architectures and datasets. Such automation would also enable more systematic exploration of the vast space of implicit priors encoded by robust classifiers.

Finally, while we demonstrate correlations between spectral concentration and denoising capability, establishing direct causal relationships would strengthen our theoretical framework. Controlled experiments that systematically perturb the spectral properties of $\mathbf{J}^\top\mathbf{J}$ through targeted interventions during training or post-hoc modifications, and measure the resulting impact on generative quality, would provide more definitive evidence for the mechanistic role of eigenvalue structure. Such experiments would help distinguish between correlation and causation in the relationship between robustness training and generative capabilities, potentially revealing new ways to enhance implicit denoising through architectural or training modifications.

6 Conclusion

This work provides evidence for a connection between adversarial robustness and generative modeling by demonstrating that robust classifiers appear to contain implicit denoising structure encoded in their Jacobian operators. Our theoretical framework suggests that the Gram operator $\mathbf{J}^\top\mathbf{J}$ can function as an implicit denoiser through the low-rank spectral structure induced by adversarial training. This mathematical insight offers a potential bridge between discriminative robustness and generative inference within a unified framework. Our empirical validation across datasets and architectures confirms the theoretical predictions. Robust classifiers exhibit stronger denoising capabilities compared to standard networks, with spectral properties that enable selective amplification of discriminative directions while suppressing noise. Visual analysis reveals that applying $\mathbf{J}^\top\mathbf{J}$ to random perturbations produces class-consistent structure in robust models but preserves random patterns in standard classifiers.

The Prior-Guided Drift Diffusion (PGDD) algorithm translates these theoretical insights into practical generation capabilities. PGDD leverages implicit denoising structure through simple inference objectives rather than explicit Jacobian computation, enabling coherent image synthesis from noise without requiring generative training or architectural modifications. Large-scale trajectory analysis demonstrates consistent convergence properties across diverse initialization conditions, indicating that PGDD reliably accesses meaningful statistical priors rather than exploiting spurious patterns. The extension to standard networks through sPGDD reveals that implicit generative structure exists beyond robustly trained classifiers, though with reduced fidelity. Notably, both approaches—adversarial training (which smooths the loss landscape during training) and sPGDD (which smooths gradients during inference)—demonstrate that regularization techniques can provide access to these implicit generative structures, suggesting that smoothing mechanisms may be a general principle for exposing the dual discriminative-generative nature of neural networks.

The finding that PGDD-generated samples occupy distinct manifolds from training data while maintaining class consistency demonstrates that robust classifiers encode implicit structural knowledge, though their current generative capacity remains limited. These findings suggest potential applications in interpretability and explainability by enabling access to the learned priors embedded within discriminative networks. The observed generative properties indicate that robust training may encode richer representations of data distributions than previously recognized, though further investigation is needed to characterize these representations fully. This work provides a foundation for exploring connections between robustness and generation. While prior work has attempted to develop training objectives that jointly optimize discriminative and generative performance, such approaches have proven difficult to optimize effectively. The implicit structure revealed here may offer new insights into why these joint objectives remain challenging and potentially suggest alternative approaches for leveraging the dual nature of these learned representations.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

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.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *International Conference on Machine Learning*, pp. 1675–1685, 2019.

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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

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.

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.

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.

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.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

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.

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.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

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). We made minimal adaptations (e.g. adding an ℓ_2-PGD attack and logging spectral metrics).

Architecture. The CNN used in all MNIST experiments has the following structure:

- Conv2d(1→32, kernel=3, padding=1), ReLU
- Conv2d(32→32, kernel=3, padding=1, stride=2), ReLU
- Conv2d(32→64, kernel=3, padding=1), ReLU
- Conv2d(64→64, kernel=3, padding=1, stride=2), ReLU
- Flatten
- Linear($7{\cdot}7{\cdot}64 \to 100$), ReLU
- Linear($100 \to 10$)

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 $Q = J^\top J$ (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 2.

Table 2: 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 $J^\top J$ 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

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

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 $J^\top J$ at test points using subspace iteration with Gram–Schmidt re-orthogonalization. Starting from k random vectors (we use $k = 8$), we repeatedly apply $J^\top J$ and orthogonalize, then compute Rayleigh quotients $\lambda_i = \langle v_i, J^\top J v_i \rangle$. 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 $J^\top J\epsilon$. This illustrates that robust models suppress noise and emphasize structured residuals, consistent with implicit denoising.

Models and evaluation. We compare a standard CNN and an ℓ_2-robustly trained CNN (see Section S1 for architecture and training details). All probes are run in PyTorch on MNIST test examples, with evaluation restricted to a single held-out image for visualizations and to batches of 32 for statistics. The reported quantities in Table 1 are averaged over the test batch.

C.1 Prior-Guided Drift Diffusion: Algorithm, Intuition, and Theory

We present the detailed algorithm for Prior-Guided Drift Diffusion (PGDD), together with the underlying intuition and theoretical justification for how PGDD grants access to the learned priors of a network. The method is designed to be both conceptually transparent and practically simple, offering a principled way to leverage the implicit generative structure in networks which were not explicitly trained for pattern generation (notably classifiers).

Algorithm 1 Prior-Guided Drift Diffusion Objective

Input: Image x_{input}, model f, target layer ℓ, constraint ϵ, step size α, noise ratio σ, iterations T
Output: Refined representations $\{x_t\}_{t=0}^{T}$
// Step 1: Feedforward pass
$x_0 \leftarrow \text{normalize}(x_{\text{input}})$
$f_\ell \leftarrow \text{extract\_layers}(f, \ell)$ {Extract model up to layer ℓ}
$x_{\text{noisy}} \leftarrow x_0 + \sigma \cdot \mathcal{N}(0, I)$
$r_{\text{anti-target}} \leftarrow f_\ell(x_{\text{noisy}})$ {Generate noisy reference representation}
for $t = 1$ to T **do**
  // Step 2: Inference objective selection
  anti-target $\leftarrow r_{\text{anti-target}}$ {Use noisy reference as target}
  // Step 3: Feedback error propagation
  $h_t \leftarrow f_\ell(x_{t-1})$ {Forward pass through target layers}
  $L_t \leftarrow \|h_t - r_{\text{anti-target}}\|^2$ {MSE loss in representation space}
  $g_t \leftarrow \nabla_{x_{t-1}} L_t$ {Gradient via feedback pathways}
  // Step 4: Constrained activation update
  $\tilde{g}_t \leftarrow \alpha \cdot g_t / (\|g_t\| + \text{1e-10})$ {Normalize gradient}
  $\eta_t \leftarrow \text{diffusion\_noise\_ratio} \cdot \mathcal{N}(0, I)$ {Add stochastic noise}
  $x'_t \leftarrow x_{t-1} + \tilde{g}_t + \eta_t$ {Move away from representation of noisy input}
  $x_t \leftarrow \text{project}(x'_t, x_0, \epsilon)$ {Enforce $\|x_t - x_0\|_\infty \leq \epsilon$}
end for
Return $\{x_t\}_{t=0}^{T}$ =0

C.2 PGDD parameters for main text figures

Table 3: PGDD parameter settings for main test figures (ImageNet ResNet-50, L_2 adversarially trained).

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

C.2.1 sPGDD IN STANDARD NETWORKS

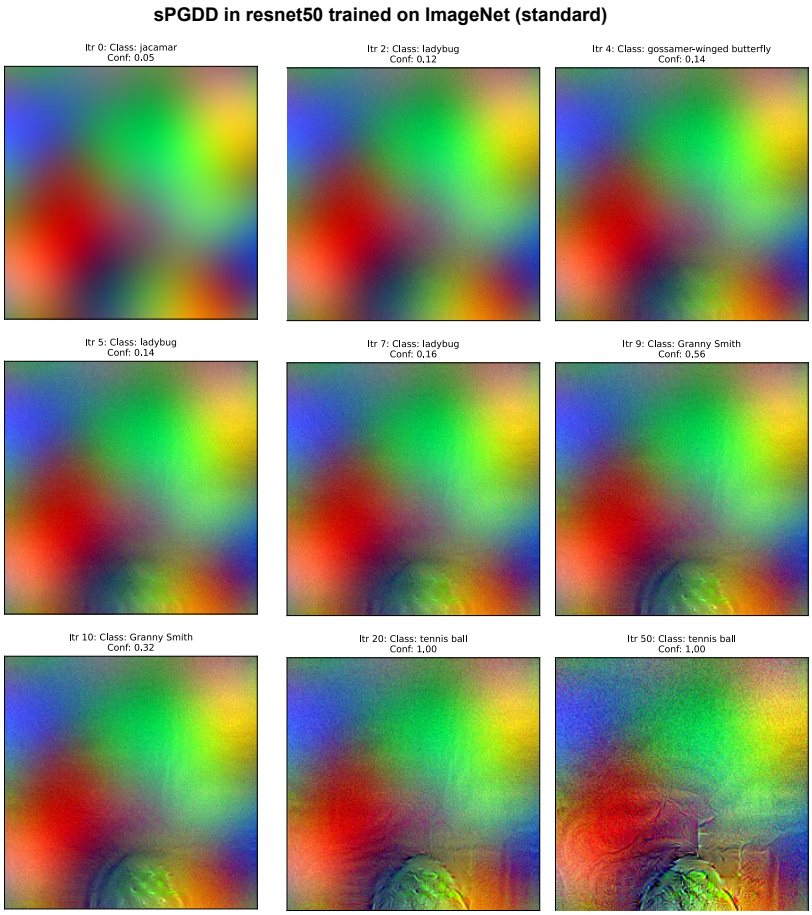

Figure S1: **Smooth PGDD (sPGDD) enables generation with standard networks.** sPGDD applied to a standard ResNet-50 (ImageNet, no adversarial training) demonstrates generative capability through gradient smoothing. The method uses multiple independently sampled noise perturbations at each iteration and averages the resulting gradients to reduce variance and emphasize stable prior information embedded in the network. Here, inference starts from a sample of Perline noise.

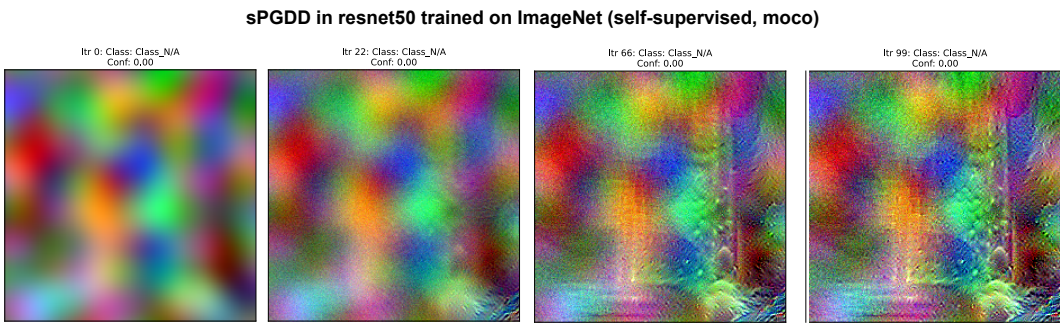

Figure S2: **sPGDD in resnet50 trained with self-supervised objective (moco)**

C.3 sPGDD PARAMETERS FOR FIGURES S1 AND S2

Table 4: sPGDD parameter settings

Figure	Model	Loss Inference	Loss Function	Drift Noise Ratio	Diffusion Noise Ratio	ϵ_{infer}	$\sigma_{smoothing}$	$n_{smoothing}$
S1	ResNet-50	sPGDD	MSE	0.2	0.01	40	0.1	100
S2	ResNet-50 moco	sPGDD	MSE	0.5	0.01	40	0.1	100