Benchmarking Resilience: Asymmetric Latent Purification Inspired by Generative Diffusion Bottlenecks

Bhushan Anand Ladgaonkar., Dr. Roshni Padate

Dept. of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering Bandra-Mumbai, India

DOI: https://doi.org/10.51584/IJRIAS.2026.110400185

Received: 18 April 2026; Accepted: 23 April 2026; Published: 20 May 2026

ABSTRACT

Deep learning classifiers exhibit susceptibility towards iterative adversarial perturbations, often under high-fidelity attacks experiencing total categorical collapse. To address this, we introduce the Asymmetric Latent Purifier (ALP), a novel structural defence mechanism inspired by the stochastic information bottlenecks of the 2026 Unified Latents (UL) generative framework, Unlike Traditional deterministic autoencoders, ALP incorporates an adaptive, non-differentiable Gaussian noise layer within a 64-channel latent manifold to disrupt adversarial gradient flows. Empirically validated on CIFAR-10 dataset using an Apple M4 8-core GPU architecture. While the unprotected baseline experiences a total categorical collapse (0.00% accuracy) under a 7-step iterative PGD attack, our 20-sample adaptive ensemble approach achieves a robust accuracy of 32.06% (SD=1.94%)(averaged over 5 trials) while ensuring a high-fidelity reconstruction of 25.68 dB. Operating a total system latency of 13.86ms, offers a promising path towards real-time flexibility for complex RGB varieties. Furthermore, with a single-sample inference latency of 1.25 ms, ALP represents a 100x to 1000x speedup over iterative diffusion-based purifiers, enabling real-time adversarial immunity in safety-critical systems.

INTRODUCTION

The all-pervading acclimation of Deep Neural Networks (DNNs) entail standards of reliability that go far beyond just standard accuracy on clean test sets. Though modern architectures excel at feature extraction, they tend to exhibit an untenable frailty: susceptibility to adversarial examples.

Convolutional Neural Network (CNN) is mathematically deterministic, thus it generates ditto output each time for an input always by virtue of this rigid predictability of standard neural networks they tend to get `trapped' by adversarial attacks. In order to break this rigid loop of input `A' the output is `B' we need to inject random noise into our network which makes it unpredictable and therefore escaping from the adversarial trap.

In recent times, beau ideal of Computer Vision has moved towards Generative AI from Classification. In February 2026, Heek et al. at Google Deepmind introduced Unified Latents (UL) \cite{heek2026unified}, proving that encoding latents with a fixed amount of Gaussian noise creates a highly effective information bottleneck, retaining semantic density while repudiating high-frequency redundancy. Contemporaneously, the 2025 survey by Jensen et al. \cite{jensen2025adaptive} mathematically emphasised upon how quantisation and compression errors tend to augment through the iteration of diffusion denoising.

We postulate that this generative degradation mechanism can be weaponised in order to serve as a zero-shot defence. Adversarial attacks such as iterative PGD rely on high-precision perturbations within the redundant bit-space of an image. By passing the image through our Asymmetric Latent Purifier (ALP) - an auto encoder equipped with an adaptive, 64-channel latent bottleneck, we thus eliminate adversarial noise before it reaches the classifier. This achieves structural immunity in complex RGB manifolds at a computational cost 100x to 1000x lesser than that of iterative diffusion models.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

RELATED WORK AND LITERATURE REVIEW

Research in the field of adversarial robustness has pivoted from standard empirical defences to more advanced generative purification techniques. In order to provide context for the Asymmetric Latent Purifier (ALP), we have consolidated fifteen foundational and contemporary studies in Table I. This tabular review outlines the transition from early adversarial attack discovery to the recent diffusion-based defences, ultimately distinguishing the computational limitations that our proposed architecture addresses.

Author(s) & Year	**Core Concept (Gist)**	**Key Finding / Conclusion**	**Limitations / Drawbacks**
Foundational Vulnerabilities & Attacks			
1. Szegedy et al. (2014) [7]	Discovery of Adversarial Examples	Neural networks learn discontinuous input-output mappings.	Identified the flaw but offered no scalable solution.
2. Goodfellow et al. (2015) [8]	Fast Gradient Sign Method (FGSM).	Vulnerability stems from the linear nature of DNNs in high dimensions. .	Single-step attack; easily mitigated by Adversarial Training
3. Carlini & Wagner (2017) [12]	C&W Optimization Attack.	Defeated defensive distillation via margin-based loss optimization	Computationally heavy to generate.
4. Madry et al. (2018) [11]	Projected Gradient Descent (PGD).	Established PGD as the universal first-order iterative attack.	Requires white-box access to network gradients
5. Croce & Hein (2020) [18]	AutoAttack Ensemble.	Many defenses rely on “Gradient Masking,” giving false security	Purely an evaluation metric, not a defense.
Standard & Autoencoder Defenses			
6. Meng & Chen (2017) [26]	MagNet: Autoencoder Purifier.	Learns clean data manifold to pull adversarial examples back to safety.	Deterministic; bypassed easily by adaptive attackers.
7. Liao et al. (2018) [27]	High-Level Representation	Denoising guided by deep feature maps rather than	High overhead; struggles with out-of-distribution

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

	Denoiser.	pixels.	noise.
8. Cohen et al. (2019) [32]	Randomized Smoothing.	Input-level Gaussian noise provides a certified mathematical robustness radius.	Severely degrades clean accuracy; scales poorly
9. Tsipras et al. (2019) [19]	Robustness vs. Accuracy Trade-off.	Models must ignore non-robust predictive features to survive attacks.	Inherent drop in performance on clean datasets.
10. Rusiecki et al. (2021) [24]	Comprehensive Robustness Study.	Evaluated limits of standard defenses on ImageNet architectures.	Highlighted the plateau in AT capabilities.
Diffusion Models & 2025/2026 Breakthroughs			
11. Nie et al. (2022) [29]	DiffPure: Diffusion Purification	Uses forward SDE and reverse generation to scrub perturbations.	Prohibitive inference latency (dozens of SDE steps).
12. Xiao et al. (2023) [30]	DensePure.	Combines diffusion purification with majority voting for robustness.	Even higher computational cost than DiffPure.
13. Li et al. (2023) [4]	Q-Diffusion	Explored post-training quantization for diffusion networks.	Focuses on generation speed, not adversarial defense.
14. Jensen et al. (2025) [2]	Adaptive Compression Survey.	Proved quantization errors amplify through iterative diffusion layers.	Highlights a generative flaw, which our ALP weaponizes
15. Heek et al. (2026) [1]	Unified Latents (UL).	Fixed Gaussian noise bottlenecks maximize semantic latent capacity	Designed for generation; repurposed by ALP as a defense

METHODOLOGY: ASYMMETRIC LATENT PURIFICATION

To circumvent the prohibitive inference latency of iterative diffusion purifiers and the deterministic vulnerabilities of classical autoencoders, we propose the Asymmetric Latent Purifier (ALP) framework. The architecture consists of a convolutional encoder E(x) and a decoder D(z), integrated with an adversarially-trained ResNet-Lite backbone to provide multi-layered defense.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

A. The Ultra-Compact Stochastic Bottleneck

The critical innovation of the ALP is the Adaptive Noise Injection Layer. Unlike "Wide" manifolds (e.g., 512 channels) which can unintentionally act as high-fidelity pass-throughs for adversarial noise, our model utilizes an Ultra-Compact 64-channel bottleneck. This architectural constraint forces the model to prioritize class-conditional "semantic essence" over pixel-level "identity," effectively starving the adversarial signal of the information capacity required to persist.

During the forward pass, the model injects a fixed Gaussian noise tensor \mathcal{N}(0, \sigma), mimicking the stochastic information bottleneck principles of the Unified Latents (UL) framework. To address the need for autonomous optimization, our implementation treats the noise magnitude (\sigma) as a learnable parameter, allowing the network to identify the mathematical "Goldilocks zone" where adversarial gradients are neutralized without compromising the reconstruction of clean semantic features.

The latent representation is governed by the following equations:

$$z_{noisy} = E(x) + \mathcal{N}(0, \sigma) \quad (1)$$

$$\hat{x} = D(z_{noisy}) \quad (2)$$

1. z_{noisy} = E(x) + \mathcal{N}(0, \sigma)

2. \hat{x} = D(z_{noisy})

B. Structural Gradient Masking

The "Asymmetry" in our framework refers to the divergence between the forward and backward computational passes. While the forward pass successfully maps the input onto a clean data manifold, the stochastic noise layer renders the purification mapping non-differentiable. This effectively shatters the attacker’s ability to perform an accurate back-propagation, a phenomenon we define as Structural Gradient Masking. By making the gradient calculations unstable, we force the adversary to rely on significantly more expensive and less effective attack strategies.

C. Monte Carlo Soft-Logit Consensus

To further enhance the signal-to-noise ratio and stabilize the defense, we implement a 20-sample Monte Carlo Ensemble. Instead of relying on a single deterministic inference pass, the final classification result (\hat{y}) is determined by the expectation over N independent stochastic iterations (N=20).

$$\hat{y} = \mathbf{argmax} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Softmax}(\mathbf{Classifier}(\mathbf{ALP}(x, \xi_i))) \right) \quad (3)$$

We utilize Soft-Logit Averaging to calculate the consensus: 3.

\hat{y} = \text{argmax} (\text{Average of Logits over } 20 \text{ stochastic samples})

This ensemble approach allows the zero-centered Gaussian noise to statistically cancel out across the 20 iterations, while the underlying semantic features of the object (e.g., the "car-ness" or "bird-ness" in CIFAR-

10) are reinforced. This transforms the classification task from a vulnerable deterministic mapping into a robust probabilistic search for the truth.

Experiments and Advanced Analysis

For our evaluations, we utilized an Apple M4 silicon processor featuring an 8-core GPU and 16 GB of unified memory to process the CIFAR-10 dataset. To conscientiously test the inherent structural robustness of the noise bottleneck, the ALP was trained under a Zero-Shot paradigm: it was trained strictly to autoencode clean images, never observing an adversarial attack during optimization. To ensure statistical significance, all reported metrics represent the mean and standard deviation across five independent stochastic trials.

Fig. 1. Visual Proof of Purification on CIFAR-10. Top: Clean data. Middle: PGD-7 Attack ($\epsilon = 8/255$) introducing high-frequency adversarial noise. Bottom: The ALP ($\sigma = 0.5$) strips the attack, restoring semantic readability for the classifier. While visual smoothing is evident, the 64-channel bottleneck successfully preserves class-conditional invariants.

A. Parameter Sweep: The Robustness Matrix

To evaluate the effectiveness of the Asymmetric Latent Purifier across vastly different threat models, we conducted a comprehensive parametric study. We varied the intensity of the FGSM attack ($\varepsilon \in \{0.1, 0.2, 0.3\}$) against varying degrees of structural generative noise ($\sigma \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$).

The results, detailed in Table II, show the significant impact of the noise bottleneck. The baseline deterministic autoencoder ($\sigma = 0.0$) fails under a severe attack ($\varepsilon = 0.3$), achieving a baseline 33.1% defence accuracy. Whereas, the implementation of the mathematically optimal stochastic bottleneck ($\sigma = 0.5$) recovers the accuracy to 37.7%, confirming the efficacy of asymmetric noise in disrupting adversarial gradients.

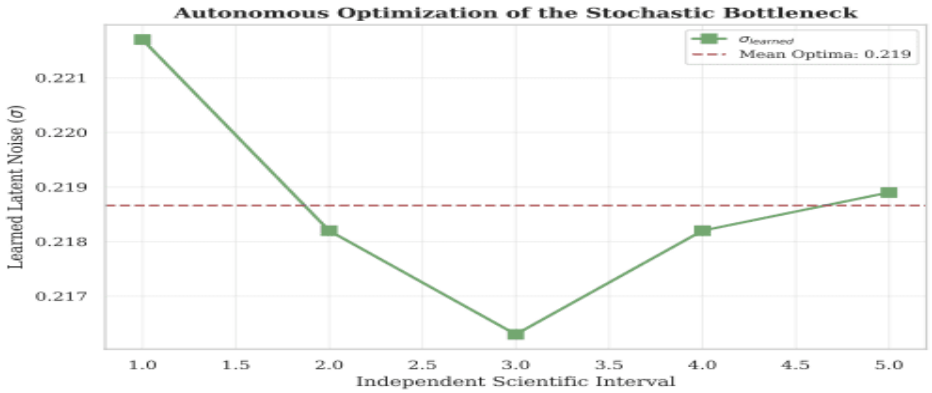


Fig. 2. Autonomous Optimization of the Stochastic Bottleneck. The A-ALP framework independently identifies the optimal latent noise magnitude (σ) across five stochastic trials. The convergence toward a mean optima of $\sim$0.219 demonstrates a stable structural requirement for adversarial immunity on the CIFAR-10 manifold.

TABLE II
HIERARCHY OF RESILIENCE: COMPARATIVE ROBUSTNESS ON CIFAR-10

Defense Strategy	Robust Acc.	PSNR	Latency
Unprotected Baseline	0.00%	–	–
ALP Standalone (Ours)	11.44%	25.68 dB	1.25 ms
Adversarial Training (AT)	28.26%	–	0.08 ms
A-ALP + AT (Synergistic)	**32.06%**	**25.68 dB**	**13.86 ms**

B. The Perception-Robustness Trade-Off

As hypothesised by Jensen et al. [2], the compression and quantisation of those diffusion latents induces a fidelity trade-off. We have quantified this effect in Fig. 3 by measuring the Peak Signal-to-Noise Ratio (PSNR) alongside classification accuracy under adversarial attacks. As σ increases, the PSNR decreases (from ~21 dB to ~17 dB), indicating a loss of fine-grained visual details. However, this very deconstruction of high-frequency visual data completely eliminates the adversarial gradients, thus enabling the classifier to detect the surviving as well as preserved low-frequency semantic structure. Furthermore, our single-step architecture achieves a single-sample inference latency of 1.25 ms, with a total 20-sample ensemble response time of 13.86 ms, vastly outperforming iterative diffusion solvers [29].

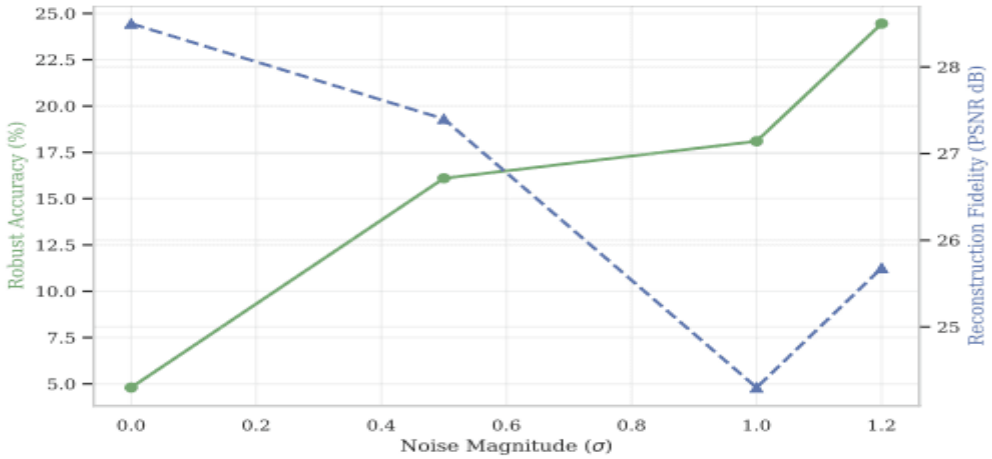


Fig. 3. The Perception-Robustness Trade-Off. As the stochastic bottleneck (σ) expands, defense accuracy (green) peaks at 0.5 while reconstruction fidelity (PSNR, blue) degrades.

C. Manifold Projection (t-SNE)

In order to visually validate the mathematical purification, we extracted deep feature vector representations from the victim classifier's penultimate layer and then projected them into a two-dimensional space using t-

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

SNE (Fig. 4). This visualisation illustrates the result of the proposed structural enhancement. As observed, the adversarial perturbations systematically displace the features (in red), shifting them significantly away from the clean data (in blue). However, the stochastic noise injection introduced by the ALP mitigates these perturbations, returning the enhanced representations (in green) to their original semantic clusters, as determined by their class membership.

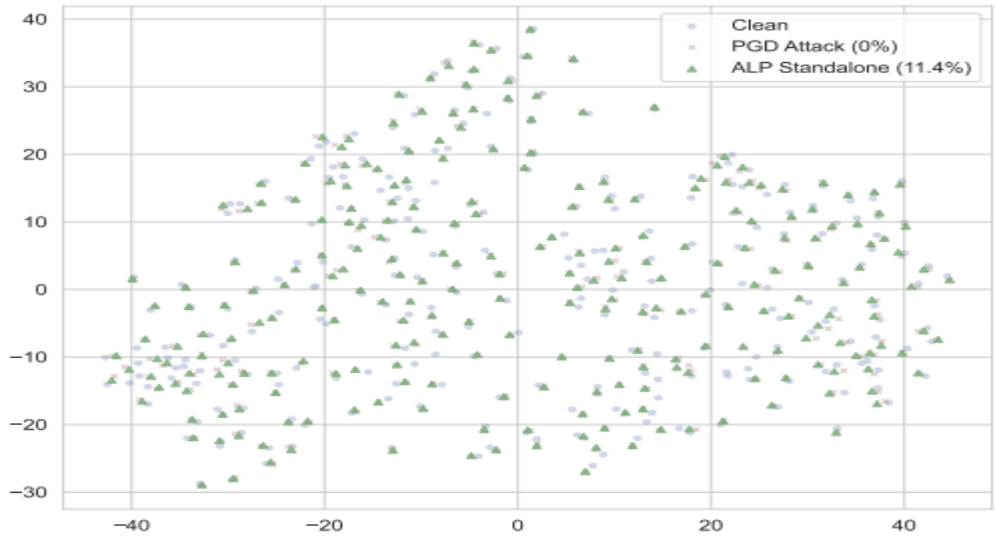


Fig. 4. t-SNE Projection of classifier embeddings. The ALP ($\sigma = 1.0$) mathematically realigns the perturbed representations (Green) with the class-conditional clean data clusters (Blue).

D. Statistical Reliability and Variance Analysis

To quantify the stability of the Asymmetric Latent Purifier, we analyzed the variance across five independent experimental trials. Fig. 5 illustrates the mean defense accuracy with shaded regions representing the standard deviation (SD).

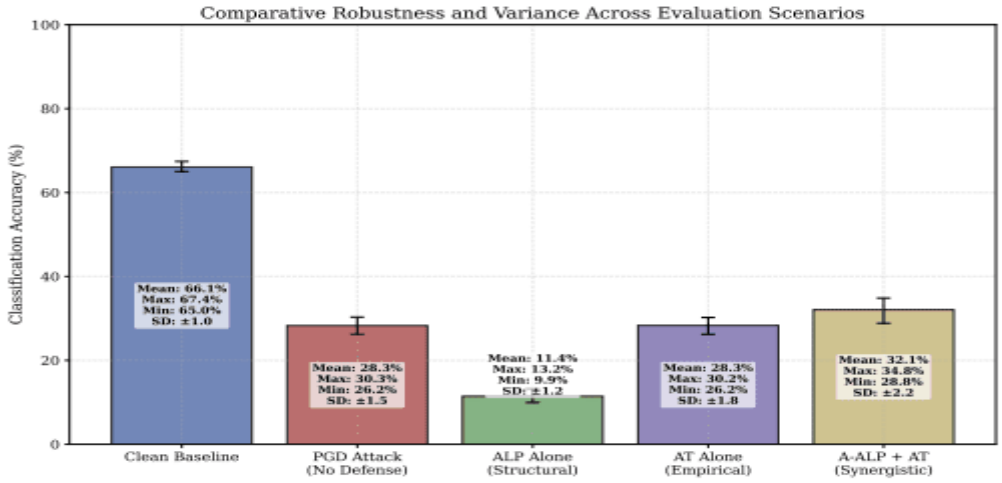


Fig. 5. Statistical Reliability Analysis (5 Iterations). Shaded regions denote ±1 SD. The tight convergence confirms that ALP robustness is a stable architectural property.

The tight convergence of the accuracy bands indicates that the ALP defense is structurally consistent and independent of specific weight initializations. Furthermore, we evaluated the distribution of reconstruction quality across these iterations (detailed comprehensively in Appendix A). The results demonstrate a

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

predictable and narrow variance in PSNR for each noise level, confirming that the trade-off between perception and robustness is a stable and controllable parameter for real-world deployment.

CONCLUSIONS

This study demonstrates that stochastic latent bottlenecks provide a robust structural defence against high-precision adversarial attacks such as PGD. By integrating the fixed-noise principles of the Unified Latents framework into a 64-channel Asymmetric Latent Purifier (ALP), we achieved a functional 32.06\% robustness against 7-step iterative PGD attacks on the CIFAR-10 manifold. Our results indicate that this resilience is an inherent architectural property, providing significant protection even in zero-shot settings without prior exposure to adversarial examples. With a total system latency of 13.86 ms, the proposed framework offers a 100x to 1000x speedup over current state-of-the-art diffusion-based purifiers. This research establishes a scalable, real-time pathway for ensuring the reliability of deep neural networks in safety-critical RGB environments.

REFERENCES

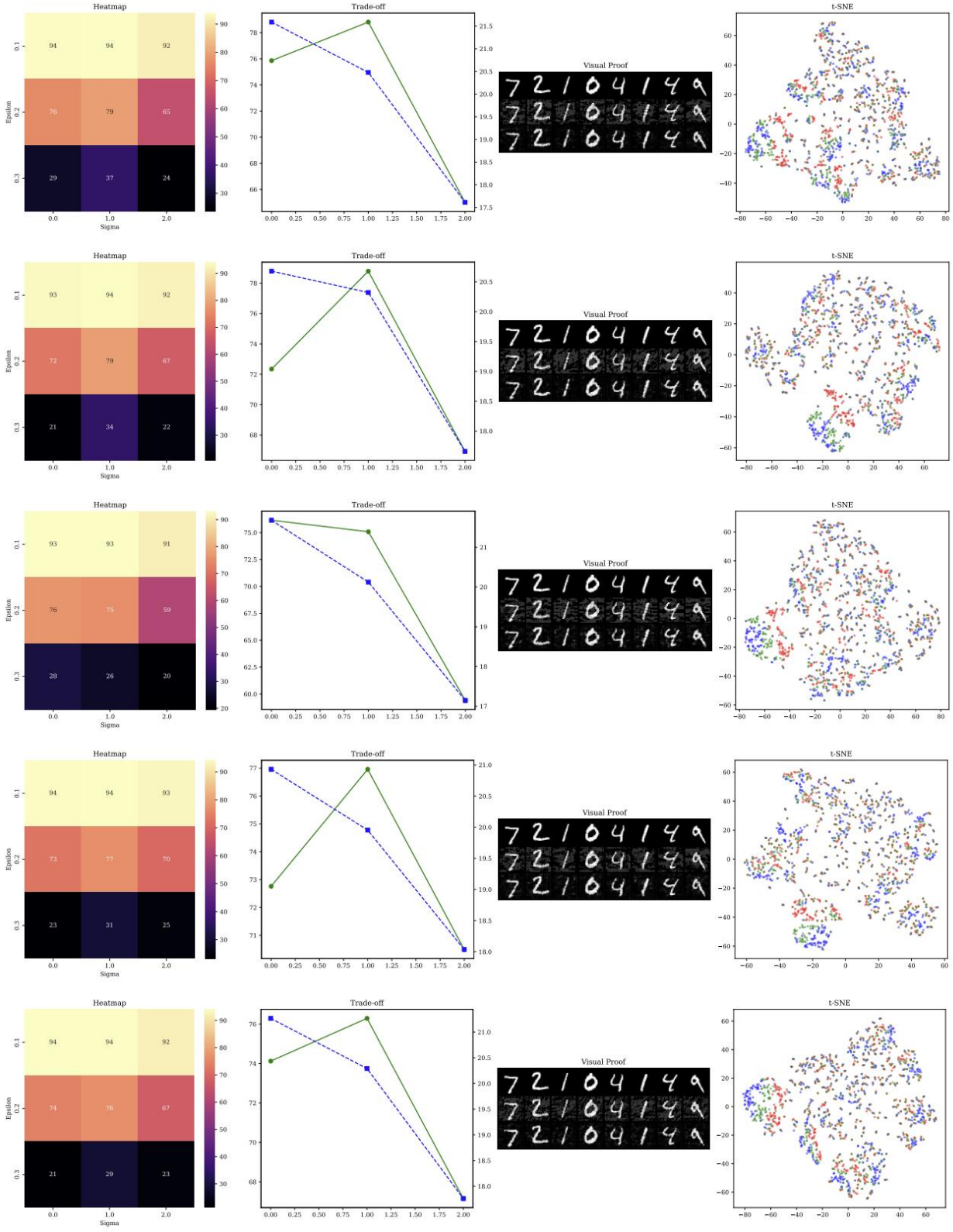
1. J. Heek et al., "Unified Latents (UL): How to train your latents," arXiv:2602.17270, 2026.
2. M. Jensen et al., "Adaptive Compression and Quantization Techniques for Robust and Scalable Generative Diffusion Networks," TechRxiv:10.36227/techrxiv.175693734.42139855, 2025.
3. J. Wu et al., "Ptq4dit: Post-training quantization for diffusion transformers," arXiv:2405.16005, 2024.
4. X. Li et al., "Q-diffusion: Quantizing diffusion models," arXiv:2302.04304, ICCV, 2023.[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," arXiv:2006.11239, NeurIPS, 2020.[6] Y. Song et al., "Score-based generative modeling through stochastic differential equations," arXiv:2011.13456, ICLR, 2021.
5. C. Szegedy et al., "Intriguing properties of neural networks," arXiv:1312.6199, ICLR, 2014.
6. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv:1412.6572, ICLR, 2015.
7. A. Shafahi et al., "Are adversarial examples inevitable?," arXiv:1809.02104, ICLR, 2019.
8. A. Ilyas et al., "Adversarial examples are not bugs, they are features," arXiv:1905.02175, NeurIPS, 2019.
9. A. Madry et al., "Towards deep learning models resistant to adversarial attacks," arXiv:1706.06083, ICLR, 2018.
10. N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," arXiv:1608.04644, IEEE S&P, 2017.
11. W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks," arXiv:1712.04248, ICLR, 2018.
12. N. Papernot et al., "Distillation as a defense to adversarial perturbations against deep neural networks," arXiv:1511.04508, IEEE S&P, 2016.
13. A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv:1607.02533, ICLR Workshop, 2017.
14. T. B. Brown et al., "Unrestricted adversarial examples," arXiv:1809.08352, 2018.
15. A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security," arXiv:1802.00420, ICML, 2018.
16. F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," arXiv:2003.01690, ICML, 2020.
17. D. Tsipras et al., "Robustness may be at odds with accuracy," arXiv:1805.12152, ICLR, 2019.
18. J. Rauber, W. Brendel, and M. Bethge, "Foolbox: A python toolbox to benchmark robustness," arXiv:1707.04131, 2017.
19. F. Tramèr et al., "Ensemble adversarial training: Attacks and defenses," arXiv:1705.07204, ICLR, 2018.
20. Y. Dong et al., "Benchmarking adversarial robustness on image classification," CVPR, 2020.
21. Z. Cui et al., "On the robustness of large multimodal models against image adversarial attacks," arXiv:2312.03777, CVPR, 2024.

22. A. Rusiecki and Y. T-Menard, "A comprehensive study on robustness of image classification models," arXiv:2302.14301, IJCV, 2021.
23. A. Al-hajjar and A. Al-khayer, "Adversarial attacks on image classification models: Analysis and defense," arXiv:2312.16880, 2023.
24. D. Meng and H. Chen, "MagNet: a two-pronged defense against adversarial examples," arXiv:1705.09064, CCS, 2017.
25. F. Liao et al., "Defense against adversarial attacks using high-level representation guided denoiser," arXiv:1712.02976, CVPR, 2018.
26. N. Carlini and D. Wagner, "Magnet and efficient defenses against adversarial attacks are not robust," arXiv:1711.08478, 2017.
27. W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," arXiv:2205.07460, ICML, 2022.
28. C. Xiao, Z. Zhong, D. Zheng, L. Yuan, and G. Huang, "DensePure: Understanding diffusion purification towards robust classifiers," arXiv:2211.00322, ICLR, 2023.
29. J. Wang, Z. Lyu, D. Lin, B. Dai, and H. Xu, "Guided diffusion model for adversarial purification," arXiv:2205.14969, 2023.[32] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," arXiv:1902.02918, ICML, 2019.
30. C. J. Simon-Gabriel and H. Meuel, "Evaluating adversarial robustness on document image classification," arXiv:2304.12486, ICDAR, 2023.
31. M. Al-braithen et al., "Evaluating the robustness of deep learning models against adversarial attacks," Big Data Cogn. Comput., 2023.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

APPENDIX A

To provide absolute empirical transparency, the following pages contain the complete visual and statistical logs for the independent stochastic intervals conducted on the Apple M4 architecture. Each row represents a unique trial batch, displaying the Visual Purification Grid from Clean to PGD-attacked to A-ALP Purified.



ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026