MANI-PURE: MAGNITUDE-ADAPTIVE NOISE INJECTION FOR ADVERSARIAL PURIFICATION

Xiaoyi Huang[1], **Junwei Wu**[2], **Kejia Zhang**[1], **Carl Yang**[2], **Zhiming Luo**[1*]
[1]Xiamen University [2]Emory University

Project Page: https://phoshowy.github.io/MANI-Pure.github.io/

ABSTRACT

Adversarial purification with diffusion models has emerged as a promising defense strategy, but existing methods typically rely on uniform noise injection, which indiscriminately perturbs all frequencies, corrupting semantic structures and undermining robustness. Our empirical study reveals that adversarial perturbations are not uniformly distributed: they are predominantly concentrated in high-frequency regions, with heterogeneous magnitude intensity patterns that vary across frequencies and attack types. Motivated by this observation, we introduce **MANI-Pure**, a magnitude-adaptive purification framework that leverages the magnitude spectrum of inputs to guide the purification process. Instead of injecting homogeneous noise, MANI-Pure adaptively applies heterogeneous, frequency-targeted noise, effectively suppressing adversarial perturbations in fragile high-frequency, low-magnitude bands while preserving semantically critical low-frequency content. Extensive experiments on CIFAR-10 and ImageNet-1K validate the effectiveness of MANI-Pure. It narrows the clean accuracy gap to within **0.59%** of the original classifier, while boosting robust accuracy by **2.15%**, and achieves the **top-1** robust accuracy on the RobustBench leaderboard, surpassing the previous state-of-the-art method.

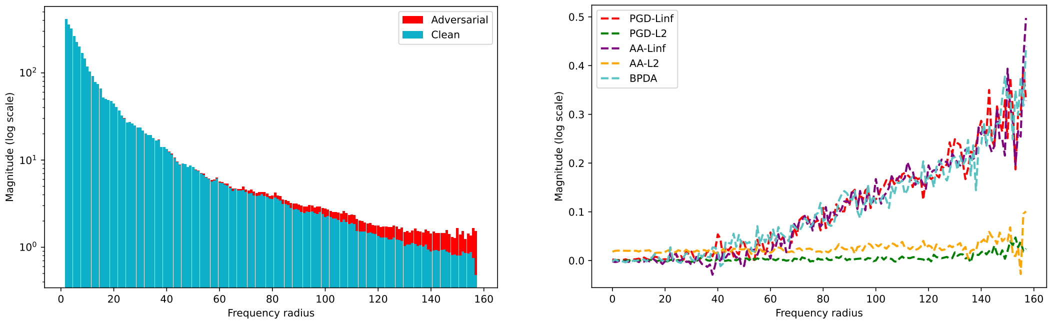
1 INTRODUCTION

Deep neural networks have achieved remarkable success across diverse applications. However, their vulnerability to adversarial perturbations remains a critical challenge (Weng et al., 2023; Tao et al., 2024; Goodfellow et al., 2014), particularly in safety-critical domains where reliability is paramount (Bortsova et al., 2021; Shao et al., 2025; Ye et al., 2024). A primary line of defense is adversarial training (AT), which augments training with adversarial examples to enhance robustness (Mao et al., 2023; Schlarmann et al., 2024). Although effective, AT incurs substantial computational costs and suffers from limited generalization, posing challenges for both large-scale and cross-domain deployment. These limitations have motivated an alternative paradigm: adversarial purification (AP). Unlike AT, AP does not require retraining classifiers; instead purifies adversarial inputs at inference, restoring them to clean representations (Samangouei et al., 2018; Nie et al., 2022). This design offers flexibility, scalability, and compatibility with off-the-shelf models.

Diffusion-based purification (DBP) has become the most effective and widely adopted approach in AP. It suppresses perturbations by injecting uniform noise in the forward process and then reconstructing images via reverse diffusion. Several variants have been proposed, such as the gradual noise scheduling (Lee & Kim, 2023) and the purification-enhanced AT method (Lin et al., 2024).

Despite these advances, existing DBP and related defense methods often assume that adversarial perturbations are uniformly distributed across the frequency domain—an assumption that is contradicted by empirical evidence. As shown in Figure 1a, radial spectral analysis reveals that perturbations are unevenly concentrated in the high-frequency region. Figure 1b reflects the heterogeneity in magnitude intensity across different frequency bands and attack strategies. As a result, uniform noise injection faces a trade-off: strong noise disrupts low-frequency semantics, reducing clean accuracy, whereas weak noise fails to suppress high-frequency perturbations, thereby compromising robustness. This motivates the need for frequency-adaptive purification that targets perturbation-prone regions while preserving semantic fidelity.

[*] Corresponding author.



(a) Magnitude distribution differences between clean and adversarial images

(b) Noise magnitude spectra of common adversarial attacks

Figure 1: Radial spectrum analysis of adversarial perturbations. Overall, adversarial noise aligns with clean samples in low-to-mid frequencies but diverges in high-frequency bands. Specifically, **Left:** adversarial samples show irregular high-frequency peaks with uneven magnitude distribution. **Right:** common attacks concentrate perturbations in high-frequency regions, yet their spectral distributions and intensities differ significantly. These observations highlight the **limitation of uniform noise injection** and directly motivate our magnitude-adaptive design.

To address this challenge, we propose MANI-Pure, a magnitude-adaptive purification framework that redesigns the diffusion process from the frequency-domain perspective. The framework comprises two complementary modules:

- **MANI** adaptively adjusts the noise injection intensity across different regions based on the magnitude spectrum, ensuring the injected noise aligns with the vulnerability to perturbations while preserving the original image semantics from excessive distortion.
- **FreqPure** (Pei et al., 2025a) employs magnitude–phase decomposition to explicitly distinguish low and high frequency components, preserving low-frequency content while focusing purification on high frequencies.

Together, MANI emphasizes magnitude-aware adaptivity, while FreqPure enforces explicit frequency constraints. Their synergy enables precise suppression of concentrated perturbations while maximally retaining semantic structure, thereby improving robustness across diverse attacks.

We conduct extensive evaluations on CIFAR-10 (Krizhevsky et al., 2010) and ImageNet-1K (Deng et al., 2009) under strong adaptive attacks, including PGD+EOT (Madry et al., 2017; Athalye et al., 2018), AutoAttack (Croce & Hein, 2020), and BPDA+EOT (Hill et al., 2021). Results show that MANI-Pure significantly enhances robustness while maintaining high clean accuracy, consistently outperforming existing DBP methods. Importantly, the framework is plug-and-play, readily applicable to modern architectures such as CLIP (Radford et al., 2021), without additional training cost.

In summary, our main contributions are briefly summarized as follows:

- We empirically verify that adversarial perturbations are concentrated in high-frequency bands and further reveal **distributional differences** between adversarial and clean samples in the magnitude spectrum.
- The proposed MANI-Pure framework combines magnitude-adaptive diffusion with frequency-domain purification, achieving a principled balance between **semantic fidelity and perturbation mitigation**, reflected in improvements to both clean and robust accuracy.
- Extensive experiments across datasets, attacks, and backbones demonstrate the superiority of our method in terms of **robustness**, **clean accuracy** and **perceptual quality**, as well as its scalability as a **plug-and-play** module.

2 Related Work

Adversarial purification provides a defense paradigm that restores adversarial inputs to clean representations at inference time, thereby avoiding the retraining cost of adversarial training.

Generative Models for Adversarial Purification. Early AP methods employed GANs, such as Defense-GAN (Samangouei et al., 2018), which projected adversarial samples onto the manifold of clean data. However, their limited generative fidelity and vulnerability to adaptive attacks significantly hindered their effectiveness. The advent of diffusion models marked a turning point: through stable likelihood-based training and high-quality reconstructions, they became the backbone of modern AP. Representative approaches include DiffPure (Nie et al., 2022), stochastic score-based denoising (Song et al., 2020), and gradient-guided purification like GDMP (Wang et al., 2022).

Precision Noise Injection. A key limitation of uniform noise injection lies in its disregard for the spectral structure of adversarial noise. Prior studies have shown that perturbations are often concentrate in high-frequency, low-magnitude regions (Yin et al., 2019). Building on this insight, FreqPure (Pei et al., 2025b) preserved low-frequency amplitude during reverse diffusion, effectively protects semantic content while targeting vulnerable high-frequency regions. These results highlight the importance of frequency-aware purification. Another line of research refines the forward noising process itself. Divide-and-Conquer (Pei et al., 2025a) integrates heterogeneous noise to better suppress adversarial perturbations, Sample-Specific Noise Injection (Sun et al., 2025) adapts noise to each input, and DiffCap (Fu et al., 2025) extends such ideas to vision–language models. While promising, these strategies remain largely fixed or heuristic, and they do not explicitly adapt to the actual spectral distribution of adversarial noise.

We unify these insights by introducing a magnitude-adaptive noise injection scheme that dynamically allocates noise to spectrally vulnerable regions, coupled with frequency-domain purification. This design enables precise suppression of perturbations while preserving semantic fidelity, thereby advancing AP toward finer-grained and more generalizable defenses.

3 Methodology

To eliminate adversarial perturbations while preserving semantic content, we propose **MANI-Pure**, a diffusion-based, frequency-domain purification framework comprising two complementary modules: **M**agnitude-**A**daptive **N**oise **I**njection (MANI) and **Freq**uency **Pur**ification (FreqPure). Figure 2 illustrates the overall structure. Before presenting the details, we briefly introduce the necessary background information.

3.1 Preliminaries

We briefly introduce diffusion model, adversarial purification, and the frequency-domain theory relevant to our method.

Diffusion Model. Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) generates data through a two-stage process: a forward noising process and a reverse denoising process.

Forward process. A sample x_0 is gradually perturbed into Gaussian noise through a Markov chain:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\Big)\,, \quad t = 1, \ldots, T, \tag{1}$$

where β_t follows a predefined variance schedule. By marginalization:

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{2}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Reverse process. To recover clean samples, the reverse distribution is approximated as

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}\big)\,. \tag{3}$$

Instead of predicting μ_θ directly, DDPM parameterizes it with a noise predictor $\epsilon_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t)\Big), \tag{4}$$

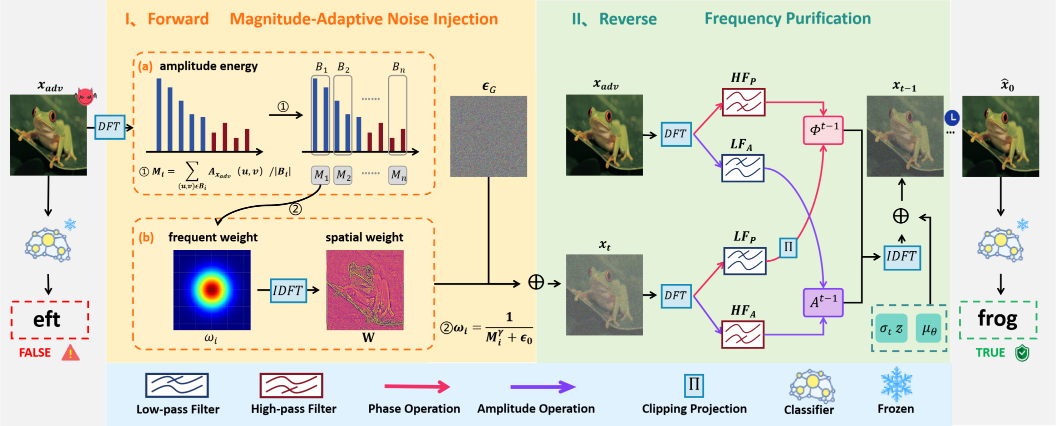


Figure 2: The pipeline of MANI-Pure. (I) **MANI**. Starting from an adversarial sample, we apply DFT to obtain its frequency representation, partition it into bands, compute average magnitudes, and derive band-wise and spatial weights. These weights modulate Gaussian noise to produce heterogeneous perturbations. (II) **FreqPure**. During the reverse process, the magnitude and phase spectra of the adversarial input and generated image are separated and recombined as shown, with the reconstructed image iteratively fed into subsequent denoising steps.

and the variance has a closed form:

$$\sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t. \tag{5}$$

Sampling. Starting from $x_T \sim \mathcal{N}(0, I)$, the model iteratively computes $x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z$ with $z \sim \mathcal{N}(0, I)$ until $\hat{x}_0$ is obtained.

Frequency-domain Theory. For an image $x \in \mathbb{R}^{H \times W}$, the discrete Fourier transform (DFT) yields

$$\mathcal{F}(x)(u, v) = \sum_{h,w} x(h, w)\, e^{-2\pi i (uh/H + vw/W)}. \tag{6}$$

Each Fourier coefficient can be expressed in polar form as

$$\mathcal{F}(x)(u, v) = A_x(u, v) \cdot e^{i\Phi_x(u,v)}, \tag{7}$$

where $A_x(u, v) = |\mathcal{F}(x)(u, v)|$ is the magnitude spectrum, reflecting the intensity of frequency components, and $\Phi_x(u, v)$ is the phase spectrum, encoding structural and semantic information.

3.2 Magnitude-Adaptive Noise Injection

Building upon the frequency-domain preliminaries introduced in Section 3.1, we leverage the magnitude spectrum of the adversarial input x_{adv} to capture the uneven distribution of frequency components. Specifically, the spectrum is partitioned into n non-overlapping frequency bands B_i. The average magnitude in each band is computed as

$$M_i = \frac{1}{|B_i|} \sum_{(u,v) \in B_i} A_{x_{\text{adv}}}(u, v), \tag{8}$$

where $|B_i|$ denotes the number of coefficients in band B_i. This corresponds to step (a) of the magnitude-adaptive noise injection on the left in Figure 2.

Low-magnitude bands are empirically more vulnerable to adversarial perturbations, while high-magnitude bands correspond to dominant semantic structures. To emphasize fragile regions, we assign larger weights to lower-magnitude bands:

$$w_i = \frac{1}{M_i^\gamma + \epsilon_0}, \tag{9}$$

where γ controls the sharpness of weighting and ϵ_0 prevents numerical instability when M_i is very small. The band-wise weights produce a frequency-domain weight distribution, which is transformed back to the spatial domain via inverse DFT to obtain a pixel-wise noise intensity map $\mathbf{W}$. In Figure 2, step (b) shows a visual representation of these two weights.

The spatial map $\mathbf{W}$ modulates Gaussian noise $\epsilon_G \sim \mathcal{N}(0, I)$ by element-wise multiplication:

$$\epsilon_t = \mathbf{W} \odot \epsilon_G, \quad \text{s.t. } \mathbf{W}, \epsilon_G \in \mathbb{R}^{H \times W \times C}. \tag{10}$$

Hence, the forward diffusion process becomes:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_{\text{adv}} + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_t, \tag{11}$$

where $\bar{\alpha}_t$ is the cumulative product of noise scheduling coefficients.

3.3 Frequency Purification

To complement MANI, we further adopt a frequency purification strategy (Pei et al., 2025b) during the reverse diffusion process. The key observation is that low-frequency magnitude components exhibit strong robustness against adversarial perturbations, whereas the phase spectrum is more easily affected across all frequencies.

For an image x_t generated during the reverse process, its DFT can be decomposed into magnitude A_t and phase Φ_t, with FreqPure handling them separately.

Magnitude purification. A low-pass filter $\mathcal{H}$ is applied to retain the low-frequency part of the adversarial input x_{adv}, while the high-frequency part is taken from the current generated image x_t:

$$A^{t-1} = \mathcal{H}(A_{\text{adv}}) + (1 - \mathcal{H})(A_t). \tag{12}$$

Phase purification. Low-frequency components are preserved through a projection operator $\Pi_\delta(\cdot)$ that restricts the generated phase within a small neighborhood of the adversarial phase:

$$\Phi^{t-1} = \mathcal{H}\big(\Pi_\delta(\Phi_t, \Phi_{\text{adv}})\big) + (1 - \mathcal{H})(\Phi_t), \tag{13}$$

where $\Pi_\delta(\Phi_t, \Phi_{\text{adv}})$ denotes clipping Φ_t into $[\Phi_{\text{adv}} - \delta,\ \Phi_{\text{adv}} + \delta]$, and δ is a hyperparameter controlling projection strength.

Reconstruction. The purified frequency representation (A^{t-1}, Φ^{t-1}) is then transformed back into the spatial domain using the inverse discrete Fourier transform (IDFT):

$$x_{t-1} = \mathcal{F}^{-1}\left(A^{t-1}, \Phi^{t-1}\right), \tag{14}$$

and iteratively participates in the reverse diffusion process until $\hat{x}_0$ is obtained. The above process is described in the corresponding module on the right side of Figure 2.

Overall, FreqPure leverages the stability of low-frequency magnitudes while constraining the phase distribution, preventing structural distortions. In contrast, MANI avoids redundant noise in robust regions and focuses perturbations on vulnerable frequency bands, enabling effective denoising with minimal semantic loss. Together, they are complementary: MANI selectively **suppresses adversarial signals** in the forward process, while FreqPure ensures frequency stability and semantic consistency in the reverse process. The above methods are summarized in Appendix E.1.

4 Experiments

4.1 Experimental Setup

Datasets and Model Architectures. We conduct experiments on two widely used datasets of different resolutions: CIFAR-10 and ImageNet-1K. Following the settings in prior works (Pei et al., 2025a; Zhang et al., 2025b), we randomly select 512 samples from CIFAR-10 and 1,000 samples from ImageNet-1K for evaluation. To better align with the development of large-scale multimodal models, we adopt CLIP as the frozen classifier to accomplish zero-shot classification tasks. For the diffusion models, we use the publicly released unconditional CIFAR-10 checkpoint of EDM (Karras et al., 2022) for CIFAR-10, and 256x256 unconditional diffusion checkpoint for ImageNet-1K.

Evaluation Metrics. We report both standard accuracy and robust accuracy. This dual evaluation provides a comprehensive view of the trade-off between preserving performance on clean data and enhancing resilience against attacks.

Attack Settings. In our experiments, we evaluate all defenses under strong adaptive attacks across both ℓ_∞ and ℓ_2 threat models. Concretely, we employ PGD and AutoAttack as primary evaluation tools, covering both ℓ_∞ and ℓ_2 perturbations. Following Lee & Kim (2023), we adopt PGD combined with expectation over transformations (PGD+EOT) to mitigate variability caused by stochastic components in the defense. In addition, we test BPDA+EOT to evaluate attacks that approximate gradients through non-differentiable or randomized components. For computational tractability while retaining attack strength, PGD and BPDA are run for 10 iterations, and EOT uses 10 samples per gradient estimate. AutoAttack is executed in its standard version. The perturbation budgets are specified as $\epsilon = 8/255$ for ℓ_∞ attacks on CIFAR-10, $\epsilon = 4/255$ for ℓ_∞ attacks on ImageNet, and $\epsilon = 0.5$ for ℓ_2 attacks on both datasets. Further experimental settings can be found in Appendix D.

4.2 Main Results

This section presents a comprehensive evaluation of MANI-Pure across multiple datasets, attack settings, and metrics, with a focus on **robustness**, **perceptual quality**, and **plug-and-play flexibility**.

4.2.1 Classification Accuracy under Adaptive Attacks

MANI-Pure consistently achieves the best trade-off between standard and robust accuracy across datasets and backbones. As summarized in Table 1 (CIFAR-10, ViT-L/14), Table 2 (CIFAR-10, RN50), and Table 3 (ImageNet-1K, ViT-L/14), we evaluate against strong adaptive attacks including PGD+EOT, AutoAttack under both ℓ_∞ and ℓ_2 norms, and BPDA+EOT.

On CIFAR-10, MANI-Pure improves robust accuracy by **2.15%** under AutoAttack (ℓ_∞) and by **2.54%** under BPDA+EOT when using ViT-L/14. Consistent improvements are also observed on RN50, confirming the backbone-agnostic nature of our framework. On ImageNet-1K, especially, MANI-Pure achieves the highest robust accuracy, outperforming all baselines by **3.8%** under BPDA+EOT, while maintaining competitive clean accuracy.

These results demonstrate that MANI-Pure not only surpasses existing AP and AT baselines (including recent leaders on RobustBench), but also exhibits strong cross-dataset generalization and backbone versatility. More results on different backbones can be found in the Appendix. E.2.

Table 1: Classification accuracy on CIFAR-10 under adversarial attacks using CLIP ViT-L/14. Zero-shot CLIP (w/o defense) is denoted by †, its standard accuracy as the upper bound. Methods from the Robustbench leaderboard are denoted by ‡. AT and AP methods are marked accordingly.

Type	**Algorithm**	**Standard**	**PGD**		**AutoAttack**		**BPDA**
			ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	
AT	*DHAT* (Zhang et al., 2025a)	85.45	63.14	66.91	56.77	57.40	54.84
	DIAT (Wang et al., 2023) ‡	92.69	71.38	85.12	70.53	84.03	69.76
	MeanSparse (Amini et al., 2024) ‡	92.98	74.02	86.41	68.85	85.98	72.87
AP	Zero-shot (w/o defense)†	94.73	2.15	55.86	0.00	0.00	0.78
	+ *DiffPure* (Nie et al., 2022)	86.52	85.55	85.74	85.35	85.55	84.96
	+ *DDPM++* (Song et al., 2020)	86.33	84.77	85.16	85.74	85.74	86.13
	+ *REAP* (Lee & Kim, 2023)	81.45	79.69	79.87	80.08	80.18	80.86
	+ *FreqPure* (Pei et al., 2025b)	91.77	90.17	91.41	90.82	91.99	87.89
	+ *CLIPure* (Zhang et al., 2025b)	93.55	89.06	92.19	90.04	92.38	83.01
	+ *Ours*	**94.14**	**91.02**	**92.58**	**92.19**	**93.16**	**88.67**

Table 2: Classification accuracy on CIFAR-10 under adversarial attacks using CLIP RN50. Zero-shot CLIP (w/o defense) is denoted by †, its standard accuracy as the upper bound. Only AP-based methods are included.

Algorithm	Standard	PGD		AutoAttack		BPDA
		ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	
Zero-shot (w/o defense)†	69.92	0.00	19.73	0.39	0.39	3.32
+ *DiffPure* (Nie et al., 2022)	61.91	59.77	61.13	59.77	60.64	60.16
+ *DDPM++* (Song et al., 2020)	56.64	56.25	56.64	56.05	56.34	55.27
+ *REAP* (Lee & Kim, 2023)	58.59	56.84	58.40	55.66	58.40	56.25
+ *FreqPure* (Pei et al., 2025b)	62.70	59.38	60.55	61.52	62.56	58.79
+ *CLIPure* (Zhang et al., 2025b)	61.33	53.71	60.55	56.84	60.55	53.32
+*Ours*	**65.23**	**61.91**	**62.50**	**62.70**	**64.84**	**60.16**

Table 3: Classification accuracy on ImageNet-1K under adversarial attacks using CLIP ViT-L/14. Zero-shot CLIP (w/o defense) is denoted by †, its standard accuracy as the upper bound. Only AP-based methods are included.

Algorithm	Standard	PGD		AutoAttack		BPDA
		ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	
Zero-shot (w/o defense)†	74.90	1.20	31.60	0.10	0.10	0.00
+ *DiffPure* (Nie et al., 2022)	71.10	43.00	43.40	42.90	44.20	42.50
+ *DDPM++* (Song et al., 2020)	70.70	66.00	70.00	68.10	70.40	63.50
+ *REAP* (Lee & Kim, 2023)	51.30	48.90	49.90	48.40	50.10	48.50
+ *OSCP* (Lei et al., 2025)	71.60	65.70	69.00	68.30	70.10	66.00
+ *Ours*	**73.10**	**67.30**	**70.80**	**68.90**	**70.90**	**67.30**

4.2.2 Perceptual Quality Evaluation

MANI-Pure produces purified images that are perceptually closest to clean images across different backbones. Since diffusion-based purification is inherently generative, we complement robustness evaluation with perceptual quality metrics, conducted on the CIFAR-10 dataset. Table 4 reports results on SSIM (Wang et al., 2004) (higher is better) and LPIPS (Zhang et al., 2018) (lower is better). On RN50, MANI-Pure achieves an SSIM of **0.9274** and an LPIPS of **0.1136**, both outperforming all baselines. Similar trends are observed with ViT-L/14. Overall, MANI-Pure consistently achieves the highest perceptual similarity, underscoring its ability to defend against adversarial perturbations while preserving image fidelity.

Table 4: To evaluate the quality of the generated images, we compute the SSIM and LPIPS scores between the images purified by different AP methods and the clean images.

Backbone	Metric	Methods				
		Adversarial	DiffPure	REAP	FreqPure	Ours
ViT-L/14	SSIM ↑	0.8204	0.8342	0.8044	0.9172	**0.9270**
	LPIPS↓	0.4403	0.2110	0.2553	0.1214	**0.1133**
RN50	SSIM ↑	0.8180	0.8344	0.8045	0.9176	**0.9274**
	LPIPS↓	0.3907	0.2110	0.2551	0.1217	**0.1136**

4.2.3 Qualitative Visualization

Visualizations confirm that MANI-Pure selectively suppresses adversarial perturbations while preserving semantics. To better validate the effectiveness of adaptive noise injection, we visualize the difference between the injected noise and adversarial noise. Figure 3 clearly shows that adaptive noise aligns much better with adversarial perturbations than uniform noise, especially in high-frequency regions that are most vulnerable to attacks. Quantitatively, KL divergence further confirms this observation: adaptive noise (**0.1628**) is substantially closer to adversarial noise than uniform noise (**0.4988**).

These findings highlight our core advantage—precise suppression of adversarially vulnerable regions while preserving semantic fidelity elsewhere.

We further compare purified samples from DDPM++ and MANI-Pure, together with pixel-wise difference heatmaps relative to clean images. Our method introduces smaller modifications in low-frequency background regions, while applying targeted changes in high-frequency regions most affected by perturbations, which provides direct evidence of MANI-Pure's frequency-adaptive design.

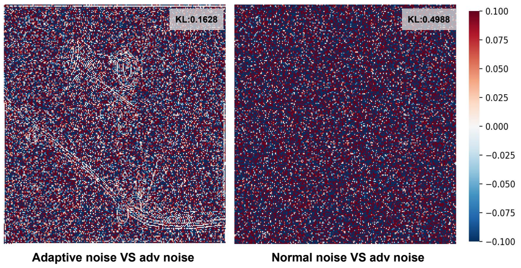


Figure 3: Difference heatmaps between adaptive noise (left) / uniform noise (right) and adversarial noise. Lighter colors indicate smaller differences.

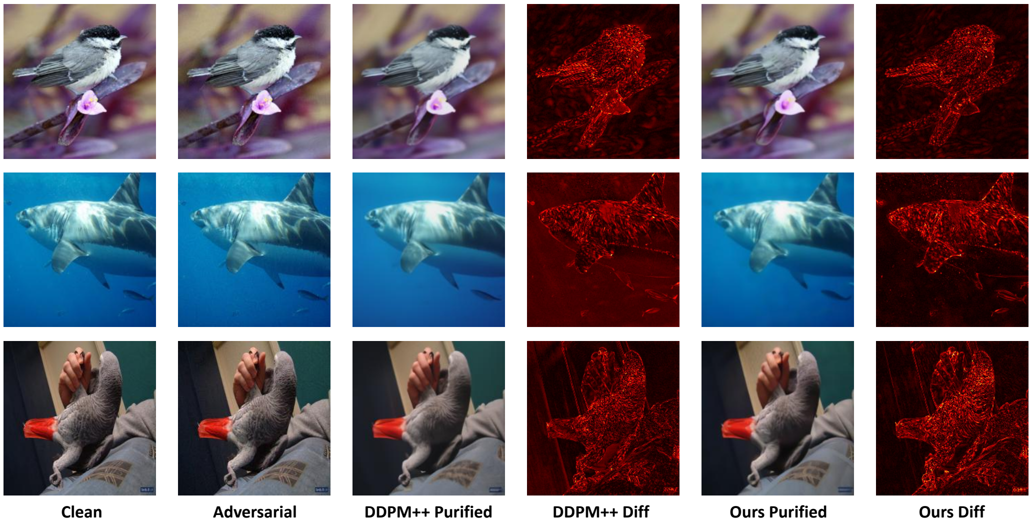


Figure 4: Visualization of purification before and after defense. The figure compares purified results from DDPM++ and MANI-Pure, together with pixel-wise difference heatmaps relative to clean images. **Overall:** MANI-Pure introduces smaller modifications in low-frequency background regions, avoiding unnecessary semantic loss. **Key effect:** it selectively alters high-frequency vulnerable regions, providing direct evidence of its frequency-adaptive design.

4.2.4 Plug-and-Play Compatibility

As a modular noise injection strategy, MANI can be seamlessly combined with various existing DBP methods. Table 5 reports results under ℓ_∞ attacks (results under ℓ_2 are listed in Appendix E.3).

We observe that **MANI consistently improves both clean and robust accuracy** across all tested AP baselines. In particular, REAP benefits the most, with its clean accuracy increased by **4.10%** and robust accuracy under AutoAttack improved by **1.95%**. More importantly, the combination of MANI with FreqPure yields the overall best performance, highlighting the **complementary design philosophy** between the two modules.These results validate MANI as a general and effective plug-in for enhancing diverse purification pipelines.

Table 5: **Plug-and-play validation of the MANI module under ℓ_∞ attacks.** We integrated MANI into various diffusion-based purification frameworks and evaluated them on CIFAR-10. Results are reported both without MANI (**w/o**) and with MANI (**w/**).

Algorithm	**Standard**		**PGD**		**AutoAttack**		**BPDA**	
	w/o	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**
+*DiffPure* (Nie et al., 2022)	86.52	87.72	85.55	86.82	85.35	86.91	84.96	85.43
+*DDPM++* (Song et al., 2020)	86.33	87.30	84.77	86.33	85.74	86.52	86.13	86.33
+*REAP* (Lee & Kim, 2023)	81.45	85.55	79.69	81.45	80.08	82.03	80.86	82.42
+*FreqPure* (Pei et al., 2025b)	91.77	94.14	90.17	91.02	90.82	92.19	87.89	88.67

4.3 Ablation Studies

We conducted ablation experiments on CIFART-10 to better understand the contributions of different design choices in MANI-Pure, primarily involving parameter analysis and module ablation.

Effect of hyperparameters. The MANI module mainly involves two hyperparameters: the weighting factor γ and the number of frequency bands n. As shown in Figure 5, both standard and robust accuracy exhibit a “rise-then-fall” trend as γ increases from 1.0. Specifically, standard accuracy peaks at $\gamma = 1.6$, while $\gamma = 1.8$ achieves a more balanced trade-off between clean and robust performance. A similar trend is observed for n, where $n = 8$ provides the best overall results.

Effect of different modules. To further assess the contribution of each component, we conduct ablation studies on MANI and FreqPure. As shown in Table 6, both modules individually enhance the baseline performance. When combined, they yield substantially larger improvements than using either module alone, achieving gains of 7.62% in clean accuracy and 5.47% in robust accuracy. These results highlight the orthogonal benefits of MANI and FreqPure, and their strong complementarity.

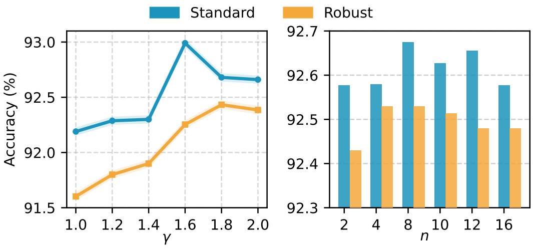


Figure 5: Standard accuracy and robust accuracy under different ratio factor γ (left) and under different number of frequency band n (right).

Table 6: Standard and robust accuracy for different block combinations. ✓ and ✗ indicate use or non-use of the module.

MANI	**FreqPure**	**Standard**	**Robust**
✗	✗	86.52	85.55
✓	✗	87.30	86.33
✗	✓	91.77	90.17
✓	✓	**94.14**	**91.02**

5 Conclusion

This work systematically analyzes the distribution of adversarial perturbations in the frequency domain and shows that existing uniform noise injection strategies may disrupt the semantic structure of clean images. To address this issue, we propose **MANI-Pure**, a diffusion-based purification framework that integrates magnitude-adaptive noise injection to emphasize vulnerable frequency bands and frequency purification to protect semantic structures. Through extensive experiments on two benchmark datasets under multiple attacks, MANI-Pure effectively suppresses adversarial noise while preserving semantic content, achieving a favorable balance between clean and robust accuracy. Moreover, the plug-and-play design of MANI highlights its compatibility with diverse purification pipelines, further broadening its applicability.

REFERENCES

Sajjad Amini, Mohammadreza Teymoorianfard, Shiqing Ma, and Amir Houmansadr. Meansparse: Post-training robustness enhancement through mean-centered feature sparsification. *arXiv preprint arXiv:2406.05927*, 2024.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018.

Mingyuan Bai, Wei Huang, Tenghui Li, Andong Wang, Junbin Gao, Cesar F Caiafa, and Qibin Zhao. Diffusion models demand contrastive guidance for adversarial purification to advance. In *ICML*, 2024.

Gerda Bortsova, Cristina González-Gonzalo, Suzanne C Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram Van Ginneken, Josien PW Pluim, Mitko Veta, et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141, 2021.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Jia Fu, Yongtao Wu, Yihang Chen, Kunyu Peng, Xiao Zhang, Volkan Cevher, Sepideh Pashami, and Anders Holst. Diffcap: Diffusion-based cumulative adversarial purification for vision language models. *arXiv preprint arXiv:2506.03933*, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *ICLR*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *ICCV*, 2023.

Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *CVPR*, 2025.

Guang Lin, Chao Li, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Adversarial training on purification (atop): Advancing both robustness and generalization. In *ICLR*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *ICML*, 2022.

Gaozheng Pei, Shaojie Lyu, Gong Chen, Ke Ma, Qianqian Xu, Yingfei Sun, and Qingming Huang. Divide and conquer: Heterogeneous noise integration for diffusion-based adversarial purification. In *CVPR*, 2025a.

Gaozheng Pei, Ke Ma, Yingfei Sun, Qianqian Xu, and Qingming Huang. Diffusion-based adversarial purification from the perspective of the frequency domain. In *ICML*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *ICLR*, 2018.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*, 2024.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

Yuhao Sun, Jiacheng Zhang, Zesheng Ye, Chaowei Xiao, and Feng Liu. Sample-specific noise injection for diffusion-based adversarial purification. In *ICML*, 2025.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Keda Tao, Jinjin Gu, Yulun Zhang, Xiucheng Wang, and Nan Cheng. Overcoming false illusions in real-world face restoration with multi-modal guided diffusion model. *arXiv preprint arXiv:2410.04161*, 2024.

Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *ICML*, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Juanjuan Weng, Zhiming Luo, Zhun Zhong, Dazhen Lin, and Shaozi Li. Exploring non-target knowledge for improving ensemble universal adversarial attacks. In *AAAI*, 2023.

Peng Ye, Yuanfang Chen, Sihang Ma, Feng Xue, Noel Crespi, Xiaohan Chen, and Xing Fang. Security in transformer visual trackers: A case study on the adversarial robustness of two models. *Sensors (Basel, Switzerland)*, 24(14):4761, 2024.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Kejia Zhang, Juanjuan Weng, Shaozi Li, and Zhiming Luo. Towards adversarial robustness via debiased high-confidence logit alignment. In *ICCV*, 2025a.

Mingkun Zhang, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Clipure: Purification in latent space via clip for adversarially robust zero-shot classification. In *ICLR*, 2025b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.