AMRM-Pure: Semantic-Preserving Adversarial Purification

Zhihao Dou[1,2,*] **Zhiqiang Gao**[1,*,†] **Dongfei Cui**[3] **Weida Wang**[4] **Qinjian Zhao**[1]
Dinggen Zhang[1] **Jun Yan**[5,†] **Zeke Xie**[6] **Shufei Zhang**[7,†]
[1] Wenzhou-Kean University; [2] Case Western Reserve University; [3] Northeast Electric Power University;
[4] Fudan University; [5] Shanghai Ocean University; [6] HKUST (Guangzhou); [7] Shanghai AI Lab

[*]Equal contribution. [†]Corresponding authors

Abstract

Adversarial purification is a defense technique that employs generative models to remove adversarial perturbations. Current methods often rely on powerful generators, typically diffusion models, and focus on reducing the gap between adversarial and clean samples in the feature space, while overlooking semantic correlation within a single sample. To address this issue, we explore adversarial purification from the perspective of preserving semantic relationships among image patches. We employ an **A**ttentive **M**ask **R**econstruction **M**odel (**AMRM**), which shows superior performance. Our theoretical and experimental analysis reveals that AMRM is highly sensitive to adversarial noise, as such noise significantly distorts patch relationships. Based on this observation, we propose AMRM-Pure, a purification framework that denoises adversarial inputs by preserving patch-level semantics, and formulate this process as a tractable optimization problem with respect to the input. To further enhance robustness, we finetune AMRM-Pure with classification loss to strengthen semantic consistency. We apply our insight to two AMRM architectures, including Mask Autoencoder (MAE) and MaskDiT. Extensive experiments confirm the effectiveness of our method, establishing new state-of-the-art performance across multiple benchmarks.

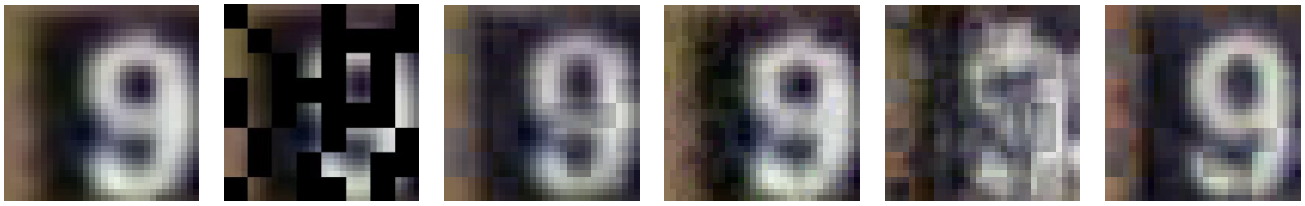
Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Deep Neural Networks (DNNs) are vulnerable to adversarial examples (Carlini and Wagner, 2017; Song et al., 2018; Fischer et al., 2017; Lyu et al., 2015), which are imperceptible to humans. However, these inputs with the malicious perturbations can cause DNNs to make erroneous predictions. Adversarial training (Madry et al., 2018; Zhang et al., 2019) is the state-of-the-art method for defending against adversarial attacks. However, the trade-off between generalization and robustness remains a concern (Zhang et al., 2019), especially against unseen adversarial examples. Furthermore, adversarial training incurs significantly higher computational costs compared to standard training.

Alternatively, another notable defense strategy is adversarial purification, which attracts widespread attention. Adversarial purification can be broadly classified into two categories, including purification with generative models (Yoon et al., 2021; Nie et al., 2022; Lin et al., 2024; Bai et al., 2024; Zhang et al., 2024) and adaptation-based purification (Shi et al., 2021). Generative model-based approaches are the most widely used methods in adversarial purification, typically harnessing the powerful capabilities of generative models (e.g., diffusion) to transform the distribution of adversarial examples to that of clean samples (Nie et al., 2022). Future efforts will aim to further enhance the denoising capabilities of the purification model through various approaches. These include leveraging contrastive guidance to steer diffusion models (Bai et al., 2024), integrating classifier confidence guidance into the denoising process (Zhang et al., 2024), and fine-tuning the purification model with adversarial loss for robust optimization (Lin et al., 2024).

The aforementioned methods primarily focus on aligning adversarial examples with the semantic distribution of clean samples, but neglect the semantic relationships among different patches within a sample. To fill this gap, we use an attentive mask reconstruction model (AMRM) to investigate how adversarial perturbations



(a) (b) (c) (d) (e) (f)

Figure 1: (a) Original image, (b) Masked image, (c) Clean image reconstruction from MAE, (d) Adversarial example under AutoAttack, (e) Reconstruction of adversarial example under AutoAttack from MAE, (f) Reconstruction of the denoised image under AutoAttack from MAE (denoised by our AMRM-Pure$_{\text{MAE}}$).

distort the semantic relationships among image patches. AMRM divides an image into multiple patches, masks a subset of them, and reconstructs the masked patches by using a self-attention (Vaswani et al., 2017) mechanism to explicitly capture dependencies among visible patches, such as the Mask Autoencoder (MAE) (He et al., 2022) or MaskDiT (Zheng et al., 2024). To design a robust purification method, we first identify an intriguing phenomenon of the simple MAE so that an easier analysis. Specifically, in the case of adversarial examples subjected to tiny, visually imperceptible perturbations, the reconstruction performance of MAE is severely compromised, dealing a devastating blow. As a typical example shown in Figure 1a and1d, although the clean example and adversarial example appear very similar, MAE's reconstruction outputs in Figures 1c and 1e exhibit significant differences. The reconstruction of perturbed data, as illustrated in Figure 1e, still displays poor quality. These findings suggest that the reconstruction capability of MAE is highly sensitive to adversarial perturbations, although these perturbations are visually imperceptible. Motivated by this observation, we consider preserving the semantic relationships among image patches as a novel mechanism for adversarial purification, a direction that has not been fully explored in existing works.

Based on such a research motivation, our proposed study aims to fill the gap. In this paper, through a series of analyses, we conjecture that this phenomenon results from adversarial perturbations that could easily distort semantic relations within patches, i.e., causing variation in the attention matrix (AMV), which leads to degraded image reconstruction quality of MAE. Through rigorous theoretical derivations and empirical experiments, we provide compelling evidence of the sensitivity of MAE to this AMV. Concretely, as the patch attention matrix essentially reflects how different semantic patches may be related to the masked patch, altering the attention matrix means a semantic change when the masked patch is reconstructed. Figures 3b and 3c show that the reconstruction of the target patch in the red square depends on the similar patches in the clean image, while for adversarial images, high importance is assigned to distant and dissimilar patches. This suggests that the adversarial perturbations alter the semantic relations among patches. Moreover, our findings reveal that the reconstruction loss of adversarial examples is lower-bounded by the sum of the loss for clean examples and the AMV. Meanwhile, we reveal that the sensitivity of AMV is transferable to a diffusion-based AMRM, e.g., MaskDiT, as shown in empirical analysis in Figure 6. Drawing inspiration from this finding, we propose a novel AMRM-Pure method which purifies adversarial perturbations by minimizing AMV, ultimately resulting in a inter-patch semantic preserving framework.

AMRM-Pure leverages the inherent sensitivity of AMV to adversarial noise, thereby achieving enhanced robustness. To realize AMRM-Pure, we introduce two variants: AMRM-Pure$_{\text{MAE}}$ based on MAE (He et al., 2022) and AMRM-Pure$_{\text{MaskDiT}}$ based on MaskDiT (Zheng et al., 2024). Furthermore, based on the insight of the previous work (Lin et al., 2024; Zhang et al., 2024), we propose a Robust AMRM-Pure based on MAE (RAMRM-Pure$_{\text{MaskDiT}}$) and Robust MRM-Pure based on MaskDiT (RAMRM-Pure$_{\text{MaskDiT}}$) that leverages classification loss to fine-tune the purification model, significantly improving its inter-patch semantic preservation capability. We have extensively evaluated our method by comparing the important adversarial training and adversarial purification methods on various challenging adaptive attack benchmarks. Our method achieves state-of-the-art (SOTA) performance on four datasets, e.g., CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and ImageNet (Deng et al., 2009).

In summary, our main contributions are as follows:

1) We investigate the susceptibility of attention-based MRM to noise interference from both theoretical and empirical perspectives and disclose that the noise induces the deviation of semantic relations among patches, resulting in a degradation of the quality of the reconstruction. On the basis of our findings, we devise a novel and efficient purification technique, called AMRM-Pure, which is theoretically proven by rigorous analysis.

2) By successfully applying our approach to MAE and MaskDiT, we introduce AMRM-Pure$_{\text{MAE}}$ and AMRM-Pure$_{\text{MaskDiT}}$. Meanwhile, we further propose RAMRM-Pure$_{\text{MAE}}$ and RAMRM-Pure$_{\text{MaskDiT}}$, which incorporates classification loss to fine-tune the purification model, significantly improving standard and robust accuracy.

3) Extensive experiments have been conducted to validate the effectiveness of our MRM-Pure on various benchmarks, showing that our approach consistently achieves favorable outcomes after denoising processes.

2 PRELIMINARIES AND RELATED WORK

2.1 Adversarial Training

Adversarial training (AT) is a technique that enhances the robustness of a neural network by augmenting training samples with additional adversarial examples (Goodfellow et al., 2015; Kurakin et al., 2017; Tramèr et al., 2018; Zhang et al., 2019; Dou et al., 2024). Since AT typically involves a high computational cost, some studies (Wong et al., 2020; Liu et al., 2021; Vivek and Babu, 2020) have focused on exploring ways to accelerate the training process by a one-step training strategy. In addition, the diffusion model has been employed for extensive data augmentation for adversarial training in many proposals (Wang et al., 2023; Gowal et al., 2021; Sehwag et al., 2021), which enlarged the original dataset and enhanced the robust generalization.

2.2 Adversarial Purification

Generative models have shown great promise in purifying adversarial examples, drawing significant attention in robustness research. The early milestone study (Samangouei et al., 2018) introduced Defense-GAN, using GANs for purification. Song et al. (Song et al., 2017) proposed the PixelDefense method, which employs the autoregressive models to mitigate the perturbations. Score-based generative models have also been applied for defense (Yoon et al., 2021). Leveraging diffusion models, DiffPure (Nie et al., 2022) uses Stochastic Differential Equation (SDE) diffusion (Song et al., 2021) for the denoising procedure, achieving robustness. Recent works (Li et al., 2025; Liu et al., 2025) further improve robustness by fine-tuning diffusion models. Lin et al. (Lin et al., 2024) proposed a hybrid approach combining adversarial training with purification. It is significant to reconstruct the data without semantic information changes. IDC (Mei et al., 2025) redesigns diffusion models from generating high-quality images to producing distinguishable label images, proposing

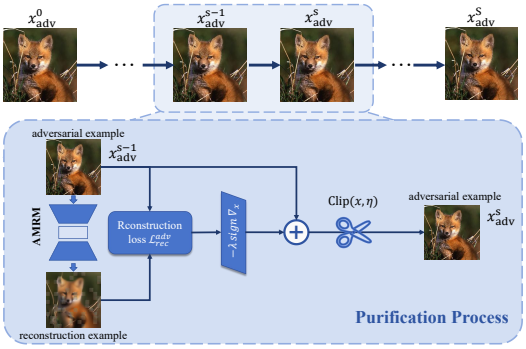


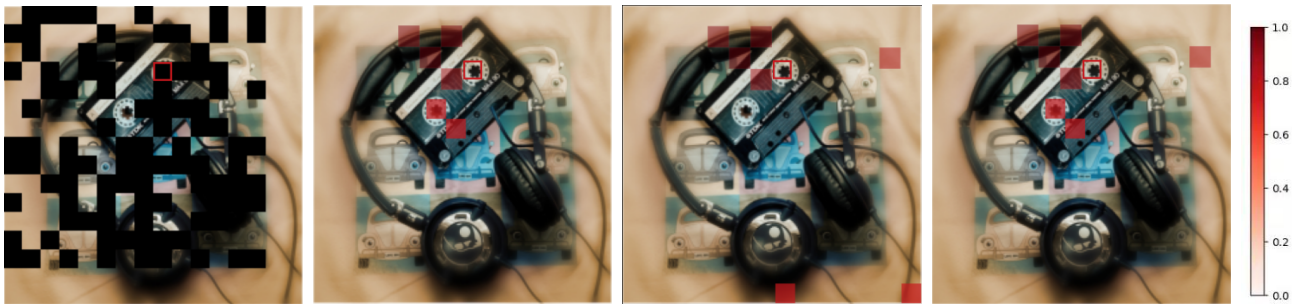
Figure 2: Overview of the proposed AMRM-Pure.

an efficient image-to-image diffusion classifier that significantly reduces computational cost while improving adversarial robustness. Thus, Bai et al. (Bai et al., 2024) introduced contrastive guidance in diffusion models to enhance purification while preserving semantics. The adversarial purification method can be combined with other machine learning paradigms. For example, the framework of Self-supervised Online Adversarial Purification (SOAP) (Shi et al., 2021) achieves notable results by integrating self-supervised tasks during training, further boosting robustness.

There are several purification based on MAE (Zhou et al., 2023; Wu et al., 2022; You et al., 2023). DIR (Zhou et al., 2023) introduces a joint training framework of classifier and MAE under adversarial training, where the MAE restores robust features from unmasked patches to mitigate adversarial noise. Following the denoising autoencoder paradigm, DMAE (Wu et al., 2022) and NIM-MAE (You et al., 2023) incorporate Gaussian noise into masked image modeling, which is eliminated through the encoding–decoding process. Specifically, DMAE focuses on achieving robust pre-training against Gaussian noise to enhance generalization and robustness, though it shows limited effectiveness against adversarial perturbations. NIM-MAE leverages pre-trained models to remove adversarial noise, yet its robustness performance still leaves room for improvement. Unlike previous methods, ARM-Pure leverages the sensitivity of semantic patch relationships to adversarial perturbations and employs optimization-based denoising to reduce AMV, effectively minimizing semantic variations in adversarial examples. This novel perspective has not been studied in previous research.

2.3 Preliminary of Masked Autoencoder (MAE)

MAE (He et al., 2022) is briefly introduced within the context of adversarial robustness in this subsection. A clean input sample x, drawn from the dataset X, is partitioned into n patch vectors of dimension d, forming $\bar{x} \in \mathbb{R}^{n \times d}$. The matrix $\bar{x}$ can be randomly divided into



(a) (b) (c) (d)

Figure 3: The first column, Figure (a) represents the Mask Matrix. The second column, Figure (b) illustrates the Attention Weights for clean samples. The third column, Figure (c) depicts the Attention Weights for adversarial examples. The fourth column, Figure (d) showcases the Attention Weights for denoised samples (by our AMRM-Pure$_{\text{MAE}}$). Patches with a deeper red color mean the elements with more attention. The data is sampled from the ImageNet dataset (Deng et al., 2009).

$m = (1-\rho)n$ masked patch vectors and $(n-m)$ visible patch vectors, where ρ is the mask ratio. MAE uses an encoder-decoder architecture. The encoder, $f(\cdot)$, produces $\mathbf{V}^{enc} \in \mathbb{R}^{m \times d_e}$, where $\mathbf{V}^{enc} = f(x_1)$, and x_1 is the visible portion of input x. Here, d_e is the dimension of each patch feature in $\mathbf{V}^{enc}$. The decoder, $g(\cdot)$, maps $\mathbf{V}^{enc}$ back to pixel space, producing $\mathbf{V}^{dec} \in \mathbb{R}^{(n-m)\times d}$, i.e., $g(\mathbf{V}^{enc}) = \mathbf{V}^{dec}$, which reconstructs masked patches x_2. Reconstruction quality is measured using Mean Squared Error (MSE) loss as follows:

$$\mathcal{L}_{\text{rec}}(x_1) = \frac{1}{N(n-m)} \sum_{i=1}^{N} ||g(f(x_{1,i})) - x_{2,i}||^2, \quad (1)$$

where N represents the sample number in X.

The MAE structure consists of multiple self-attention layers, where attention captures semantic relationships between input patches. At the t-th layer, the input features are $\mathbf{Z}^t \in \mathbb{R}^{n_t \times d_t}$, with n_t patches and d_t-dimensional patch features. Weight matrices $\mathbf{W}_Q^t$, $\mathbf{W}_K^t$, and $\mathbf{W}_V^t$ generate the query $\mathbf{Q}^t$, key $\mathbf{K}^t$, and value $\mathbf{V}^t$ matrices, all in $\mathbb{R}^{n_t \times d_t}$.

The self-attention matrix $\mathbf{A}^t$ is computed as:

$$\mathbf{A}^t = \text{softmax}\left(\frac{\mathbf{Q}^t(\mathbf{K}^t)^T}{\sqrt{d_t}}\right),$$

quantifying similarities between $\mathbf{Q}^t$ and $\mathbf{K}^t$. The j-th output patch feature e_j is a weighted sum of value vectors:

$$e_j = \sum_{i=1}^{n} a_{ji}^t v_i^t, \quad a_{ji}^t = \frac{q_{ji}^t k_{ji}^t}{\sum_{o=1}^{n} q_{jo}^t},$$

where a_{ji}^t indicates how much v_i^t contributes to e_j.

3 THEORETICAL ANALYSIS

In this section, we theoretically analyze how adversarial perturbations affect semantic relationships among patches. We take MAE as a representative AMRM architecture due to its simplicity and typical design, and investigate the link between AMV variation and decoder reconstruction loss, both theoretically and empirically. Given the structural similarity, we further extend this analysis to MaskDiT (Sec. 3.4), and our results confirm that the theory also holds, demonstrating strong transferability.

3.1 Adversarial Perturbation Induces Attention Matrix Variation in MAE

Theorem 3.1. *Let $\delta_t = \boldsymbol{Z}_{adv}^t - \boldsymbol{Z}^t$ denotes the latent feature shift caused by the adversarial perturbation at layer t in MAE. With a set $\{\omega_i\}_{i=0}^k$ and kernel coefficient $\omega_i \in \mathcal{N}(0, \mathbf{I}_d)$, it holds that:*

$$||\mathbf{A}_{adv}^t - \mathbf{A}^t||_2 \geq \gamma \left\| \left[\left(\mathbf{Y} - \mathbf{B}\mathbf{Q}^t\right)^\top \mathbf{W}_Q^t + \left(\mathbf{Y} - \mathbf{B}\mathbf{K}^t\right)^\top \mathbf{W}_K^t \right] \delta_t \right\|_2,$$

$$\mathbf{B} = \sum_{i=0}^{k} \exp\left(\omega_i^\top (\mathbf{Q}^t + \mathbf{K}^t)\right), \quad \mathbf{Y} = \sum_{i=0}^{k} \exp\left(\omega_i^\top (\mathbf{Q}^t + \mathbf{K}^t)\right) \omega_i,$$

$$\gamma = \frac{\exp\left(-\frac{\|\mathbf{Q}^t\|^2 + \|\mathbf{K}^t\|^2}{2}\right)}{m}.$$

Proof. The proof can be seen in Appendix I.2. □

Theorem 3.1 suggests that even minor shifts in the latent features (δ_t) may have the ability to cause disproportionately large changes in AMV, especially due to the high dimensionality d_t of internal projection matrices. Notably, in MAE/AMRM, where $d_t \gg$ input dimension, the sensitivity is further amplified. This analysis reveals the intrinsic vulnerability of MAE's attention mechanism under adversarial conditions, offering a theoretical foundation for the empirical trends shown in Figure 4 (a-c) of Section 3.3.

3.2 Impact of Decoder Attention Shifts on Adversarial Reconstruction in MAE

To deepen the understanding of how attention pattern distortions affect output quality in MAE, we present a theoretical lower bound on the reconstruction loss under adversarial conditions. This analysis extends the discussion of AMV sensitivity in Section 3.1 and reveals how attention shifts in the decoder layer affect reconstruction loss. First, let $\mathcal{L}_{\text{rec}}^{adv}$ denote the average reconstruction loss for adversarial examples, corresponding to the reconstruction loss $\mathcal{L}_{\text{rec}}$ for clean samples as Eq. (1) in Appendix 2.3.

Theorem 3.2. *Let $\mathbf{A}_{i,t}^{\text{dec}}$ denote the attention matrix at the t-th layer of the MAE decoder for the i-th sample in the dataset, and let $\mathbf{A}_{\text{adv},i,t}^{\text{dec}}$ denote the corresponding attention matrix for the adversarial examples. With ratio constants C_A and H, it holds that:*

$$\mathcal{L}_{rec}^{\text{adv}} \geq \frac{1}{2}\mathcal{L}_{rec} + \frac{1}{2NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left[HC_A\left\|\mathbf{A}_{\text{adv},i,t}^{\text{dec}} - \mathbf{A}_{i,t}^{\text{dec}}\right\|^2 - c_{rec}\right]$$

c_{rec} is the reconstruction bias, which symbolizes the disparity between the output of MAE and the original, unmasked image. The definition of c_{rec} can be found in Appendix I.1.

Proof. The proof can be seen in the Appendix I.3. □

It shows that the lower bound of the reconstruction loss for adversarial data can be decomposed into three components: the average reconstruction loss for clean data $\mathcal{L}_{\text{rec}}$, the average attention matrix variation for the MAE decoder at each layer $\frac{1}{2NT}\sum_{t=1}^{T}\sum_{i=1}^{N}||(A_{adv_{i,t}}^{dec} - A_{i,t}^{dec})||^2$, and constant terms. Theorem 3.2 demonstrates that adversarial distortions in AMV of decoder lead to increases in reconstruction loss $\mathcal{L}_{\text{rec}}$, and Figure 4 provides further evidence of this phenomenon. Notably, the reconstruction loss is shown to be consistent with the degree of AMV, confirming a strong correlation between inter-patch semantic relationships and output degradation. This theorem builds a theoretical foundation of robust MAE-based purification.

3.3 Empirical Validation

This section empirically validates our theoretical analysis. We first assess the impact of MAE reconstruction on adversarial perturbations, then analyze the correlation between attention matrix variation and reconstruction loss. Finally, we show that adversarial perturbations alter semantic relationships within MAE patches and confirm the sensitivity of AMV to such perturbations.

Perturbation Leads to Degraded Reconstruction Quality. To empirically investigate how perturbation influences reconstruction, an image is randomly selected from the SVHN dataset. We then compare the reconstruction results of the clean data and the adversarial example as shown in Figure 1. The adversarial example generated by the AutoAttack procedure (Croce and Hein, 2020b) in Figure 1d looks almost identical to its clean counterpart in Figure 1a visually. However, its reconstruction (Figure 1e) is significantly different from that of the clean sample (Figure 1c). Likewise, its reconstruction result also shows significant differences with a reconstruction of its clean sample (Figure 1c). These phenomena emphasize the substantial impact of adversarial perturbations on the overall outcome of the reconstruction.

Visualization of Attention Matrix Variation. To check how the semantic relationship between different patches changes under perturbation, we show the degree of importance of visible patches for reconstructing the masked patch in Figure 3. We select an image from ImageNet (Deng et al., 2009) and randomly generate a mask matrix. A specific masked patch is considered as the target patch (marked red in a box) in Figure 3a. Then, the visible patches are fed into the MAE to perform the reconstruction. The degree of importance of visible patches for reconstructing the target patch, which is determined by the corresponding values of the attention matrix, is illustrated in Figure 3b (e.g., the last layer attention of the MAE decoder). Meanwhile, we also display the corresponding visualizations of the adversarial example and the denoised example for the same mask matrix and target patch position.

In generating attention weights for the target patch (red box), the approach involves using the patch itself as the query vector and the remaining patches as key vectors. Through self-attention, the attention weights are determined. Higher weights indicate a stronger semantic similarity between the target patch and the patch itself. Figures 3b, 3c, and 3d show the importance degree of visible patches in the clean sample, adversarial example, and denoised example, respectively. Patches that are closer to red are more significant. As demonstrated in Figure 3b, when a hole in the tape is used as the target patch, the visible patch of another similar hole and the surrounding patches serve as the most important basis for reconstructing the target patch. However, as illustrated in Figure 3c, some distant and irrelevant visible patches with large color and shape differences are taken or focused to reconstruct the adversarial perturbed target patch. This indicates that the adversarial perturbation leads MAE to erroneous attention. More examples can be seen in Appendix G.

Analysis of AMV Sensitivity and Consistency with Reconstruction Loss. To validate Theorem 3.2 and the effect of AMV on the reconstruction qual-

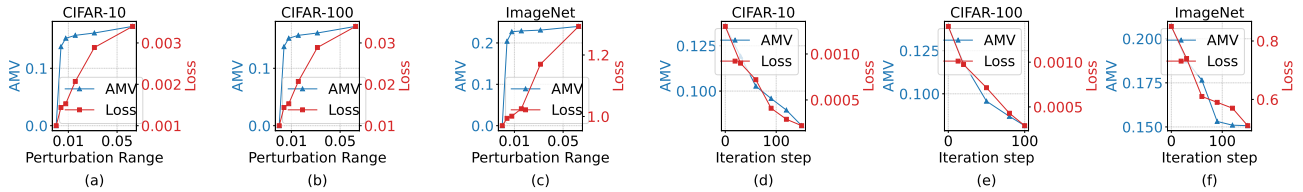


Figure 4: Trends of MAE reconstruction loss and attention matrix variation under AutoAttack and during purification across multiple datasets. (a–c): Under AutoAttack on CIFAR-10, CIFAR-100, and ImageNet with different attack budget. (d–f): During the purification process with AMRM-Pure$_{\text{MAE}}$ on CIFAR-10, CIFAR-100, and ImageNet.

ity, we visualize the changes of reconstruction loss and AMV values with respect to the intensity of adversarial noise in Figure 4. We evaluate reconstruction loss and average AMV $\frac{1}{N}\sum_{i=1}^{N}||\mathbf{A}_{adv,i}^{dec}-\mathbf{A}_{i}^{dec}||_2$ on 150 images from CIFAR-10, CIFAR-100, and ImageNet with a 0.5 mask ratio, using AutoAttack to generate adversarial examples. As shown in Figure 4 (a–c), AMV is highly sensitive, rising sharply even under small perturbations (e.g., $\delta = 0.01$), and then steadily increasing until reaching its softmax-bounded upper limit. Reconstruction loss and AMV follow the same trend, aligning with our theoretical analysis in Theorem 3.2 of their relationship and AMV's sensitivity to perturbations.

3.4 A General AMRM Framework

The aforementioned analysis based on simple MAE reveals *a key vulnerability in attentive masked image modeling: inter-patch semantic information (captured by AMV) is sensitive to noise.* To show the universality of this observation, we further illustrate this sensitivity of AMV on a more powerful generative model, such as MaskDiT (Zheng et al., 2024), which is a powerful diffusion-based AMRM that preserves its masked autoencoder architecture while integrating forward and reverse diffusion steps for high-quality image generation. Given their architectural alignment, the MaskDiT indeed agrees with the analysis presented in Theorem 3.1 in Sec 3.1. As shown in Figure 7 of Appendix E, empirical comparisons show that MaskDiT exhibits consistent patterns with MAE under adversarial settings. Specifically, the AMV of MaskDiT remains highly sensitive to noise, so that its reconstruction loss and AMV consistently maintain alignment. As such, MaskDiT continues to satisfy Theorem 3.1 and 3.2. Motivated by this phenomenon, we propose AMRM-Pure$_{\text{MaskDiT}}$, which leverages MaskDiT's generative capacity while retaining MAE's inter-patch semantic sensitivity, further improving purification performance by optimizing its reconstruction loss (Eq. (3)).

4 METHOD

4.1 Adversarial Purification with AMRM

As discussed in Section 3, the attention mechanism of AMRM is highly sensitive to adversarial perturbations, which distort inter-patch semantic relationships and degrade reconstruction quality. Based on this sensitivity, we propose a purification scheme, AMRM-Pure, which formulates denoising as an optimization problem that minimizes semantic variations.

We denote the clean data as x, and the adversarial example as x_{adv}. In the context of denoising, the objective is to induce a modification Δ on x_{adv} such that the attention matrix of the denoised examples, $atten(x_{adv}+\Delta)$, closely aligns with the attention matrix of clean data samples $atten(x)$. Therefore, the learning objective of denoising can be formed as an Attention Matrix Variation Minimization problem, which is denoted as:

$$\begin{aligned} &\min_{\Delta} \mathcal{L}(\Delta) = ||\text{atten}(x_{adv}+\Delta) - \text{atten}(x)||_2, \\ &\text{s.t.} \quad ||\Delta||_\infty \le C_e, \end{aligned} \tag{2}$$

where C_e is a small constant.

In addressing the AMV Minimization problem, the denoising process strives to mitigate adversarial perturbations on x_{adv}. Its goal is to achieve a consistent attention matrix between the denoised images and the clean images of MAE. This alignment ensures that the patch relationship of the denoised image closely approximates that of clean samples, thereby minimizing the impact of adversarial perturbations in the denoising outcome.

Since clean attention $atten(x)$ is unavailable during inference, directly minimizing AMV Eq. (2) is intractable. Instead, inspired by Theorem 3.2, we minimize the AMRM reconstruction loss as a tractable surrogate. Reconstruction loss not only enables efficient optimization without requiring clean attention (tractability), but also can reduce AMV caused by perturbations (owing to the consistent trend presented in Theorem 3.2),

thereby aligning with the inter-patch semantic structure of the clean input. As such, our approach instead minimizes the AMRM reconstruction loss for adversarial purification.

$$min_{\Delta}\mathcal{L}_{\text{rec}}(x_{adv}+\Delta) \iff min_{\Delta}\mathcal{L}(\Delta), \quad s.t. \quad ||\Delta||_{\infty} \leq C_e. \tag{3}$$

To address this problem, we employ the standard Projected Gradient Descent (PGD) method (Madry et al., 2018). In this approach, the modifications are iteratively added to adversarial examples, and the total number of iterations is denoted as S. At the s-th iteration, the denoising process is denoted as:

$$\begin{aligned} x_{adv}^{s} &= Clip(x_{adv}^{s-1} - \lambda \cdot \Delta_s, \eta), \\ \Delta_s &= \text{sign}(\nabla_x L_{\text{rec}}(x_{adv}^{s-1})). \end{aligned} \tag{4}$$

Table 1: Clean and robust accuracy (%) on CIFAR-10 obtained by different purification methods. WideResNet is commonly abbreviated as WRN.

Method	Classifier	Std Acc	Robust Acc	
			ℓ_∞	ℓ_2
Shi et al. (Shi et al., 2021)	WRN-28-10	91.89	4.56	7.25
Yoon et al. (Yoon et al., 2021)	WRN-70-16	87.93	37.65	57.81
Zhang et al. (Zhang et al., 2023)	WRN-70-16	**93.16**	22.07	35.74
Diffpure (Nie et al., 2022)	WRN-70-16	92.50	42.20	60.80
COUP (Zhang et al., 2024)	WRN-28-10	90.33	41.72	57.25
ADBM (Li et al., 2025)	WRN-70-16	91.90	47.70	63.30
$\text{ADDT}_{w/\text{Diffpure}}$ (Liu et al., 2025)	WRN-28-10	89.94	55.76	-
$\text{AMRM-Pure}_{\text{MAE}}$	WRN-28-10	88.57	40.53	53.50
$\text{AMRM-Pure}_{\text{MaskDiT}}$	WRN-28-10	92.03	50.57	64.53
$\text{RAMRM-Pure}_{\text{MAE}}$	WRN-28-10	90.09	45.15	60.72
$\text{RAMRM-Pure}_{\text{MaskDiT}}$	WRN-28-10	93.11	**62.13**	**73.57**

Here $\mathcal{L}_{\text{rec}}$ signifies the AMRM reconstruction loss defined in Eq. (1), with λ representing the step size and η as the clipping threshold. The overall modification Δ is composed of individual iteration modification Δ_s for $s \in [1, S]$. The purpose of Δ is to guide the attention distribution of adversarial examples $atten(x_{adv})$ towards the clean sample distribution $atten(x)$. The denoising algorithm and pipeline of our AMRM-Pure can be seen in Algorithm 1 (Supplement F) and the whole process is in Figure 2. For AMRM-Pure, when the mechanism is applied to the MAE framework, we denote it as $\text{AMRM-Pure}_{\text{MAE}}$; when applied to the MaskDiT framework, we denote it as $\text{AMRM-Pure}_{\text{MaskDiT}}$. To empirically validate the theory of AMRM-Pure, we also plot the trends of the MAE reconstruction loss and AMV with increasing denoising iterations in Figure 4 (d-f) and Fig 7 (b) for MaskDiT. We randomly selected 100 adversarial examples from each dataset, perturbed using AutoAttack. For CIFAR10 and CIFAR100, the perturbation magnitude is set to $\ell_\infty = \frac{8}{255}$, while for ImageNet, it is set to $\ell_\infty = \frac{4}{255}$. As observed, both the AMV and loss exhibit a similar downward trend as the number of purification iterations increases. This supports the validity of our theory.

Furthermore, we provide strict convergence analysis within the Appendix H.

4.2 Robust Purification Model

As the study of AToP (Lin et al., 2024) shows, further fine-tuning a purification model using classification loss can enhance its robustness against both seen and unseen attacks. Following this insight, we propose a two-stage fine-tuning method to develop Robust AMRM-Pure_{MAE} (RAMRM-Pure_{MAE}) and Robust AMRM-$\text{Pure}_{\text{MaskDiT}}$ (RAMRM-$\text{Pure}_{\text{MaskDiT}}$) variants to enhance the semantic relationship-preserving capabilities. For more details about our method, refer to Appendix D.1.

Figure 6a (see Appendix D.1) shows a quantitative analysis of enhanced semantic relationships using AMV as an evaluation metric. We randomly select 100 images from the CIFAR-10 dataset and examined the AMV of AMRM-Pure_{MAE} and RAMRM-Pure_{MAE} under the AutoAttack with an attack budget of $\frac{8}{255}$ across different purification iterations. Specifically, the initial AMV (without purification) of RAMRM-Pure_{MAE} is higher than that of AMRM-Pure_{MAE}. This suggests that Robust MAE is more sensitive to interpatch semantic information changes caused by adversarial attacks. However, with the progression of purification iterations, the AMV of RAMRM-Pure_{MAE} decreases significantly, highlighting its superior capability in preserving semantic integrity compared to AMRM-Pure_{MAE}. Furthermore, as shown in Figure 6b (Appendix D.1), both AMRM-$\text{Pure}_{\text{MaskDiT}}$ and RAMRM-$\text{Pure}_{\text{MaskDiT}}$ exhibit similar trends.

5 EXPERIMENT

5.1 Experimental Setting

Datasets and Classifier. In this section, we validate the robustness of our purification method, AMRM-Pure, on four benchmark datasets, including CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and ImageNet (Deng et al., 2009). We use WideResNet-28-10 (Zagoruyko and Komodakis, 2016) as the main classifier for CIFAR-10, CIFAR-100, and SVHN, and ResNet-101 (He et al., 2016) as the main classifier for ImageNet.

Adversarial Attacks. Several studies (Chen et al., 2024; Li et al., 2025; Liu et al., 2025) show that the AutoAttack method (Croce and Hein, 2020b) tends to overestimate the robustness of diffusion models, primarily due to the presence of gradient obfuscation, which prevents the attack from effectively exploiting the true vulnerabilities of the model. To address this

Table 2: Clean and robust accuracy (%) on CIFAR-100 obtained by different purification methods. The experiment are implemented on WideResNet-28-10.

Method	Std Acc	Robust Acc	
		ℓ_∞	ℓ_2
Diffpure (Nie et al., 2022)	45.23	11.57	31.53
COUP (Zhang et al., 2024)	65.71	15.22	34.28
$\text{ADDT}_{w/\text{DDPM}}$ (Liu et al., 2025)	66.02	18.85	36.57
$\text{AMRM-Pure}_{\text{MAE}}$	65.34	14.28	29.29
$\text{AMRM-Pure}_{\text{MaskDiT}}$	**70.03**	24.39	36.51
$\text{RAMRM-Pure}_{\text{MAE}}$	66.28	19.53	31.58
$\text{RAMRM-Pure}_{\text{MaskDiT}}$	69.87	**29.91**	**43.27**

Table 3: Clean and robust accuracy (%) on SVHN obtained by different purification methods. The experiment are implemented on WideResNet-28-10.

Method	Std Acc	Robust Acc	
		ℓ_∞	ℓ_2
Diffpure (Nie et al., 2022)	93.90	39.70	63.30
COUP (Zhang et al., 2024)	92.07	41.62	63.97
ADBM (Li et al., 2025)	93.50	47.90	65.70
$\text{AMRM-Pure}_{\text{MAE}}$	94.54	27.59	55.29
$\text{AMRM-Pure}_{\text{MaskDiT}}$	94.91	46.57	66.38
$\text{RAMRM-Pure}_{\text{MAE}}$	94.47	39.15	60.51
$\text{RAMRM-Pure}_{\text{MaskDiT}}$	**95.39**	**55.90**	**70.18**

issue, recent studies (Chen et al., 2024; Li et al., 2025; Liu et al., 2025) have adopted the gradient checkpointing technique to efficiently extract complete gradients throughout the diffusion process. Furthermore, Li et al. (Li et al., 2025) have further demonstrated that, compared to AutoAttack, the combination of PGD + EOT is more effective in evaluating the adaptive defense mechanisms of diffusion models. In line with these studies (Li et al., 2025), we employ the PGD200 + EOT20 configuration with $\ell_\infty(\epsilon = \frac{8}{255})$ and $\ell_2(\epsilon = 1)$, utilizing the exact gradient computation method described in (Li et al., 2025; Liu et al., 2025) to ensure a more robust evaluation of defense performance in our experiments. Our threat model is purifier followed by a classifier. To ensure fair comparison with adversarial training methods, we use AutoAttack with full gradient settings (Chen et al., 2024) as the evaluation protocol, ensuring objectivity and comparability.

Evaluation Metrics. To evaluate the model's performance, we employ two metrics for classification: robust accuracy (**Robust Acc**) and standard accuracy (**Std Acc**), which are tested respectively on adversarial examples and clean samples. Due to the high computational cost of testing models with multiple attacks, we follow previous work (Nie et al., 2022; Lin et al., 2024; Li et al., 2025) and randomly select 512 test samples from each testing dataset. All results are from 5 different random seeds and we use its average values. The standard deviations of the experiments will be provided separately in the Supplementary Materials (not in the Appendix).

5.2 Compare with the State-of-the-art

We compare our results with state-of-the-art methods across four datasets: CIFAR-10, CIFAR-100, SVHN, and ImageNet. **Due to the space limitations, we provide detailed comparisons under the PGD200 + EoT20 attack for CIFAR-10, CIFAR-100, and SVHN in main paper. More experimental results, including performance on ImageNet, transferability of the fine-tuned purification model, defense against extra attacks, ablation studies, sensitivity analysis, and evaluations between different classifiers, are also presented in Appendix C.**

CIFAR-10. Table 1 highlights the performance of various purification methods on the CIFAR-10 dataset in terms of Std Acc and Robust Acc. $\text{RAMRM-Pure}_{\text{MaskDiT}}$ excels with 62. 13% robust accuracy in $\ell\infty$ attacks, 73. 57% in ℓ_2 attacks, and strong standard accuracy of 93. 11%. In contrast, traditional methods like the previous method (Shi et al., 2021) perform poorly, achieving only 4.56% under ℓ_∞ attacks. In general, our method greatly improves adversarial robustness, with additional WideResNet-70-16 results provided in Supplementary C.5.

CIFAR100. Table 2 summarizes the performance of various purification methods on CIFAR-100. Our proposed method achieves strong performance, with $\text{AMRM-Pure}_{\text{MaskDiT}}$ reaching the highest standard accuracy of 70.03% and a robust accuracy of 24.39% under ℓ_∞ attacks and 36.51% under ℓ_2 attacks. $\text{RAMRM-Pure}_{\text{MaskDiT}}$ further improves robustness, achieving the best ℓ_∞ robust accuracy of 29.91% and ℓ_2 robust accuracy of 43.27%, outperforming other methods like Diffpure and COUP.

SVHN. Table 3 shows that ADBM achieves strong robust accuracy (47.90% under ℓ_∞ attacks and 65.70% under ℓ_2) attacks but is outperformed by $\text{RAMRM-Pure}_{\text{MaskDiT}}$, which achieves the best robust accuracy (55.90% and 70.18%) with comparable standard accuracy. While $\text{RAMRM-Pure}_{\text{MAE}}$ leads in standard accuracy (95.39%), its robustness is lower. Overall, our MaskDiT-based methods better balance clean and robust performance than ADBM.

5.3 Comparison with adversarial training

As shown in Table 4, our methods achieve competitive or superior robustness compared to adversarial training baselines, **without using any extra data**. In contrast, prior works (Rebuffi et al., 2021; Pang et al., 2022; Wang et al., 2023) rely on **1M additional training samples**. Notably, **$\text{RAMRM-Pure}_{\text{MaskDiT}}$ achieves 75.83% robust accuracy on CIFAR-**

Table 4: Comparison with adversarial training under Autoattack ($\epsilon = \frac{8}{255}$)

Method	Extra data	Architecture	CIFAR10		CIFAR100		SVHN	
			Std Acc	ℓ_∞	Std Acc	ℓ_∞	Std Acc	ℓ_∞
Rebuffi et al.(Rebuffi et al., 2021)	✓	WRN-28-10	87.33	60.73	62.41	32.06	94.34	60.90
Pang et al. (Pang et al., 2022)	✓	WRN-28-10	88.10	61.51	62.08	31.40	–	–
Wang et al.(Wang et al., 2023)	✓	WRN-28-10	91.12	63.35	68.06	35.65	95.19	61.85
AMRM-Pure$_{\text{MAE}}$	✗	WRN-28-10	88.57	40.65	65.34	16.77	94.54	47.03
AMRM-Pure$_{\text{MaskDiT}}$	✗	WRN-28-10	92.03	64.97	**70.03**	33.51	94.91	64.15
RAMRM-Pure$_{\text{MAE}}$	✗	WRN-28-10	90.09	47.15	66.28	22.15	94.47	50.03
RAMRM-Pure$_{\text{MaskDiT}}$	✗	WRN-28-10	**93.11**	**75.83**	69.87	**40.13**	**95.39**	**66.12**

10, outperforming all baselines and highlighting the effectiveness of our data-free approach.

5.4 Inference Time Comparison

Table 5: Inference time (s) consumption comparison across different defense models on CIFAR-10 and ImageNet datasets.

Defense Model	CIFAR10	ImageNet
Diffpure (Nie et al., 2022)	12.39	81.54
AMRM-Pure$_{\text{MAE}}$	18.25	**31.51**
AMRM-Pure$_{\text{MaskDiT}}$	32.85	79.27
RAMRM-Pure$_{\text{MAE}}$	**11.77**	27.38
RAMRM-Pure$_{\text{MaskDiT}}$	29.73	62.52

Table 5 compares the inference time between different defense models in CIFAR-10 and ImageNet. We calculate the run-time for all methods with a batch size of 32, and our experiments are conducted on an A40 GPU. For CIFAR-10, **RAMRM-Pure$_{\text{MAE}}$** achieves the fastest time (**11.77s**), followed by Diffpure (**12.39s**) and AMRM-Pure$_{\text{MAE}}$ (**18.25s**). On ImageNet, **AMRM-Pure$_{\text{MAE}}$** is the most efficient (**31.51s**), significantly outperforming Diffpure (**81.54s**). The results highlight that the DiffPure model has the advantage of inference time for smaller datasets, while our proposed model performs better on the metrics of inference time on the ImageNet dataset.

6 CONCLUSION

This paper reveals the vulnerability of AMRM to subtle adversarial attacks, caused by adversarial noises that disrupt semantic relations in image patches. To address this, we propose the AMRM-Pure pipeline, a denoising method using Attention Matrix Variation Minimization, which iteratively refines adversarial examples by minimizing reconstruction loss to converge to clean images. We further enhance AMRM-Pure with classifier loss, introducing RAMRM-Pure$_{\text{MAE}}$. We apply two AMRM (MAE and MaskDiT) to our method. Extensive experiments demonstrate that our methods achieve state-of-the-art performance on multiple benchmarks.

7 ACKNOWLEDGEMENT

The work was partially supported by the following: the Zhejiang Provincial Natural Science Foundation – Exploration Project under No. LMS26F020007, the Wenzhou Applied Fundamental Research Program (Basic Research) under No. GG20250198, the WKU 2026 International Frontier Interdisciplinary Research Institute Talent Program under No. WKUTP2026002, the WKU 2025 International Collaborative Research Program under No. ICRPSP2025001.

References

M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020.

A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293, 2018.

M. Bai, W. Huang, T. Li, A. Wang, J. Gao, C. F. Caiafa, and Q. Zhao. Diffusion models demand contrastive guidance for adversarial purification to advance. *Internal Conference on Machine Learning*, 2024.

S. Cao, P. Xu, and D. A. Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. in 2017 ieee symposium on security and privacy (sp). *in 2017 IEEE Symposium on Security and Privacy (SP), pp.*, page 39–57, 2017.

H. Chen, Y. Dong, Z. Wang, X. Yang, C. Duan, H. Su, and J. Zhu. Robust classification via a single diffusion model. *ICML*, 2024.

J. Chen and Q. Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020.

K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al. Rethinking attention with performers. *International Conference on Learning Representations*, 2021.

F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020a.

F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020b.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

O. Dou, Z. Gao, H. Shen, Z. Yuan, S. Zhang, and K. Huang. Improving robust generalization with diverging spanned latent space. *Transactions on Machine Learning Research*, 2024.

V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. *International Conference on Computer Vision*, 2017.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *n International Conference on Learning Representations*, 2017.

Y. Lee, J. R. Willette, J. Kim, and S. J. Hwang. Visualizing the loss landscape of self-supervised vision transformer. *arXiv preprint arXiv:2405.18042*, 2024.

X. Li, W. Sun, H. Chen, Q. Li, Y. Liu, Y. He, J. Shi, and X. Hu. Adbm: Adversarial diffusion bridge model for reliable adversarial purification. *ICLR*, 2025.

G. Lin, C. Li, J. Zhang, T. Tanaka, and Q. Zhao. Adversarial training on purification (atop): Advancing both robustness and generalization. *International Conference on Learning Representations*, 2024.

G. Liu, I. Khalil, and A. Khreishah. Using single-step adversarial training to defend iterative adversarial examples. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 17–27, 2021.

Y. Liu, K. Liu, Y. Xiao, Z. Dong, X. Xu, P. Wei, and L. Lin. Towards understanding the robustness of diffusion-based purification: A stochastic perspective. In *The Thirteenth International Conference on Learning Representations*, 2025.

C. Lyu, K. Huang, and H.-N. Liang. A unified gradient regularization family for adversarial examples. *2015 IEEE international conference on data mining*, pages 301–309, 2015.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

H. Mei, M. Dong, and C. Xu. Efficient image-to-image diffusion classifier for adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6081–6089, 2025.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, Spain, 2011.

T. Nguyen, M. Pham, T. Nguyen, K. Nguyen, S. Osher, and N. Ho. Fourierformer: Transformer meets generalized fourier integral theorem. *Advances in Neural Information Processing Systems*, 35:29319–29335, 2022.

W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. Diffusion models for adversarial purification. *International Conference on Machine Learning*, 2022.

T. Pang, M. Lin, X. Yang, J. Zhu, and S. Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, pages 17258–17277. PMLR, 2022.

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *International Conference on Learning Representations*, 2018.

V. Sehwag, S. Mahloujifar, T. Handina, S. Dai, C. Xiang, M. Chiang, and P. Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *International Conference on Learning Representations*, 2021.

C. Shi, C. Holtz, and G. Mishne. Online adversarial purification based on self-supervision. *International Conference on Learning Representations*, 2021.

D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, and T. Kohno. Physical adversarial examples for object detectors. *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, Aug. 2018.

Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *International Conference on Learning Representations*, 2017.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations*, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

B. Vivek and R. V. Babu. Single-step adversarial training with dropout scheduling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 947–956. IEEE, 2020.

Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan. Better diffusion models further improve adversarial training. *International Conference on Machine Learning*, 2023.

E. Wong, L. Rice, and J. Z. Kolter. Fast is better than free: Revisiting adversarial training. *International Conference on Learning Representations*, 2020.

Q. Wu, H. Ye, Y. Gu, H. Zhang, L. Wang, and D. He. Denoising masked autoencoders help robust classification. *ICLR*, 2022.

J. Yoon, S. J. Hwang, and J. Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pages 12062–12072. PMLR, 2021.

Z. You, D. Liu, B. Han, and C. Xu. Beyond pretrained features: noisy image modeling provides adversarial defense. *Advances in Neural Information Processing Systems*, 36, 2023.

S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*, 2016.

B. Zhang, W. Luo, and Z. Zhang. Enhancing adversarial robustness via score-based optimization. *Advances in Neural Information Processing Systems*, 36:51810–51829, 2023.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482, 2019.

M. Zhang, J. Li, W. Chen, J. Guo, and X. Cheng. Classifier guidance enhances diffusion-based adversarial purification by preserving predictive information. In *ECAI 2024*, pages 2234–2241. IOS Press, 2024.

Q. Zhang, Y. Wang, and Y. Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.

H. Zheng, W. Nie, A. Vahdat, and A. Anandkumar. Fast training of diffusion models with masked transformers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

D. Zhou, Y. Chen, N. Wang, D. Liu, X. Gao, and T. Liu. Eliminating adversarial noise via information discard and robust representation restoration. In *International Conference on Machine Learning*, pages 42517–42530. PMLR, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

AMRM-Pure: Semantic-Preserving Adversarial Purification: Supplementary Materials

A Limitation

One limitation of AMRM-Pure$_{\text{MAE}}$ lies in its linear memory consumption with respect to batch size, which may restrict scalability under limited GPU resources.

B Broader impact

We are the first to systematically explore the relationship between adversarial noise and inter-patch semantic information. While existing defense methods primarily focus on suppressing pixel-level perturbations or enhancing model robustness structurally, our work takes a novel perspective by analyzing reconstruction consistency and semantic alignment. We reveal how adversarial perturbations disrupt inter-patch semantic relations and propose a reconstruction paradigm that restores this consistency. This new angle provides a valuable direction for future adversarial defense research and advances the theoretical and practical understanding of robustness from a structure-aware perspective.

C Supplement Experiment

We have enhanced this section with additional experiments to provide a more comprehensive evaluation of our work. Specifically, we present ImageNet (Deng et al., 2009) results under PGD200 + EoT20 (Madry et al., 2018; Athalye et al., 2018) with the perturbation budgets $\epsilon = \frac{4}{255}$ for ℓ_∞ attack and $\epsilon = 0.5$ for ℓ_2 attack, using ResNet-101 (He et al., 2016) as the classifier. We also performed ablation studies on CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and SVHN (Netzer et al., 2011) using different backbones and evaluated diverse attack scenarios.

C.1 Performance on ImageNet.

Table 6 presents the standard accuracy and robust accuracy of different purification methods on the ImageNet dataset under the ℓ_∞ attack. As shown in the table, the ADDT method achieves the highest standard accuracy at **80.20%**, slightly outperforming the other methods. However, in terms of robustness, RAMRM-Pure$_{\text{MaskDiT}}$ stands out with a robust accuracy of **36.87%**, surpassing all other methods, including ADDT. In contrast, AMRM-Pure$_{\text{MAE}}$ and AMRM-Pure$_{\text{MaskDiT}}$ demonstrate relatively lower robustness, achieving 24.75% and 32.29%, respectively.

C.2 Transferability of finetuned purification on new classifiers

We fine-tune the RAMRM-Pure$_{\text{MAE}}$/RAMRM-Pure$_{\text{MaskDiT}}$ model based on WideResNet-28-10 (Zagoruyko and Komodakis, 2016) and replace it with different classifiers for testing experiments. The WideResNet-70-16 (Zagoruyko and Komodakis, 2016) and ResNet-50 (He et al., 2016) are selected for testing process to observe the transferability of our proposed method across different classifiers.

The results in Table 7 highlight the superior performance of our method, particularly RAMRM-Pure$_{\text{MaskDiT}}$, which achieves the highest robust accuracy across both classifiers and attack norms. Notably, the fine-tuned models, initially trained on WideResNet-28-10, demonstrate strong transferability when applied to WideResNet-70-16 and ResNet-50 without the need for retraining. This finding underscores the practicality and scalability of our method, as its fine-tuned models can be seamlessly adapted to new classifiers, providing an efficient and robust defense against adversarial attacks.

Table 6: Clean and robust accuracy (%) with ℓ_∞ and ℓ_2 attack on ImageNet obtained by different purification methods.

Defense model	Std Acc	Robust Acc	
		ℓ_∞	ℓ_2
Diffpure (Nie et al., 2022)	77.51	30.15	44.15
ADDT (Liu et al., 2021)	**80.20**	35.83	-
AMRM-Pure$_{\text{MAE}}$	67.53	24.75	35.95
AMRM-Pure$_{\text{MaskDiT}}$	75.52	32.29	45.57
RAMRM-Pure$_{\text{MAE}}$	78.85	29.48	42.25
RAMRM-Pure$_{\text{MaskDiT}}$	79.52	**36.87**	**51.17**

Table 7: Robust accuracy (%) of different purification methods against $\ell_\infty(\epsilon = \frac{8}{255})$ and $\ell_2(\epsilon = 1)$ adversarial attacks across two classifiers: WideResNet-70-16 and ResNet-50. Here, our method are derived by fine-tuning on the WideResNet-28-10 classifier.

Classifier	WideResNet-70-16		ResNet-50	
	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2
Diffpure (Nie et al., 2022)	42.20	60.80	38.02	54.74
RAMRM-Pure$_{\text{MAE}}$	44.27	62.42	43.72	60.08
RAMRM-Pure$_{\text{MaskDiT}}$	**58.37**	**68.55**	**54.22**	**67.15**

C.3 Performance on unseen threats

Table 8: Robust accuracy (%) against unseen threats with the setting of $\ell_1(\epsilon = 12)$ and $\ell_2(\epsilon = 1)$.

Defense model	CIFAR-10		CIFAR-100		SVHN	
	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2
Diffpure (Nie et al., 2022)	44.30	60.80	13.51	27.53	46.10	63.30
ADBM (Li et al., 2025)	49.60	63.30	-	-	51.20	65.70
RAMRM-Pure$_{\text{MAE}}$	44.41	60.72	12.97	29.58	47.09	60.51
RAMRM-Pure$_{\text{MaskDiT}}$	**65.11**	**73.57**	**41.15**	**43.27**	**55.53**	**70.18**

For the three methods, ADBM, RAMRM-Pure$_{\text{MAE}}$, and RAMRM-Pure$_{\text{MaskDiT}}$, all of which are fine-tuned under the ℓ_∞ norm, the ℓ_1 and ℓ_2 norms are considered as unseen threats. To verify the robustness of the proposed method, we will now conduct testing under these unseen threats. For a fair comparison on the CIFAR-10 dataset, we employ the WideResNet-70-16 architecture, and we use the WideResNet-28-10 architecture on the CIFAR-100 and SVHN datasets.

Table 8 presents the robust accuracy (%) of different defense models against unseen threats (ℓ_1 and ℓ_2 attacks) on the CIFAR-10, CIFAR-100, and SVHN datasets. As a baseline method, DiffPure (Nie et al., 2022) performs moderately on CIFAR-10 and SVHN but poorly on CIFAR-100, especially under ℓ_1 attacks (13.51%). ADBM outperforms DiffPure on CIFAR-10 and SVHN, but no data is provided for CIFAR-100, suggesting potential limitations or untested performance on this dataset. Our RAMRM-Pure$_{\text{MAE}}$ method slightly outperforms DiffPure on CIFAR-10 and SVHN but underperforms on CIFAR-100 (12.97% vs. 13.51%), indicating some limitations on more complex datasets. In contrast, RAMRM-Pure$_{\text{MaskDiT}}$ significantly outperforms all other methods across all datasets and attack types. On CIFAR-10, RAMRM-Pure$_{\text{MaskDiT}}$ achieves accuracies of 65.11% and 73.57% under ℓ_1 and ℓ_2 attacks, respectively, far surpassing other methods. On CIFAR-100, although its performance under ℓ_2 attacks is slightly lower than under ℓ_1, it still outperforms other methods. On the SVHN dataset, RAMRM-Pure$_{\text{MaskDiT}}$ also demonstrates considerable robustness performance, particularly under

ℓ_2 attacks (70.18%). Overall, RAMRM-Pure$_{\text{MaskDiT}}$ exhibits the strongest robustness against unseen threats, especially on CIFAR-10 and CIFAR-100, showcasing its superior generalization and defense capabilities, while RAMRM-Pure$_{\text{MAE}}$, though slightly less effective, still outperforms baseline methods in certain scenarios.

C.4 Defense against adaptive attacks

Table 9 presents the robust accuracy (%) of various defense methods under different adversarial attacks in the adaptive $\ell_2(\epsilon = 1)$-norm setting on the CIFAR-10 dataset. The evaluated adaptive attacks include C&W (Carlini and Wagner, 2017)+EOT (Athalye et al., 2018), DeepFool (Moosavi-Dezfooli et al., 2016)+EOT, AutoAttack (Croce and Hein, 2020b)+EOT, and PGD (Madry et al., 2018)+EOT. Among the methods, Diffpure and ADBM represent baseline defense approaches, with ADBM generally outperforming Diffpure across all attacks. For instance, ADBM achieves 78.40% robust accuracy against C&W+EOT compared to Diffpure's 74.80%. The pure methods (non-adversarial training approaches) show varying performance: AMRM-Pure$_{\text{MAE}}$ exhibits the lowest robust accuracy across all attacks, while AMRM-Pure$_{\text{MaskDiT}}$ demonstrates stronger performance, particularly against DeepFool+EOT (82.8%). RAMRM-Pure$_{\text{MAE}}$ shows moderate results, and RAMRM-Pure$_{\text{MaskDiT}}$ consistently outperforms all other methods, achieving the highest robust accuracy against every attack, with 80.58% for C&W+EOT, 86.11% for DeepFool+EOT, 73.59% for AutoAttack+EOT, and 69.57% for PGD+EOT. This indicates that RAMRM-Pure$_{\text{MaskDiT}}$ is the most effective defense method in this setting, offering superior robustness across diverse adversarial attacks.

Table 9: Robust Accuracy (%) of various defense methods under different attacks in the $\ell_2(\epsilon = 1)$-norm setting using the exact gradient with WideResNet-70-16 on CIFAR-10.

Method	C&W+EOT	DeepFool+EOT	AutoAttack+EOT	PGD+EOT
Diffpure(Nie et al., 2022)	74.80	78.40	63.90	60.80
ADBM(Li et al., 2025)	78.40	84.30	66.80	66.30
AMRM-Pure$_{\text{MAE}}$	62.54	65.15	52.75	53.50
AMRM-Pure$_{\text{MaskDiT}}$	79.15	82.80	66.43	64.53
RAMRM-Pure$_{\text{MAE}}$	72.20	72.29	59.11	60.72
RAMRM-Pure$_{\text{MaskDiT}}$	**80.58**	**86.11**	**73.59**	**69.57**

C.5 Extra experiments on different classifier

Table 10 shows our standard and robust accuracy using WideResNet-70-16 under CIFAR-10 and SVHN. Compared with WideResNet-28-10, it shows better results. It means the overparameterization contributes model's robustness. Among all the methods, the effectiveness of the RAMRM-Pure$_{\text{MaskDiT}}$ method is most notable.

Table 10: Performance of standard accuracy and robust accuracy (%) using WideResNet-70-16.

Method	Architecture	CIFAR10			SVHN		
		Std Acc	ℓ_∞	ℓ_2	Std Acc	ℓ_∞	ℓ_2
AMRM-Pure$_{\text{MAE}}$	WideResNet-70-16	89.66	42.21	56.77	94.97	17.11	32.75
AMRM-Pure$_{\text{MaskDiT}}$	WideResNet-70-16	94.91	52.07	66.55	94.93	46.03	63.77
RAMRM-Pure$_{\text{MAE}}$	WideResNet-70-16	91.07	47.92	60.95	94.78	40.79	60.51
RAMRM-Pure$_{\text{MaskDiT}}$	WideResNet-70-16	**93.85**	**63.94**	**75.50**	**95.91**	**56.02**	**69.18**

C.6 Performance on Black-box Attack

To evaluate the effectiveness of against black-box attacks, we adopt three black-box attack methods: FAB (Croce and Hein, 2020a), Square (Andriushchenko et al., 2020), and Rays (Chen and Gu, 2020) on CIFAR-10 and SVHN. The black-box scenario implies that the attacker has no knowledge of the defense method. Table 11 shows

the robustness of various methods against black-box ℓ_∞ attacks with the perturbation budget $\epsilon = \frac{8}{255}$ using WideResNet-28-10. RMaskDiT still achieves the best results.

Table 11: Robust accuracy (%) against different black-box attacks $\ell_\infty(\epsilon = \frac{8}{255})$ with WideResNet-28-10. The "Vanilla" setting represents the model trained on clean datasets without any defense.

Method	Architecture	CIFAR-10				SVHN			
		Std Acc	Robust Acc			Std Acc	Robust Acc		
			Square	FAB	RayS		Square	FAB	RayS
Vanilla	WideResNet-28-10	96.75	19.15	0.00	1.23	98.11	9.08	14.78	16.89
Diffpure (Nie et al., 2022)	WideResNet-28-10	89.15	89.15	88.29	90.51	93.93	92.15	93.13	92.97
ADBM (Li et al., 2025)	WideResNet-28-10	-	-	-	-	93.49	93.32	92.98	93.16
AMRM-Pure$_{\text{MAE}}$	WideResNet-28-10	88.57	78.59	76.43	77.29	94.54	92.57	93.36	93.41
AMRM-Pure$_{\text{MaskDiT}}$	WideResNet-28-10	92.03	90.96	92.25	**93.39**	94.91	92.80	92.59	92.73
RAMRM-Pure$_{\text{MAE}}$	WideResNet-28-10	90.09	90.25	89.15	92.31	94.47	92.73	93.27	93.36
RAMRM-Pure$_{\text{MaskDiT}}$	WideResNet-28-10	**93.11**	**93.27**	**93.38**	93.03	**95.39**	**94.15**	**94.18**	**94.88**

C.7 Robust under BPDA attack

We evaluate the robustness of our model under a strong white-box attack setting using BPDA combined with EoT set to 20. The results show as follow:

Table 12: Clean and robust accuracy (%) under BPDA attack ($\epsilon = \frac{8}{255}$) on CIFAR-10.

Method	Architecture	Std Acc	Robust Acc (ℓ_∞)
Diffpure (Nie et al., 2022)	WRN-28-10	89.20	78.53
AMRM-Pure$_{\text{MAE}}$	WRN-28-10	88.57	78.89
AMRM-Pure$_{\text{MaskDiT}}$	WRN-28-10	92.03	83.44
RAMRM-Pure$_{\text{MAE}}$	WRN-28-10	90.09	80.17
RAMRM-Pure$_{\text{MaskDiT}}$	WRN-28-10	**93.11**	**85.41**

Table 12 presents the comparison of clean and robust accuracy under BPDA attack ($\epsilon = \frac{8}{255}$) on the CIFAR-10 dataset. While the conventional Diffpure method achieves decent robustness, our proposed methods demonstrate significant improvements in both clean and robust accuracy. In particular, RAMRM-Pure$_{\text{MaskDiT}}$ achieves the highest clean accuracy (93.11%) and robust accuracy (85.41%), highlighting its superior purification capability and enhanced resistance to adversarial attacks. These results validate the effectiveness of our approach in improving semantic reconstruction and adversarial robustness.

C.8 A specific attacks

We consider a white-box adversarial attack that jointly optimizes reconstruction fidelity and classification error by minimizing a weighted combination of reconstruction loss and negative classification loss. Specifically, given an input x, the attacker generates an adversarial example x_{adv} by solving the following optimization problem:

$$x_{\text{adv}} = \arg\min_{x'} \, (1-\alpha)\, \mathcal{L}_{\text{rec}}(x') \; - \; \alpha\, \mathcal{L}_{\text{cls}}(x') \tag{5}$$

where $\mathcal{L}_{\text{rec}}$ denotes the reconstruction loss and $\mathcal{L}_{\text{cls}}$ is the classification loss that encourages misclassification. The trade-off parameter $\alpha \in [0, 1]$ controls the balance between preserving reconstruction quality and inducing misclassification. The attack is implemented using a strong PGD-based procedure with sufficient iterations and step size tuning to avoid gradient obfuscation. This formulation generalizes standard adversarial attacks by incorporating a reconstruction constraint, aiming to produce adversarial examples that remain visually consistent while still fooling the classifier.

Results and Analysis. We evaluate the effectiveness of the proposed joint-loss attack under different settings of the trade-off coefficient α and step size β.

Effect of α. We vary $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to study the trade-off between reconstruction fidelity and adversarial strength. The results are shown in Table 13.

Table 13: Effect of the trade-off parameter α on the joint-loss attack.

α	0.1	0.3	0.5	0.7	0.9
AMRM-Pure$_{\text{MaskDiT}}$	89.77	88.29	87.54	87.10	87.24
RAMRM-Pure$_{\text{MaskDiT}}$	90.17	89.25	89.57	89.11	88.27

As α increases, the adversarial objective becomes stronger, but the reconstruction quality degrades. Specifically, when $\alpha = 0.1$, the reconstruction loss is low ($\mathcal{L}_{rec}^{adv} = 0.04$), but the negative classification loss remains small ($-\mathcal{L}_{cls} = -0.19$), indicating a weak attack. In contrast, when $\alpha = 0.9$, the attack becomes much stronger ($-\mathcal{L}_{cls} = 4.4$), but the reconstruction loss increases significantly ($\mathcal{L}_{rec}^{adv} = 0.19$), leading to severely degraded reconstructions. These results reveal a clear trade-off between reconstruction fidelity and adversarial effectiveness.

Effect of β. We further vary the step size $\beta \in \{0.01, 0.05, 0.1, 0.5, 1\}$, and report the results in Table 14.

Table 14: Effect of the step size β on the joint-loss attack.

β	0.01	0.05	0.1	0.5	1
AMRM-Pure$_{\text{MaskDiT}}$	84.49	83.22	87.54	88.59	89.27
RAMRM-Pure$_{\text{MaskDiT}}$	85.59	87.92	89.57	88.62	89.54

In table 14, we observe that the attack becomes more effective when β is around 0.05–0.1. However, even under the best step size, the proposed attack still does not outperform the standard full-gradient white-box attack used in our paper.

From the above results, we conclude that the joint-loss attack struggles to achieve a satisfactory balance between reconstruction fidelity and adversarial effectiveness. In particular, no choice of α can simultaneously ensure low reconstruction loss and strong adversarial impact. This suggests that the reconstruction objective and classification objective are inherently conflicting under this formulation, making the attack weaker than a standard full-gradient white-box attack.

C.9 Ablation study

C.9.1 Impact of time step number

To investigate the impact of time steps on the denoising process in MaskDiT, we conduct experiments by observing the robust accuracy at different time steps, aiming to understand how varying time steps influence the model's ability to effectively remove noise and improve overall performance.

Table 15: Impact of time step on CIFAR-10, and all configurations align with Table 1.

Time steps	15	20	25	30
AMRM-Pure$_{\text{MaskDiT}}$	49.27	50.41	50.57	**51.69**
RAMRM-Pure$_{\text{MaskDiT}}$	49.22	51.34	**62.13**	58.22

As shown in Table 15, our method exhibits a stable increase in robust accuracy as time steps increase, peaking at **51.69** when the metric of time steps is set as 30. In contrast, RAMRM-Pure$_{\text{MaskDiT}}$ achieves its highest accuracy of **62.13** when the metric of time steps is set as 25, but experiences a slight drop at the 30th step, indicating its sensitivity to the optimal time step selection.

C.9.2 Ablation on the Effectiveness of MaskDiT framework

To explore the origins of the enhanced performance observed with AMRM-Pure$_{\text{MaskDiT}}$ and RAMRM-Pure$_{\text{MaskDiT}}$, we conducted a series of controlled experiments in Table 16. We begin with training two models, e.g., MaskDiT and RMaskDiT, used for purification. Firstly, following the DiffPure (Nie et al., 2022), we conduct the reconstruction-based purification processes for purification, which are MaskDiT$_{/\text{purification}}$ and RMaskDiT$_{/\text{purification}}$. This

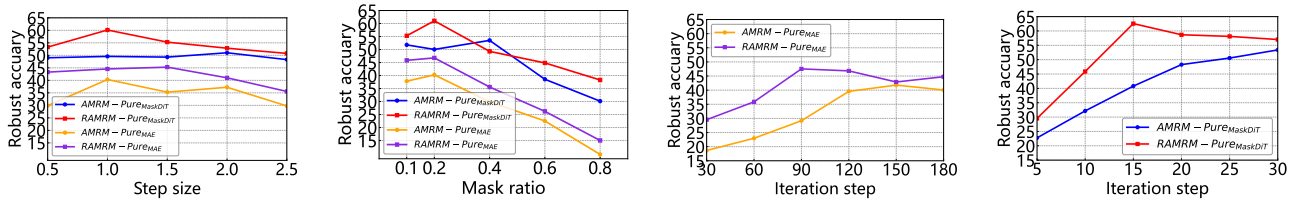
reconstruction based purification involves a forward (noise addition) and backward (denoising) pass. The reconstructed images are treated as purified results to investigate whether the superior performance stems from generative ability of selected models. On the other hand, we then compared the above performances with the proposed AMV-based purification methods by using the same models, where the purification processes are named as AMRM-Pure$_{\text{MaskDiT}}$ and RAMRM-Pure$_{\text{MaskDiT}}$. By keeping the model architecture consistent across all variants, we enabled a direct comparison of the effectiveness of different denoising strategies. It is note that MaskDiT and RMaskDiT indeed possess a certain capability to withstand stronger adversarial attacks; however, their effectiveness is still much lower than that of our proposed AMRM-Pure$_{\text{MaskDiT}}$ and RAMRM-Pure$_{\text{MaskDiT}}$.

Table 16: Ablation analysis on different denoising components across various datasets.

Method	Architecture	CIFAR10		CIFAR100		SVHN	
		Std acc	Robust acc	Std acc	Robust acc	Std acc	Robust acc
MaskDiT$_{\text{/purification}}$	WRN-28-10	91.13	42.99	63.29	13.57	92.28	40.07
RMaskDiT$_{\text{/purification}}$	WRN-28-10	90.57	47.57	64.15	18.55	93.03	45.19
AMRM-Pure$_{\text{MaskDiT}}$	WRN-28-10	92.03	50.57	**70.03**	24.39	94.91	46.57
RAMRM-Pure$_{\text{MaskDiT}}$	WRN-28-10	**93.11**	**62.13**	69.87	**29.91**	**95.39**	**55.90**

C.10 Sensitivity Analysis

In this subsection, taking the CIFAR10 under same setting with Table 1, we analyze the impact of step size, mask ratio and step size. The result is described in Fig. 5. It confirms our theoretical analysis.



(a) Step size (b) Mask ratio (c) Iteration /w MAE (d) Iteration /w MaskDiT

Figure 5: Sensitivity analysis

C.10.1 Is non-attentive MRM effective?

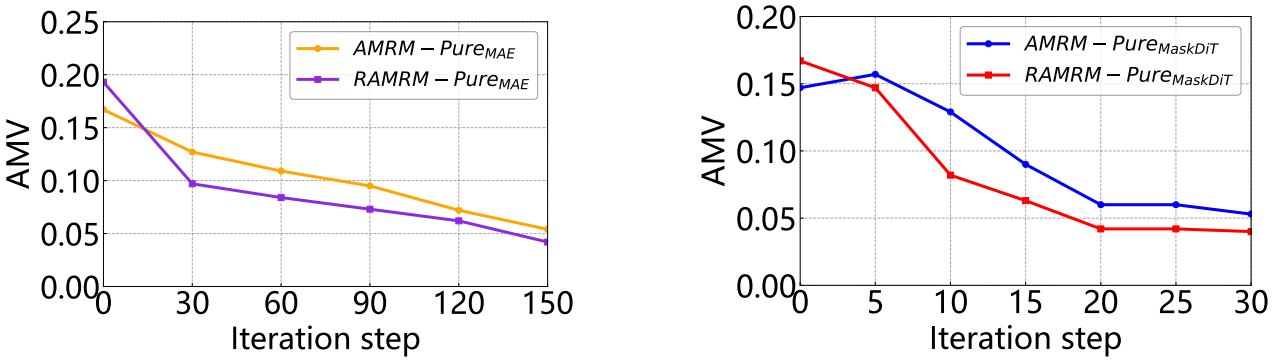
To investigate whether non-attentive MRMs are effective, we adopt a CNN-based architecture similar to MAE, namely Context Encoders (CE) (Pathak et al., 2016), and propose CE-Pure in a manner analogous to AMRM-Pure$_{\text{MAE}}$. We show their results in Table 17 Context Encoders, like MAE, are self-supervised methods based on masking and reconstruction, but differ in architectural design: CE relies on CNNs that excel at capturing local structural information and are widely used for image inpainting, while MAE is Transformer-based and better at modeling global contextual information. We evaluate CE-Pure on CIFAR-10 and SVHN using WRN-28 as the classifier, and adopt PGD200+EOT20 with full gradient for consistency with the main paper. The results are as follows:

Table 17: Comparison of CE-Pure and AMRM-Pure$_{\text{MAE}}$ results on CIFAR-10 and SVHN using WRN-28 as the classifier with Sta Acc and Robust Acc ($\ell_\infty = \frac{8}{255}$).

Method	CIFAR-10		SVHN	
	Std Acc	Robust Acc	Std Acc	Robust Acc
CE-Pure	82.59	6.75	85.33	8.59
AMRM-Pure$_{\text{MAE}}$	**88.57**	**40.53**	**95.39**	**55.90**

The results show that the robustness of AMRM-Pure$_{\text{MAE}}$ is significantly stronger than that of CE-Pure. This mechanism cannot be transferred to CNNs, thus proving that AMRM has better robustness.

D Robust MAE and MaskDiT analysis



(a) (b)

Figure 6: Impact of purification iterations on the AMV using CIFAR-10.

D.1 Fine-tuning of RMRM-Pure

Stage I: We begin by generating adversarial example dataset X'_{adv}, and it achieves the purifier-classifier system as follows:

$$\mathrm{X}'_{adv} = \max_{\delta} \left[\sum_{(x,y)\in \mathrm{X}} \mathcal{L}_{\mathcal{C}} \left(\mathcal{P}_{\theta}(x' + \delta), y \right) \right], \tag{6}$$

where x' represents the perturbed sample x with added Gaussian noise, and y denotes its corresponding label from the training dataset X. The functions $\mathcal{P}_{\theta}(\cdot)$ and $\mathcal{L}_{\mathcal{C}}(\cdot)$ correspond to the purification process and the loss of the classifier, respectively. x'_{adv} is an adversarial sample designed to target the entire purifier-classifier system. It is utilized during the subsequent fine-tuning stage to improve the overall robustness of the system.

Stage II: We choose to use the generated adversarial example x'_{adv} from adversarial dataset X'_{adv} for fine-tuning the purification model as follows:

$$\min_{\theta} \mathcal{L}_{fine}(x'_{adv}, y, \theta) = \min_{\theta} \sum_{(x'_{adv},y)\in \mathrm{X}'_{adv}} \mathcal{L}_{\mathcal{C}} \left(x'_{adv}, y \right), \tag{7}$$

where θ represents the weight of the purification model.

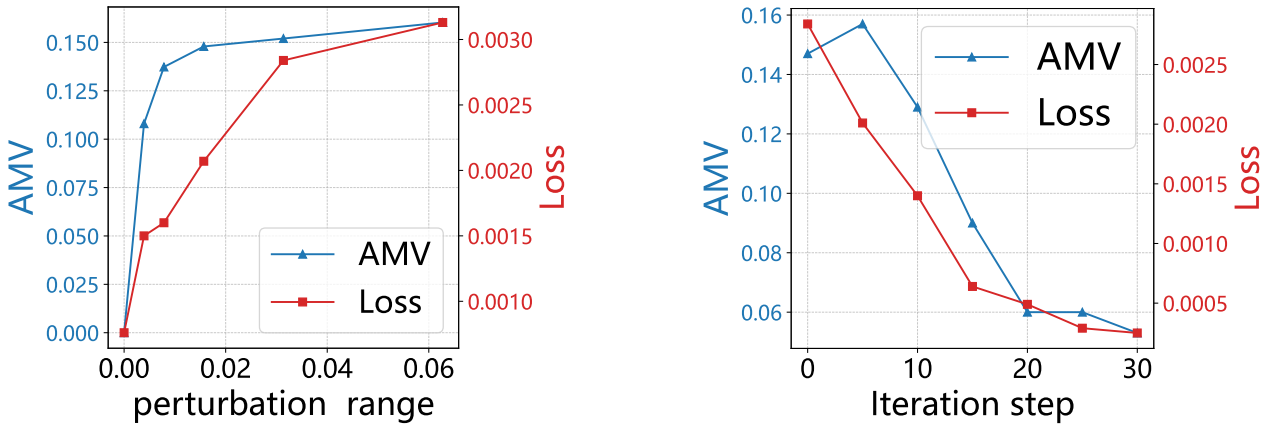
Compared to AMRM-Pure, RAMRM-Pure optimizes Eq. (6) to steer denoised images toward the classifier domain of natural datasets. This operation ensures that the attention distribution of denoised images more closely resembles that of natural samples, thereby reducing AMV and preserving inter-patch semantic relationships. As a result, RAMRM-Pure achieves superior robustness and generalization.

E Validation of MaskDiT

Figure 7 compares the reconstruction loss, AMV, and perturbation of MaskDiT with those of MAE shown in Figures 4, revealing similar patterns. Like MAE, MaskDiT is highly sensitive to noise, with AMV increasing sharply under minimal perturbations (e.g., $\delta = \frac{1}{255}$). Based on this observation, we extend AMRM-Pure$_{\mathrm{MAE}}$ manner to MaskDiT.

F Algorithm

Algorithm 1 describes the AMRM-Pure defense procedure. The fine-tuning process of RAMRM-Pure consists of two stages.



(a) (b)

Figure 7: The relationship between the trends of MaskDit loss, AMV, and perturbation.

Algorithm 1 AMRM-Pure.

Input: Adversarial Example x_{adv}, Step Size λ, Number of iteration S, clipping threshold η.
Output: Denoised data x_{den}.

```
s ← 0
x^s_adv ← x_adv
while s ≤ S do
    s ← s + 1
    Gain the AMRM reconstruction loss for adversarial examples L_rec(x^s_adv)
    Δ_s = sign(∇_x L_rec(x^s_adv))
    x^s_adv ← clip(x^s_adv − λΔ_s, η)
end while
x_den ← x^S_adv
```

G Visualization Analysis

The visualization analysis of our proposed method is illustrated as Fig. 8. It qualitatively verifies the effectiveness of our proposed method.

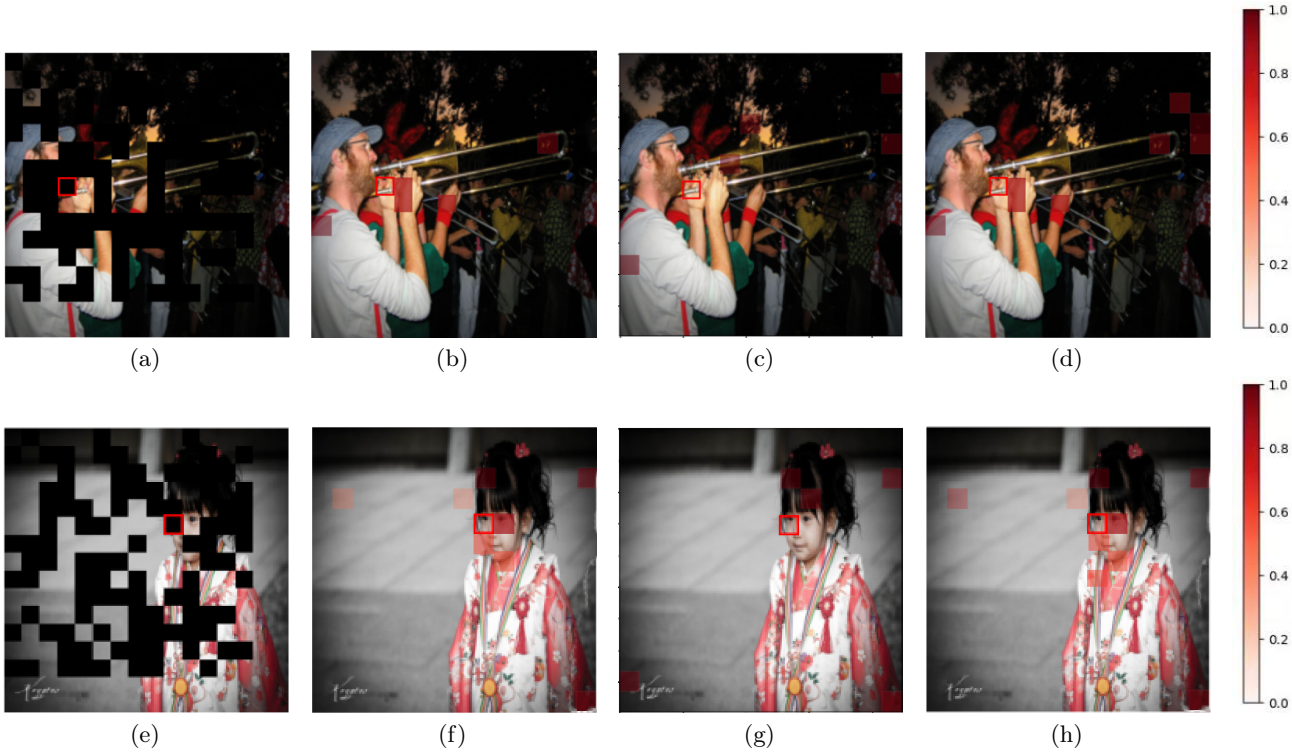


Figure 8: The first column, Fig. (a) and (e), represents the Mask Matrix. The second column, Fig. (b) and (f), illustrates the Attention Weights for clean samples. The third column, Fig. (c) and (k), depicts the Attention Weights for adversarial examples. The fourth column, Fig. (d) and (l), showcases the Attention Weights for denoised samples (by our AMRM-Pure$_{\text{MAE}}$). Patches with a deeper red color mean the elements with more attention. The data is sampled from the ImageNet dataset (Deng et al., 2009).

H Convergence Analysis of Purification Process.

For a well-trained MAE model, the reconstruction loss $\mathcal{L}_{\text{rec}}(x)$ is expected to reach a local minimum $\mathcal{L}^*_{\text{rec}}$,where the input x corresponds to a clean example, i.e., $\mathcal{L}_{\text{rec}}(x) \approx \mathcal{L}^*_{\text{rec}}$. Motivated by prior analysis (Lee et al., 2024; Zhang et al., 2022) that the loss landscape of MAE is smoother and exhibits wider convex regions, the reconstruction loss $\mathcal{L}_{\text{rec}}$ can be regarded as weakly convex within the neighborhood $[x-\delta, x+\delta]$ around the clean input x. Purified samples at the s-th iteration of the denoising process are represented as x^s_{adv}, and their corresponding reconstruction loss is denoted as $\mathcal{L}^s_{\text{rec}} = \mathcal{L}_{\text{rec}}(x^s_{adv}, \mathbf{U})$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$ represents the MAE mask.

Theorem H.1. *Let $\{\mathbf{U}_i\}_{i=1}^E$ be mask set which contains all possible masks with mask ratio ρ. After S optimization iterations according to Eq.4, it holds that:*

$$\frac{1}{(1-\rho)E}\sum_{e=1}^{E}[\mathcal{L}_{rec}(\frac{1}{S+1}\sum_{s=0}^{S}x^s_{adv},\mathbf{U}_e)-\mathcal{L}^*_{rec}] \leq \frac{1}{(1-\rho)E}\sum_{e=1}^{E}[\frac{||x_{adv}-x||_2^2}{2\lambda(S+1)}+\frac{\lambda}{2(S+1)}\sum_{s=0}^{S}||\nabla_x\mathcal{L}_{rec}(x^s_{adv},\mathbf{U}_e)||_2^2],$$

where λ is the step size and η is the clipping threshold.

Proof. The proof is relegated to Supplementary Material I.4. □

Theorem H.1 provides an upper bound on the gap between the averaged reconstruction loss during purification and the optimal loss $\mathcal{L}^*_{\text{rec}}$. As S increases, the first term vanishes at $\mathcal{O}(1/S)$ and the second term decreases as the gradient norm $\|\nabla_x\mathcal{L}_{\text{rec}}(x^s_{adv}, \mathbf{U}_e)\|_2^2$ diminishes. Hence, the RHS tends to zero, implying that the LHS also converges to zero. As a result, $\mathcal{L}_{\text{rec}}\left(\frac{1}{S+1}\sum_{s=0}^{S}x^s_{adv}, \mathbf{U}_e\right) \to \mathcal{L}^*_{\text{rec}}(x)$, indicating that the denoised sample progressively approximates the clean example in the reconstruction space.

I Proof of Theoretical Analysis

I.1 Assumption

Assumption 1. *There exists a pseudo-inverse encoder f_g that satisfies $\|g(f_g(a)) - a\|_2 \leq c_{rec}$ for any non-degenerate decoder, where a can be ether x_1 or x_2.* Here, x_1 and x_2 indicate visible portion and masked patches of input image x. c_{rec} is reconstruction bias, which symbolizes the disparity between the output of MAE and the original, unmasked image.

Remark. To facilitate the theoretical analysis of MAE, we borrow the above reasonable assumption from the previous study (Zhang et al., 2022). Intuitively, this assumption states that within the MAE framework, the decoder g is non-degenerate—i.e., its outputs retain meaningful information—and there exists a corresponding pseudo-inverse encoder f_g, such that their composition $h_g = g \circ f_g$ can approximately reconstruct either the visible patches x_1 or the masked patches x_2 of the input image. Physically, this implies that the decoder in MAE has sufficient capacity for recovering local structures from latent representations, a property empirically verified in many Transformer-based autoencoding models. This assumption provides the theoretical foundation for connecting MAE's reconstruction loss with alignment loss and helps interpret MAE as implicitly performing contrastive alignment through its masking mechanism.

Assumption 2 (Lipschitz Continuity). Let $\mathbf{A}_i^{\text{dec}} = \Phi\big(\{\mathbf{A}_{i,t}^{\text{dec}}\}_{t=1}^T\big)$ denote the final decoder attention matrix obtained from the per-layer attention matrices $\{\mathbf{A}_{i,t}^{\text{dec}}\}_{t=1}^T$. We assume Φ is L-Lipschitz in the domain of interest, i.e., there exists a constant $L > 0$ such that for any sets of matrices $\{\mathbf{X}_t\}_{t=1}^T$ and $\{\mathbf{Y}_t\}_{t=1}^T$:

$$\big\|\Phi(\mathbf{X}_1, \ldots, \mathbf{X}_T) - \Phi(\mathbf{Y}_1, \ldots, \mathbf{Y}_T)\big\| \;\leq\; L \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{Y}_t\|.$$

Under this assumption, for the adversarially perturbed attention matrices $\{\mathbf{A}_{\text{adv},i,t}^{\text{dec}}\}$, the final attention matrix satisfies:

$$\big\|\mathbf{A}_{\text{adv},i}^{\text{dec}} - \mathbf{A}_i^{\text{dec}}\big\|^2 \;\leq\; L^2\, T \sum_{t=1}^{T} \Big\|\mathbf{A}_{\text{adv},i,t}^{\text{dec}} - \mathbf{A}_{i,t}^{\text{dec}}\Big\|^2.$$

Hence, we can further write as:

$$\big\|\mathbf{A}_{\text{adv},i}^{\text{dec}} - \mathbf{A}_i^{\text{dec}}\big\|^2 \;\leq\; \frac{H}{T} \sum_{t=1}^{T} \Big\|\mathbf{A}_{\text{adv},i,t}^{\text{dec}} - \mathbf{A}_{i,t}^{\text{dec}}\Big\|^2,$$

by setting $H = L^2 T^2$, thereby bounding the overall adversarial effect via the per-layer differences.

I.2 Proof of Theorem 3.1

Proof:

$$||\mathbf{A}_{\text{adv}}^t - \mathbf{A}^t||_2 = \big\|\text{softmax}(\mathbf{Q}_{\text{adv}}^t(\mathbf{K}_{\text{adv}}^t)^T) - \text{softmax}(\mathbf{Q}^t(\mathbf{K}^t)^T)\big\|_2$$

To streamline the equation, we employ kernel methods as a substitute for the softmax$(\cdot)$ function (Choromanski et al., 2021). Let the kernel function be denoted as $\phi(\cdot)$:

$$\text{softmax}(\mathbf{Q}\mathbf{K}^T) \approx \langle \phi(\mathbf{Q}), \phi(\mathbf{K}) \rangle$$

The definition is written as follows:

$$\phi(x) = \frac{d(x)}{\sqrt{k}} \{f(\omega_1^\top x), ..., f(\omega_k^\top x)\}, \quad d(x) = \exp\left(-\frac{||x||^2}{2}\right), \quad f(x) = \exp(x)$$

Thus, we obtain:

$$\mathbf{A}^t_{\text{adv}} - \mathbf{A}^t \approx \frac{1}{m}\left[\exp\left(-\frac{||\mathbf{Q}^t_{\text{adv}}||^2 + ||\mathbf{K}^t_{\text{adv}}||^2}{2}\right)\sum_{i=0}^{k}\exp\left(\omega_i^\top(\mathbf{Q}^t_{\text{adv}} + \mathbf{K}^t_{\text{adv}})\right)\right.$$
$$\left. - \exp\left(-\frac{||\mathbf{Q}^t||^2 + ||\mathbf{K}^t||^2}{2}\right)\sum_{i=0}^{k}\exp\left(\omega_i^\top(\mathbf{Q}^t + \mathbf{K}^t)\right)\right]$$

For: $\mathbf{Q}^t_{\text{adv}} = \mathbf{Q}^t + \Delta\mathbf{Q}^t$, $\mathbf{K}^t_{\text{adv}} = \mathbf{K}^t + \Delta\mathbf{K}^t$

Following (Nguyen et al., 2022; Moosavi-Dezfooli et al., 2016), using Taylor expansion approximations, the formulation can be deduced as follows:

$$\exp\left(-\frac{\|\mathbf{Q}^t_{\text{adv}}\|^2 + \|\mathbf{K}^t_{\text{adv}}\|^2}{2}\right) \geq \exp\left(-\frac{\|\mathbf{Q}^t\|^2 + \|\mathbf{K}^t\|^2}{2}\right)\left(1 - (\mathbf{Q}^t)^\top\Delta\mathbf{Q}^t - (\mathbf{K}^t)^\top\Delta\mathbf{K}^t\right)$$

Let $\mathbf{S} = \mathbf{Q}^t + \mathbf{K}^t$, $\Delta\mathbf{S} = \Delta\mathbf{Q}^t + \Delta\mathbf{K}^t$, there exists a following equation:

$$\sum_{i=0}^{k}\exp\left(\omega_i^\top(\mathbf{S} + \Delta\mathbf{S})\right) \geq \sum_{i=0}^{k}\exp\left(\omega_i^\top\mathbf{S}\right) + \sum_{i=0}^{k}\exp\left(\omega_i^\top\mathbf{S}\right)\omega_i^\top\Delta\mathbf{S}$$

The definition is written as follows:

$$\mathbf{B} = \sum_{i=0}^{k}\exp\left(\omega_i^\top\mathbf{S}\right), \quad \mathbf{Y} = \sum_{i=0}^{k}\exp\left(\omega_i^\top\mathbf{S}\right)\omega_i$$

Then, we deduce the following sum equation:

$$\sum_{i=0}^{k}\exp\left(\omega_i^\top(\mathbf{Q}^t_{\text{adv}} + \mathbf{K}^t_{\text{adv}})\right) \approx \mathbf{B} + \mathbf{Y}^\top\Delta\mathbf{S} = \mathbf{B} + \mathbf{Y}^\top(\Delta\mathbf{Q}^t + \Delta\mathbf{K}^t)$$

The substitution into difference is:

$$D \approx \frac{1}{m}\exp\left(-\frac{\|\mathbf{Q}^t\|^2 + \|\mathbf{K}^t\|^2}{2}\right)\left[\left(1 - (\mathbf{Q}^t)^\top\Delta\mathbf{Q}^t - (\mathbf{K}^t)^\top\Delta\mathbf{K}^t\right)\left(\mathbf{B} + \mathbf{Y}^\top(\Delta\mathbf{Q}^t + \Delta\mathbf{K}^t)\right) - \mathbf{B}\right]$$
$$\approx \frac{1}{m}\exp\left(-\frac{\|\mathbf{Q}^t\|^2 + \|\mathbf{K}^t\|^2}{2}\right)\left[\mathbf{Y}^\top(\Delta\mathbf{Q}^t + \Delta\mathbf{K}^t) - \mathbf{B}\left((\mathbf{Q}^t)^\top\Delta\mathbf{Q}^t + (\mathbf{K}^t)^\top\Delta\mathbf{K}^t\right)\right]$$

We group and finalize to get a concluding formulation:

$$||\mathbf{A}^t_{\text{adv}} - \mathbf{A}^t||_2 \approx \frac{\exp\left(-\frac{\|\mathbf{Q}^t\|^2 + \|\mathbf{K}^t\|^2}{2}\right)}{m}\left\|\left(\mathbf{Y} - \mathbf{B}\mathbf{Q}^t\right)^\top\Delta\mathbf{Q}^t + \left(\mathbf{Y} - \mathbf{B}\mathbf{K}^t\right)^\top\Delta\mathbf{K}^t\right\|_2$$
$$= \gamma\left\|\left(\mathbf{Y} - \mathbf{B}\mathbf{Q}^t\right)^\top\Delta\mathbf{Q}^t + \left(\mathbf{Y} - \mathbf{B}\mathbf{K}^t\right)^\top\Delta\mathbf{K}^t\right\|_2$$

where the variables' definitions are:

- $\mathbf{B} = \sum_{i=0}^{k}\exp\left(\omega_i^\top(\mathbf{Q}^t + \mathbf{K}^t)\right)$,
- $\mathbf{Y} = \sum_{i=0}^{k}\exp\left(\omega_i^\top(\mathbf{Q}^t + \mathbf{K}^t)\right)\omega_i$,

- $\gamma = \frac{\exp\left(-\frac{\|\mathbf{Q}^t\|^2+\|\mathbf{K}^t\|^2}{2}\right)}{m}$.

Therefore, we get

$$||\mathbf{A}^t_{adv} - \mathbf{A}^t||_2 \geq \gamma \left\| \left[\left(\mathbf{Y} - \mathbf{B}\mathbf{Q}^t\right)^\top \mathbf{W}^t_Q + \left(\mathbf{Y} - \mathbf{B}\mathbf{K}^t\right)^\top \mathbf{W}^t_K \right] \delta_t \right\|_2, \tag{8}$$

The proof is complete.

I.3 Proof of Theorem 3.2

Proof:

$$\mathcal{L}^{adv}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} ||g(f(x^{adv}_{i_1})) - x^{adv}_{i_2}||^2$$

$$\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}[||g(f(x^{adv}_{i_1})) - x^{adv}_{i_2}||_2 + c_{\text{rec}} - c_{\text{rec}}] \\
&\geq \frac{1}{N}\sum_{i=1}^{N}[||g(f(x^{adv}_{i_1})) - x^{adv}_{i_2}||^2 + ||g(f_g(x_{i_2})) - x_{i_2}||^2 - c_{\text{rec}}] \\
&= \frac{1}{N}\sum_{1}^{N}[||g(f(x^{adv}_{i_1})) - x^{adv}_{i_2} + g(f(x_{i_1})) - g(f(x_{i_1}))||^2 + ||g(f_g(x_{i_2})) - x_{i_2}||^2 - c_{\text{rec}}] \\
&\geq \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{2}||g(f(x^{adv}_{i_1})) - x^{adv}_{i_2} + g(f(x_{i_1})) - g(f(x_{i_1})) + g(f_g(x_{i_2})) - x_{i_2}||^2 - c_{\text{rec}}] \\
&\geq \frac{1}{2N}\sum_{i=1}^{N}[||g(f(x^{adv}_{i_1})) - x^{adv}_{i_2} + g(f(x_{i_1})) - g(f(x_{i_1})) + g(f_g(x_{i_2})) - x_{i_2}||^2 - 2c_{\text{rec}}] \\
&\geq \frac{1}{2N}\sum_{i=1}^{N}[||g(f(x_{i_1})) - x_{i_2}||^2 + ||g(f(x^{adv}_{i_1})) - g(f(x_{i_1}))||^2 + ||g(f_g(x_{i_2}) - x^{adv}_{i_2}||^2 - 2c_{\text{rec}}] \\
&\geq \frac{1}{2}\mathcal{L}_{\text{rec}} + \frac{1}{2N}\sum_{i=1}^{N}[||g(f(x^{adv}_{i_1})) - g(f(x_{i_1}))||^2 + ||g(f_g(x_{i_2}) - x^{adv}_{i_2}||^2 - 2c_{\text{rec}}].
\end{aligned}$$

Since we have $||g(f_g(x_{i_2}) - x^{adv}_{i_2}||^2 = ||g(f_g(x_{i_2}) - x_{i_2} + \delta||^2$, and $||\delta||_2$ is tiny, it leads $||g(f_g(x_{i_2}) - x_{i_2} + \delta||^2 \geq ||g(f_g(x_{i_2}) - x_{i_2})||^2 - ||\delta||^2$. Following the previous study (Cao et al., 2022), the decoder $g(\cdot)$ in a MAE can be represented as an interpolation of the encoder output, denoted by $g(f(x)) = \mathbf{A}_{\text{dec}}\mathbf{V}_{\text{enco}}$. A_{dec} represents the interpolation weights, and $\mathbf{V}_{\text{enco}}$ is the encoder output.

$$\mathcal{L}^{adv}_{\text{rec}} \geq \frac{1}{2}\mathcal{L}_{\text{rec}} + \frac{1}{2N}\sum_{i=1}^{N}[||\mathbf{A}^{dec}_{adv_i}\mathbf{V}^{enco}_{adv_i} - \mathbf{A}^{dec}_i\mathbf{V}^{enco}_i||^2 - ||\delta||^2 - c_{\text{rec}}]$$

$$\begin{aligned}
&= \frac{1}{2}\mathcal{L}_{\text{rec}} + \frac{1}{2N}\sum_{i=1}^{N}[||\mathbf{A}^{dec}_{adv_i}\mathbf{V}^{enco}_{adv_i} - \mathbf{A}^{dec}_i\mathbf{V}^{enco}_i + \mathbf{A}^{dec}_i\mathbf{V}^{enco}_{adv_i} - \mathbf{A}^{dec}_i\mathbf{V}^{enco}_{adv_i}||^2 - ||\delta||^2 - c_{\text{rec}}] \\
&\geq \frac{1}{2}\mathcal{L}_{\text{rec}} + \frac{1}{2N}\sum_{i=1}^{N}[||\mathbf{V}^{enco}_{adv_i}(\mathbf{A}^{dec}_{adv_i} - \mathbf{A}^{dec}_i)||^2 + ||\mathbf{A}^{dec}_i(\mathbf{V}^{enco}_i - \mathbf{V}^{enco}_{adv_i})||^2 - ||\delta||^2 - c_{\text{rec}}].
\end{aligned}$$

Since $||\mathbf{A}_i^{dec}(\mathbf{V}_i^{enco} - \mathbf{V}_{adv_i}^{enco})||^2 \geq ||\delta||^2$ and a ratio constant C_A, we can get the following equation:

$$\mathcal{L}_{\text{rec}}^{adv} \geq \frac{1}{2}\mathcal{L}_{\text{rec}} + \frac{1}{2N}\sum_{i=1}^{N}[C_A||(\mathbf{A}_{adv_i}^{dec} - \mathbf{A}_i^{dec})||^2 - c_{\text{rec}}].$$

Based on Assumption 2, there exists a constant H such that $\left\|\mathbf{A}_{\text{adv},i}^{\text{dec}} - \mathbf{A}_i^{\text{dec}}\right\|^2 \approx \frac{H}{T}\sum_{t=1}^{T}\left\|\mathbf{A}_{\text{adv},i,t}^{\text{dec}} - \mathbf{A}_{i,t}^{\text{dec}}\right\|^2$. T represents the number of layers in the decoder, and $\mathbf{A}_{adv_{i,t}}^{dec}$ and $\mathbf{A}_{i,t}^{dec}$ are the attention matrices at the t-th decoder layer corresponding to i-th adversarial example and the clean example, respectively. Therefore, we can conclude the proof:

$$\mathcal{L}_{\text{rec}}^{adv} \geq \frac{1}{2}\mathcal{L}_{\text{rec}} + \frac{1}{2NT}\sum_{t=1}^{T}\sum_{i=1}^{N}[HC_A||(\mathbf{A}_{adv_{i,t}}^{dec} - \mathbf{A}_{i,t}^{dec})||^2 - c_{\text{rec}}].$$

The proof is complete.

I.4 Proof of Theorem 4.1

Proof:

We assume $\mathcal{L}_{\text{rec}}$ function is convex at the area $[x-\delta, x+\delta]$. For fixed mask U_0, we can get the following inequality:

$$\begin{aligned}
&\mathcal{L}_{\text{rec}}(x_{adv}, U_0) \leq \mathcal{L}_{\text{rec}}(x, U_0) + < \nabla_x L(x, U_0), x_{adv} - x > \\
&\iff \mathcal{L}_{\text{rec}}(x_{adv}, U_0) - L_{\text{rec}}(x, U_0) \leq < \nabla_x L(x, U_0), x_{adv} - x > .
\end{aligned}$$

We define $y_{adv}^1 = x_{adv} - \nabla_x \mathcal{L}(x_{adv}, U_0)$,

$$\begin{aligned}
&\iff \mathcal{L}_{\text{rec}}(x_{adv}, U_0) - \mathcal{L}_{\text{rec}}(x, U_0) \leq < \frac{x_{adv} - y_{adv}^1}{\lambda}, x_{adv} - x > \\
&\Rightarrow \mathcal{L}_{\text{rec}}(x_{adv}, U_0) - \mathcal{L}_{\text{rec}}(x, U_0) \leq \frac{< x_{adv} - y_{adv}^1, x_{adv} - x >}{\lambda} \\
&= \frac{(x_{adv})^2 - x_{adv}x - x_{adv}y_{adv}^1 + y_{adv}^1 x}{\lambda} \\
&= \frac{2(x_{adv})^2 - 2x_{adv}x - 2x_{adv}y_{adv}^1 + 2y_{adv}^1 x}{2\lambda} \\
&\frac{(X_{adv})^2 - 2x_{adv}x + (x_{adv})^2 - 2y_{adv}x_{adv}^1 + x^2 - x^2 + (y_{adv}^1)^2 - (y_{adv}^1)^2}{2\lambda} \\
&= \frac{||x_{adv} - x||_2^2 + ||x_{adv} - y_{adv}^1||_2^2 - ||y_{adv}^1 - x||_2^2}{2\lambda}.
\end{aligned}$$

Therefore, we can get the inequality:

$$\mathcal{L}_{\text{rec}}(x_{adv}, U_0) - \mathcal{L}_{\text{rec}}(x, U_0) \leq \frac{||x_{adv} - x||_2^2 - ||y_{adv}^1 - x||_2^2}{2\lambda} + \frac{\lambda}{2}||\nabla_X \mathcal{L}(x_{adv}, U_0)||_2^2.$$

Since $||y_{adv}^1 - x||_2^2 \leq ||clip(y_{adv}^1, \eta) - x||_2^2$, and η is the clipping threshold, we define $x_{adv}^1 = clip(y_{adv}^1, \eta)$ and get a derivation:

$$\mathcal{L}_{\text{rec}}(x_{adv}, U_0) - \mathcal{L}_{\text{rec}}(x, U_0) \leq \frac{||x_{adv} - x||_2^2 - ||x_{adv}^1 - x||_2^2}{2\lambda} + \frac{\lambda}{2}||\nabla_x \mathcal{L}(x_{adv}, U_0)||_2^2.$$

As the same theory, the law is written as the following formulation:

$$\begin{cases} s=0 \\ \mathcal{L}_{\text{rec}}(x_{adv},U_0)-\mathcal{L}_{\text{rec}}(x,U_0)\leq \frac{||x_{adv}-x||_2^2-||x_{adv}^1-x||_2^2}{2\lambda}+\frac{\lambda}{2}||\nabla_X\mathcal{L}(x_{adv}^1,U_0)||_2^2 \\ s=1 \\ \mathcal{L}_{\text{rec}}(x_{adv}^1,U_0)-\mathcal{L}_{\text{rec}}(x,U_0)\leq \frac{||x_{adv}-x||_2^2-||x_{adv}^2-x||_2^2}{2\lambda}+\frac{\lambda}{2}||\nabla_x\mathcal{L}(x_{adv}^2,U_0)||_2^2 \\ s=2 \\ \mathcal{L}_{\text{rec}}(x_{adv}^2,U_0)-\mathcal{L}_{\text{rec}}(x,U_0)\leq \frac{||x_{adv}-x||_2^2-||x_{adv}^3-x||_2^2}{2\lambda}+\frac{\lambda}{2}||\nabla_x\mathcal{L}(x_{adv}^3,U_0)||_2^2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \ldots\ldots \\ s=S \\ \mathcal{L}_{\text{rec}}(x_{adv}^S,U_0)-\mathcal{L}_{\text{rec}}(x,U_0)\leq \frac{||x_{adv}-x||_2^2-||x_{adv}^{S+1}-x||_2^2}{2\lambda}+\frac{\lambda}{2}||\nabla_x\mathcal{L}(x_{adv}^S,U_0)||_2^2 \end{cases}$$

We sum all the above terms as follows:

$$\sum_{s=0}^{S}(\mathcal{L}_{\text{rec}}(x_{adv}^s,U_0)-\mathcal{L}_{\text{rec}}(x,U_0))\leq \frac{||x_{adv}-x||_2^2-||x_{adv}^{S+1}-x||_2^2}{2\lambda}+\frac{\lambda}{2}\sum_{s=0}^{S}||\nabla_x\mathcal{L}(x_{adv}^s,U_0)||_2^2$$
$$\leq \frac{||x_{adv}-x||_2^2}{2\lambda}+\frac{\lambda}{2}\sum_{s=0}^{S}||\nabla_x\mathcal{L}(x_{adv}^s,U_0)||_2^2$$

The average loss upper bound $\frac{1}{S+1}\sum_{s=0}^{S}(\mathcal{L}_{\text{rec}}(x_{adv}^s,U_0)-\mathcal{L}_{\text{rec}}(x,U_0))$ is given as the following inequality:

$$\frac{1}{S+1}\sum_{s=0}^{S}(\mathcal{L}_{\text{rec}}(x_{adv}^s,U_0)-\mathcal{L}_{\text{rec}}(x,U_0))\leq \frac{||x_{adv}-x||_2^2}{2\lambda(S+1)}+\frac{\lambda}{2(S+1)}\sum_{s=0}^{S}||\nabla_x\mathcal{L}(x_{adv}^s,U_0)||_2^2.$$

Due to the weak convexity of $\mathcal{L}_{\text{rec}}(\cdot)$, it is evident that $\frac{1}{S+1}\sum_{s=0}^{S}\mathcal{L}_{\text{rec}}(x_{\text{adv}}^s,U_0)\geq\mathcal{L}_{\text{rec}}\left(\frac{1}{S+1}\sum_{s=0}^{S}X_{\text{adv}}^s,U_0\right)$ and we can obtain the formulation as follows:

$$\mathcal{L}_{\text{rec}}\left(\frac{1}{S+1}\sum_{s=0}^{S}x_{\text{adv}}^s,U_0\right)-\mathcal{L}_{\text{rec}}(x,U_0))\leq \frac{||x_{adv}-x||_2^2}{2\lambda(S+1)}+\frac{\lambda}{2(S+1)}\sum_{s=0}^{S+1}||\nabla_x\mathcal{L}(x_{adv}^s,U_0)||_2^2.$$

For any mask U_e and $e\in[1,E]$, we can obtain similar results. Thus, we can draw the results:

$$\frac{1}{(1-\rho)E}\sum_{e=1}^{E}[\mathcal{L}_{\text{rec}}\left(\frac{1}{S+1}\sum_{s=0}^{S}x_{\text{adv}}^s,U_e\right)-\mathcal{L}_{MAE}(x,U_e))]\leq\frac{1}{(1-\rho)E}\sum_{e=1}^{E}[\frac{||x_{adv}-x||_2^2}{2\lambda(S+1)}$$
$$+\frac{\lambda}{2(S+1)}\sum_{s=0}^{S}||\nabla_x\mathcal{L}(x_{adv}^s,U_e)||_2^2].$$

For a fixed masked rate p, the proof has been completed. $\mathcal{L}_{\text{rec}}(X,U_e))\approx\mathcal{L}_{\text{rec}}^*(X)$ for any $e\in[1,E]$, and we can conclude the proof:

$$\frac{1}{(1-\rho)E}\sum_{e=1}^{E}[\mathcal{L}_{\text{rec}}\left(\frac{1}{S+1}\sum_{s=0}^{S}x_{\text{adv}}^s,U_e\right)-\mathcal{L}_{\text{rec}}^*(x)]\leq\frac{1}{(1-\rho)E}\sum_{e=1}^{E}[\frac{||x_{adv}-x||_2^2}{2\lambda(S+1)}$$
$$+\frac{\lambda}{2(S+1)}\sum_{s=0}^{S+1}||\nabla_x\mathcal{L}(x_{adv}^s,U_e)||_2^2].$$

The proof is complete.

J Details of Mask Autoencoder

MAE: For CIFAR-10, CIFAR-100, and SVHN, we use Base-MAE (He et al., 2022), setting the image size to 32 and the patch size to 4, while leaving the other parameters unchanged. For ImageNet, we directly used Large-MAE.

For MAE training on CIFAR-10, CIFAR-100, and ImageNet, we use A100 GPUs, training for 2000 epochs with the learning rate of $1e^{-3}$, followed by an additional 1000 epochs with the learning rate of $1.5e^{-4}$. Our training batch size is 64. For ImageNet, we directly use the checkpoint provided by the author.

MaskDiT: For SVHN, CIFAR100, and CIFAR10, the network first divides the 32x32 CIFAR-10 image into non-overlapping 8x8 patches, with each patch sized 4x4, resulting in a total of 64 patches. Each patch is mapped to a 128-dimensional embedding space through a linear projection layer, and learnable positional encodings are added. During training, 50% of the patches are randomly masked, and only the unmasked patches are fed into an encoder composed of 6 Transformer blocks, each consisting of multi-head self-attention and a feed-forward network. Subsequently, the encoded unmasked patches are concatenated with learnable mask tokens and passed into a decoder composed of 3 lightweight Transformer blocks. Finally, a linear projection layer maps the decoded patches back to the 4x4x3 patch space, completing image reconstruction and denoising score prediction. For ImageNet, we use patch size as 16× 16 and other parameter are same with original work (Zheng et al., 2024). Both network training process and parameters are aligned.

J.1 Details of Robust fine-tuning

For CIFAR-10, CIFAR-100, and SVHN, AMRM-Pure$_{\text{MAE}}$ is fine-tuned for 100 epochs, while for ImageNet, it is fine-tuned for 3 epochs. In contrast, MaskDiT is fine-tuned for 25 epochs on CIFAR-10, CIFAR-100, and SVHN, and 1 epoch on ImageNet.

J.2 Details of Experiment

J.2.1 Hyperparameters in Denoising Process

All hyperparameters related to the denoising process are encompassed in Table 18 and 19, covering the number of denoising iterations, step size, mask rate, and patch size for each dataset. The asterisk (*) indicates that the step size decays by 0.5 after more than half of the iterations have been completed.

Table 18: Hyperparameters of Step Size and Patch Size

Dataset	Step Size				Patch Size			
	MAE	RMAE	MaskDiT	RMaskDiT	MAE	RMAE	MaskDiT	RMaskDiT
CIFAR-10	1*	1*	1	1	4	4	4	4
CIFAR-100	1	1	1	1	4	4	4	4
SVHN	1*	1	1	1	4	4	4	4
ImageNet	1*	1	1	1	16	16	16	16

Table 19: Hyperparameters of Mask Rate and Iteration Numbers

Dataset	Mask Rate				Iterations			
	MAE	RMAE	MaskDiT	RMaskDiT	MAE	RMAE	MaskDiT	RMaskDiT
CIFAR-10	0.25	0.25	0.50	0.50	150	90	25	15
CIFAR-100	0.30	0.25	0.50	0.50	100	65	20	20
SVHN	0.30	0.25	0.50	0.50	150	80	20	20
ImageNet	0.20	0.25	0.30	0.30	150	20	25	20