Enhancing Variational Autoencoders with Smooth Robust Latent Encoding

Hyomin Lee[1*†]   Minseon Kim[2*]   Sangwon Jang[3]   Jongheon Jeong[1]   Sung Ju Hwang[3,4]

[1]Korea University, [2]Microsoft, [3]KAIST, [4]DeepAuto.ai

{lhm1024, jonghj}@korea.ac.kr, minseonkim@microsoft.com,
{sangwon.jang, sungju.hwang}@kaist.ac.kr

Abstract

Variational Autoencoders (VAEs) have played a key role in scaling up diffusion-based generative models, as in Stable Diffusion, yet questions regarding their robustness remain largely underexplored. Although adversarial training has been an established technique for enhancing robustness in predictive models, it has been overlooked for generative models due to concerns about potential fidelity degradation by the nature of trade-offs between performance and robustness. In this work, we challenge this presumption, introducing Smooth Robust Latent VAE (SRL-VAE), a novel adversarial training framework that boosts both generation quality and robustness. In contrast to conventional adversarial training, which focuses on robustness only, our approach smooths the latent space via adversarial perturbations, promoting more generalizable representations while regularizing with originality representation to sustain original fidelity. Applied as a post-training step on pre-trained VAEs, SRL-VAE improves image robustness and fidelity with minimal computational overhead. Experiments show that SRL-VAE improves both generation quality, in image reconstruction and text-guided image editing, and robustness, against Nightshade attacks and image editing attacks. These results establish a new paradigm, showing that adversarial training, once thought to be detrimental to generative models, can instead enhance both fidelity and robustness.

1. Introduction

Variational Autoencoders (VAEs) [12] have been employed as a compressor in the success of latent generative models [15, 19, 22, 24], which have demonstrated surprising capabilities in generating high-quality images. The VAEs compress high-dimensional images into a latent space that retains semantic and structural information, which is continuous [24] or discrete [4] space for high-quality gener-

(a) Diffusion-based generative process of clean image

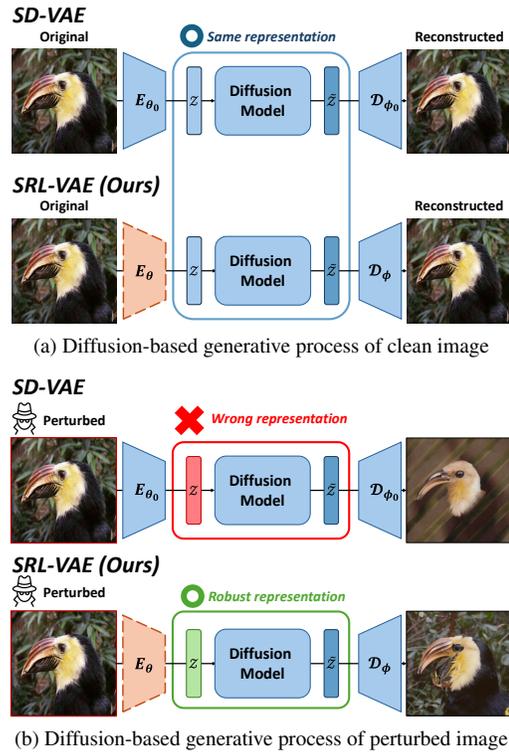(b) Diffusion-based generative process of perturbed image

Figure 1. **Concept figure of SRL-VAE.** Compared to SD-VAE, SRL-VAE maintains similar representations for clean examples while achieving robust representation against perturbed examples.

ative modeling. Despite their effectiveness as a compressor in generative models, prior work has largely overlooked the representational role of VAEs, primarily focusing on generative aspects to improve performance by proposing architectures [22], training objectives [19, 37], or regularization [38]. However, obtaining an effective compressor is one of the key components to achieving higher-quality generations while also ensuring robustness with efficient computational costs.

To obtain effective VAEs for both higher fidelity and better robustness, representation space of VAEs needs to be robust and capable of capturing better structural latent. We

---

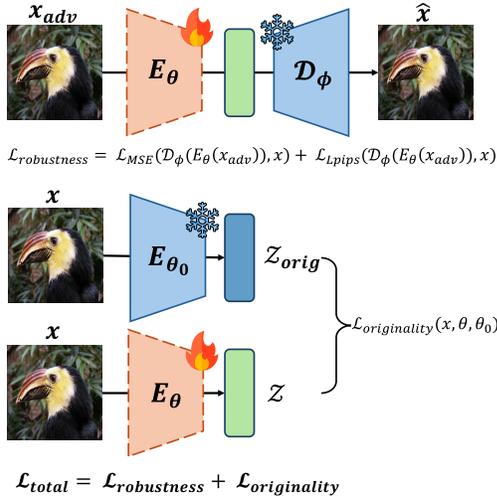*Equal contribution
†Work done during an internship at KAIST

Figure 2. **Training objective of Smooth Robust Latent Variational Autoencoders (ours).** A novel adversarial training approach in the latent space of VAE with originality regularization.

were inspired by adversarial training [20, 39] to build robust representations. Adversarial training [20, 39] has initially been recognized as an effective method for improving adversarial robustness in predictive models against adversarial attacks, particularly in classification tasks. In fact, adversarial training builds robust and smooth representations [14] so that it could have better generalization [10, 36] from leveraging a min-max formulation. However, adversarial training in generative modeling remains largely unexplored, partly because generative models have traditionally prioritized fidelity and diversity over robustness. Moreover, adversarial training is challenging to optimize and often leads to large performance degradation where trade-offs are clear [2, 39]. As a result, the adoption of adversarial training in VAEs and similar generative models has been limited, as it is frequently perceived as a significant trade-off.

In this work, we introduce a novel adversarial training framework for VAEs that enhances both generation quality and robustness by constructing smooth latent space. Unlike conventional adversarial training, which primarily targets predictive robustness, our approach leverages adversarial perturbations to smooth the latent space and promote more generalizable representations of VAEs (Figure 1). Our approach consists of two key steps: (1) maximizing the VAE loss to introduce adversarial perturbations in the latent space, exposing the model to challenging variations, and (2) minimizing both the VAE loss and an originality loss to preserve the original representation structure while building smooth latent space, ensuring a stable training and robustness as shown in Figure 2. Furthermore, our approach is applied as a post-training step on pre-trained VAEs, requiring only a small amount of additional computational resources,

making it an efficient and practical solution for improving generative models.

By bridging adversarial training and generative modeling, our work introduces a new perspective on the importance of obtaining robust, high-quality representations of VAEs in generative models. Our method enables VAEs to generate outputs of comparable or, in some cases, even higher quality (Figure 1a) while being extremely effective against various types of adversarial attacks during the generation process (Figure 1b). This demonstrates that adversarial training is not merely a defensive mechanism that sacrifices performance but rather a powerful strategy for enhancing generative models with robustness, paving the way for future advancements in robust generative learning.

The main contributions can be summarized as follows:

- Unlike prior adversarial training methods, which predominantly focus on classification models, we introduce the first adversarial training approach tailored for VAEs, demonstrating its ability to improve both generation quality and robustness simultaneously.

- We show that **adversarial training**, when combined with an originality loss, **fosters a smoother latent space, leading to more stable and generalizable representations,** which enhance image fidelity and deliver surprisingly strong performance against various types of attacks.

- Through extensive experiments, we demonstrate significant improvements in both fidelity and robustness across multiple tasks. Specifically, we evaluate image quality through image reconstruction and generation, and assess robustness by evaluation against adversarial attacks on text-guided image editing and adversarial poisoning attacks, establishing the effectiveness of our approach in enhancing both generative performance and robustness.

2. Related Works

Latent Generative Models Variational Autoencoders (VAEs) [12] are generative models that learn compact latent representations by regularizing the latent distribution through a Kullback–Leibler (KL) divergence term. While VAEs enable smooth interpolation and efficient sampling, they often produce blurry outputs due to limitations in the latent space. To address this, Vector Quantized VAE (VQ-VAE) [33] introduces a discrete codebook of embeddings, preventing latent space collapse and improving reconstruction quality. VQ-GAN [4] further enhances this framework with adversarial training, guiding the decoder towards sharper and more realistic outputs. Building on these approaches, Latent Diffusion Models (LDMs) [24] apply diffusion processes in a compressed latent space obtained from a pretrained autoencoder, significantly reducing computational costs while preserving high fidelity. Cross-attention mechanisms enable flexible conditional genera-

tion from various inputs, as demonstrated in Stable Diffusion [24]. Recent work also explores enhancing the interaction between autoencoders and diffusion models, addressing spectral properties of latent spaces [31] and introducing equivariance regularization for better generative performance [13]. However, prior works have largely overlooked the quality of latent representations in VAEs in terms of fidelity and robustness. We address this by applying adversarial training to enhance the latent space, improving both generation quality and robustness.

Adversarial Training Szegedy et al. [32] first revealed the vulnerability of deep neural networks (DNNs) to adversarial attacks, showing that imperceptible perturbations could mislead models. Goodfellow et al. [5] introduced adversarial training with the Fast Gradient Sign Method (FGSM), demonstrating that training on adversarial and clean samples improves robustness. Madry et al. [20] extended this with Projected Gradient Descent (PGD) adversarial training, formulating it as a minimax optimization problem. TRADES [39] further refined robustness by enforcing consistency between clean and adversarial samples via Kullback-Leibler divergence minimization. Recent works leveraged unlabeled data [2] or generative models [6] to enhance adversarial robustness by exposing diverse distribution. Beyond supervised settings, adversarial self-supervised learning (SSL) emerged as an alternative perspective to obtain robust representation [9, 11], using contrastive learning or self-supervised learning by introducing adversarial examples that maximize given losses without any class information. However, all these methods focus on building a robust representation for predictive models to have robust decision boundaries, which did not consider generative models. Unlike prior works, our approach suggests adversarial training for VAEs, demonstrating that it can enhance both generation quality and robustness.

3. Variation Autoencoders with Smooth Robust Latent Encoding

In this section, we first revisit the preliminary of Variational Autoencoder (VAE) and adversarial training (AT) in section 3.1. Then, we propose our smooth robust latent VAE approach with theoretical motivation in Section 3.2.

3.1. Preliminary

Variational autoencoder (VAE) A Variational autoencoder (VAE) is a latent space compressor that encodes high-dimensional image data into a lower-dimensional latent space. Given an input image x, the encoder $E_\theta(x)$ compresses the image to a latent variable z, and the decoder $D_\phi(z)$ reconstructs x as $\hat{x} = D_\phi(z)$. We employ

a VAE [24] that is optimized primarily for high-fidelity reconstruction. The training objective is defined as follows:

$$\mathcal{L}_{\texttt{VAE}}(x) = \mathcal{L}_{\texttt{rec}}(x, \hat{x}) + \mathcal{L}_{\texttt{gan}}(\hat{x}) + \mathcal{L}_{\texttt{reg}}(x), \quad (1)$$

where $\mathcal{L}_{\texttt{rec}}$ combines pixel-wise loss (L_1 distance loss or L_2 distance loss) and perceptual loss (LPIPS loss). LPIPS loss is a similarity loss of learned perceptual image patches that calculates the similarity distance based on features extracted from a pre-trained VGG model [30]. The $\mathcal{L}_{\texttt{gan}}$ is an adversarial loss that encourages the generation of more realistic outputs by leveraging a discriminator network. Lastly, $\mathcal{L}_{\texttt{reg}}$ is a regularization term which is following KL-divergence:

$$\mathcal{L}_{\texttt{reg}}(x) = \mathcal{L}_{\texttt{KL}}(q_\theta(z|x)||p(z)), \quad (2)$$

where $p(z)$ is the prior distribution, and set to a standard normal distribution $\mathcal{N}(0, I)$. KL regularization in recent latent diffusion models ensures smooth sampling and interpolation by encouraging a well-structured latent space.

Adversarial training Adversarial training is a technique for obtaining robust models against adversarial attacks by solving a min-max optimization problem. First, we define an adversarial perturbation, δ, which is applied to an input x. Then, the min-max optimization problem is formulated with 1) generating the perturbation δ by maximizing the model's given loss $\mathcal{L}$, while 2) simultaneously minimizing the training loss under this perturbation. Several approaches [34, 39] exist for generating effective adversarial perturbation in the min-max optimization problem. Here, we employ projected gradient descent attacks [20] that maximize the training loss, as follows.

$$\delta^{t+1} = \Pi_{B(0,\epsilon)}\Big(\delta^t + \alpha \texttt{sign}\big(\nabla_{\delta^t}\mathcal{L}\big)\Big), \quad (3)$$

where $B(0, \epsilon)$ is the ℓ_∞ norm-ball of radius ϵ, Π is the projection function to the norm-ball, α is the step size of the attacks, and $\texttt{sign}(\cdot)$ is the sign of the vector. Also, δ represents the perturbations accumulated by $\alpha \texttt{sign}(\cdot)$ over multiple iterations t. Then, minimization is defined as follows,

$$\min_\omega \mathbb{E}_{x \sim \mathcal{D}} \big[\mathcal{L}(x + \delta^t)\big], \quad (4)$$

where ω is a parameter of model f, x is training samples from dataset $\mathcal{D}$ and $\mathcal{L}$ is training objectives which employ perturbed samples $x + \delta$.

3.2. Smooth Robust Latent VAE

Theoretical motivation Adversarial training has been widely used in predictive models to improve robustness against adversarial attacks, yet its application in generative models remains under-explored. In predictive tasks, robustness is achieved by ensuring that perturbed inputs $x + \delta$ produce outputs similar to those of the original inputs x.
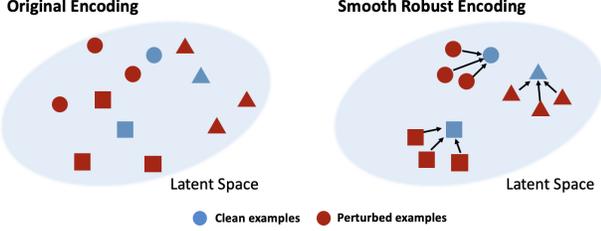
Figure 3. **Concept of smooth latent space.** A smooth latent space ensures that perturbed examples are mapped closely to their original counterparts, enabling the VAE to extract robust features.

This is typically enforced by a Lipschitz constraint, which guarantees that small changes within an ϵ-norm ball result in only minor variations in the output, thereby creating a smooth representation space. Motivated by this, we propose that applying adversarial training to VAEs can similarly promote a smooth and well-structured latent space, leading to improved generation quality and robustness (Figure 3).

Furthermore, in latent-based generative models, the encoder acts as an information bottleneck, compressing high-dimensional inputs into latent codes $z = E_\theta(x)$. Motivated by the Information Bottleneck (IB) principle, we believe that an optimal latent representation should capture only the essential features for accurate reconstruction while discarding noisy features and clearly distinguishing different inputs. In VAEs, this results in a latent space that is both expressive and compact. By applying adversarial training, our approach forces the encoder to extract only the crucial features for high-quality reconstruction and to maintain clear separations among examples. Specifically, adversarial training encourages each input's latent representation to be confined within a secure ϵ-ball, creating a large margin between different examples. This leads to a tighter information bottleneck and a more structured latent space, ultimately enhancing both image fidelity and generalization.

Smooth Robust Latent VAE   We propose *Smooth Robust Latent VAE* (SRL-VAE), which enhances latent representation quality by applying adversarial training to the encoder. Additionally, to ensure compatibility with pre-trained diffusion models such as the UNet in latent diffusion models (LDMs), our approach emphasizes preserving the original latent structure while refining it for improved performance.

We formulate a min-max optimization framework for VAEs. In particular, to generate adversarial perturbations, we define the maximization step using a projected gradient descent (PGD) formulation as follows:

$$\delta^{t+1} = \Pi_{B(0,\epsilon)}\Big(\delta^t + \alpha\, \texttt{sign}\big(\nabla_{\delta^t}(\mathcal{L}_{\texttt{MSE}}(D_\phi(E_\theta(x+\delta^t)), x)$$
$$+ \lambda \cdot \mathcal{L}_{\texttt{LPIPS}}(D_\phi(E_\theta(x+\delta^t)), x))\big)\Big),$$
$$(5)$$

where $\mathcal{L}_{\texttt{MSE}}$, and $\mathcal{L}_{\texttt{LPIPS}}$ is L_2 distance loss and LPIPS perceptual loss between adversarial examples and original ex-

amples, respectively.

Then, the minimization step is formulated to ensure that the outputs from the perturbed examples are similar to the original examples, while preserving the original latent space distribution. This is expressed as follows:

$$\mathcal{L}_{\texttt{total}} = \alpha \mathcal{L}_{\texttt{orig}}(x, \theta, \theta_0) + \mathcal{L}_{\texttt{MSE}}(D_\phi(E_\theta(x_{\texttt{adv}})), x)$$
$$+ \lambda \mathcal{L}_{\texttt{LPIPS}}(D_\phi(E_\theta(x_{\texttt{adv}})), x)$$
$$(6)$$

where $x_{\texttt{adv}} = x + \delta$ from equation 5, and α, λ control the balance between latent consistency and reconstruction quality.

The originality loss $\mathcal{L}_{\texttt{orig}}$ acts as regularization to preserve the latent distribution of clean inputs by minimizing the difference from the pre-trained encoder:

$$\mathcal{L}_{\texttt{orig}}(x, \theta, \theta_0) = \|\mu - \mu_{\texttt{orig}}\|_2^2 + \|\log \sigma^2 - \log \sigma^2_{\texttt{orig}}\|_2^2,$$
$$(7)$$

where μ and σ^2 represent the mean and variance of the latent distribution produced by the current encoder parameterized by θ, and $\mu_{\texttt{orig}}$ and $\sigma^2_{\texttt{orig}}$ represent the mean and variance produced by the pre-trained encoder parameterized by θ_0, respectively, when given input x. By minimizing this objective, the encoder learns robust latent representations that maintain the original latent distribution, ensuring compatibility with downstream components and enabling stable, high-fidelity generation with enhanced robustness.

4. Experiment

In this section, we first describe our experimental setup, including datasets, training details, and evaluation details in Section 4.1. We then present the image quality performance of our SRL-VAE in Section 4.2, demonstrating both its reconstruction quality and diffusion generation quality. In Section 4.3, we evaluate the robustness of our latent space against various types of perturbations and different attacks in diffusion models. Lastly, we conduct ablation studies and analyze the latent space of SRL-VAE in Section 4.4.

4.1. Setup

Training details   We further fine-tune a pre-trained Stable Diffusion Variational Autoencoder (SD-VAE) on a subset of 100K images from the LAION-Aesthetic dataset [27], resized to 256×256 resolution. During fine-tuning, only the encoder is updated while keeping the decoder frozen to maintain compatibility with the pre-trained diffusion model. The model is optimized with a batch size of 20 for a total of 5K steps. For adversarial training, we use the Projected Gradient Descent (PGD) attack under an ℓ_∞ perturbation bound of $\epsilon = 8/255$ with 10 iterations per attack and a step size of 0.02. We apply the originality loss with a weight of $\alpha = 0.01$, selected through hyperparameter tuning.

Dataset	VAE	PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓
COCO	SD-VAE	23.68	0.74	**0.14**	8.79
	Ours	**24.46**	**0.76**	0.15	**7.92**
ImageNet	SD-VAE	23.53	0.73	**0.15**	3.71
	Ours	**24.48**	**0.74**	0.16	**3.09**

Table 1. **Quantitative evaluation of reconstruction quality on the MS-COCO and ImageNet validation sets.** SRL-VAE consistently outperforms the baseline VAE across PSNR, SSIM, and FID metrics, indicating superior fidelity.

Evaluation To assess the image quality of SRL-VAE, we evaluate both reconstruction quality and generation quality. For reconstruction quality, we use the MS-COCO [17] validation set (5,000 images) and the ImageNet [3] validation set (50,000 images), with all images resized to 256×256. We measure Fréchet Inception Distance (FID) [7], Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [35], and Learned Perceptual Image Patch Similarity (LPIPS) [40] to quantify the model's ability to accurately reconstruct images. For generation quality, we compare SRL-VAE with SD-VAE within the Diffusion Transformer [22] (DiT-B/2) framework. We train DiT models on ImageNet-1000k (1,280K images) for 10 epochs (50K steps) with a batch size of 256 and evaluate their performance using Inception Score (IS) [25] and FID.

To validate the robustness of SRL-VAE against adversarial perturbations, we conducted two experiments. First, we evaluated the model's resilience to adversarial attacks that target the training process to maliciously manipulate the diffusion model [18, 29]. Specifically, we tested Nightshade [29], which poisons a concept C so that it generates images resembling a destination concept A. To determine attack success, we measured the CLIP [23] similarity between the generated image and its generating prompt C, assessing how far the image deviates from the intended concept of C. If the CLIP score was lower than the threshold $\tau = 0.25$, we considered the attack successful. Second, we assessed the robustness of SRL-VAE against defensive perturbations from PhotoGuard [26], MIST [16] and Glaze [28], which prevent unauthorized image edits. In a realistic image-to-image editing scenario, we compared SRL-VAE and SD-VAE, measuring FID and CLIP similarity with the original generation results to evaluate robustness.

4.2. Smooth Robust Latent VAE

Image reconstruction quality To evaluate the effectiveness of our proposed SRL-VAE, we measure its performance on image reconstruction tasks compared to the baseline VAE in both COCO dataset [17] and ImageNet dataset (Table 1). Our SRL-VAE consistently achieves superior performance across most of the metrics compared to the original VAE. The higher PSNR and SSIM scores indi-
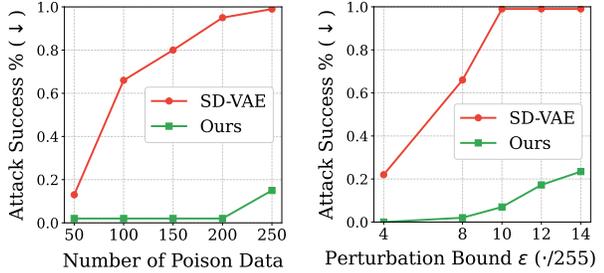
cate that images reconstructed by SRL-VAE preserve more structural and visual details, whereas the lower FID scores demonstrate enhanced perceptual realism of the generated images. These results collectively confirm that adversarial training within our SRL-VAE significantly enhances image fidelity, underscoring the improved quality and robustness of the learned latent representations.

Diffusion generation quality We further evaluate the diffusion generation capabilities with our SRL-VAE within the Diffusion Transformer (DiT) framework [22]. Our primary objective is to confirm that our adversarial training approach does not compromise generation performance, particularly in the diffusion process. The original DiT model with a standard SD-VAE achieves an IS of 12.49 and a FID of 91.54. In comparison, DiT with our SRL-VAE achieves slightly improved performance, with an IS of 12.87 and an FID of 91.27, demonstrating that our method does not degrade diffusion generation quality. This result highlights that integrating SRL-VAE into the diffusion generation process is both seamless and adaptable, preserving image quality while simultaneously providing additional benefits such as improved latent space representation and enhanced robustness against perturbations, as discussed in Section 4.3.
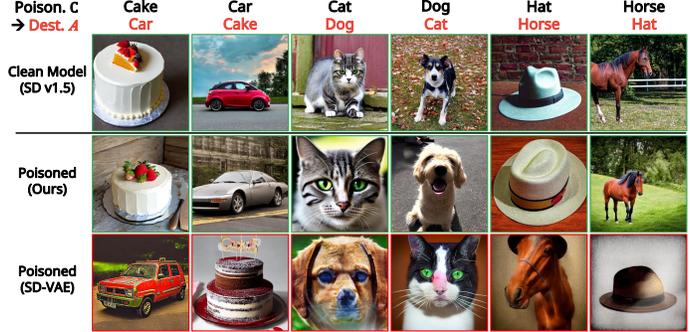
4.3. Robustness on Perturbations

In this section, we assess the robustness of SRL-VAE by integrating our encoder into existing frameworks and evaluating against two distinct categories of perturbation-based approaches. First, we test against the Nightshade attack [29], a malicious adversarial perturbation-based data poisoning technique designed to disrupt specific outputs of diffusion models by injecting a few poison samples into training data. Second, we evaluate robustness by measuring the neutralization scale against defensive perturbation methods such as PhotoGuard [26], Glaze [28], and Mist [16], which are initially designed to protect intellectual property by adding imperceptible perturbations. Our experiments demonstrate that our SRL-VAE effectively neutralizes both types of perturbation-based approaches.

Robustness in Nightshade malicious attack Nightshade [29] is a prompt-specific poisoning attack that can maliciously control generative outputs with only a small number of adversarial samples. To demonstrate the robustness of our method against this attack, we fine-tune a pre-trained diffusion model on 10K images from the LAION-Aesthetic while varying the poisoning ratio of Nightshade poisoned samples or perturbation bound ϵ. In Figure 4(a), we fixed $\epsilon = 8/255$ and varied the poisoning ratio. With our SRL-VAE, the diffusion model remained resistant to the attack, whereas the model using SD-VAE was easily attacked even at low poisoning ratios. In Figure 4(b), we fixed the number of poisoned samples at 100 and changed ϵ.

(a) Attack success rate depending on **number of poisoned data injected**

(b) Attack success rate depending on **perturbation bound** ϵ

(c) Qualitative examples generated by Nightshade attacked models.

Figure 4. **Robustness on Nightshade attack.** (a) and (b) demonstrate robustness evaluation against the Nightshade poisoning attack. SRL-VAE maintains low attack success rates across varying poisoning ratios and perturbation bounds. (c) Qualitative examples demonstrate that SRL-VAE preserves intended generation even under attacks.
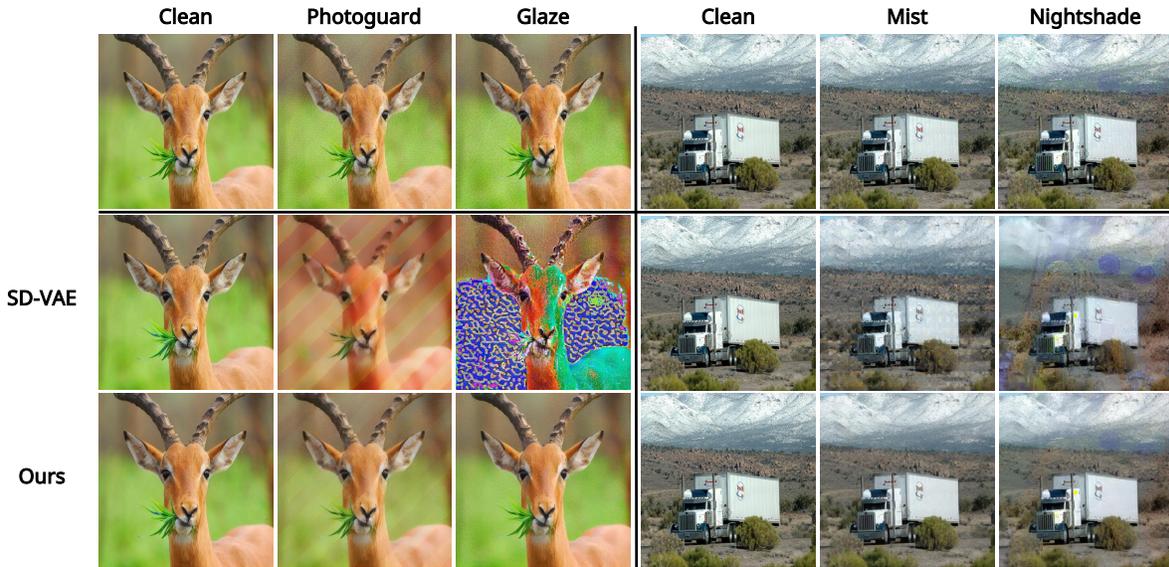

Figure 5. **Visual examples of reconstruction under various perturbations.** SD-VAE struggles to reconstruct images with added perturbed noise, whereas ours robustly handles both clean and various perturbed images.

Even under a higher perturbation bound (ϵ=15/255, which is clearly visible), SRL-VAE retained its robustness beyond its training bound. The qualitative examples demonstrate that our SRL-VAE successfully prevents a poisoned cake image from being transformed into a car, preserving its original appearance as a cake, as shown in Figure 4(c).

These results highlight the effectiveness of our approach in mitigating poisoning attacks without introducing additional overhead. While purification-based defenses [1, 8, 21, 41] can serve as a solution for poisoning attacks, applying them to every image in large datasets incurs high computational costs due to the difficulty of identifying poisoned images within the dataset. Moreover, with the rise of publicly available data and the trend of data sharing, the size of training datasets keeps growing, making it inefficient, sometimes nearly impossible, to purify every single image. In contrast, our method modifies only the VAE and does

VAE	Metric	Photoguard	MIST	Glaze
SD-VAE	FID ↓	221.1	146.1	86.40
	CLIP ↑	0.7231	0.7909	0.8410
Ours	FID ↓	**68.42**	**60.50**	**57.32**
	CLIP ↑	**0.8832**	**0.8933**	**0.9065**

Table 2. **Evaluation of image-to-image editing robustness under various perturbation defenses.** SRL-VAE significantly improves FID and CLIP scores across all methods, indicating better visual quality and semantic alignment.

not introduce any additional runtime overhead compared to purification-based defenses.

Robustness against various type of perturbations To demonstrate the robustness of our SRL-VAE against various types of adversarial perturbations, we evaluated its
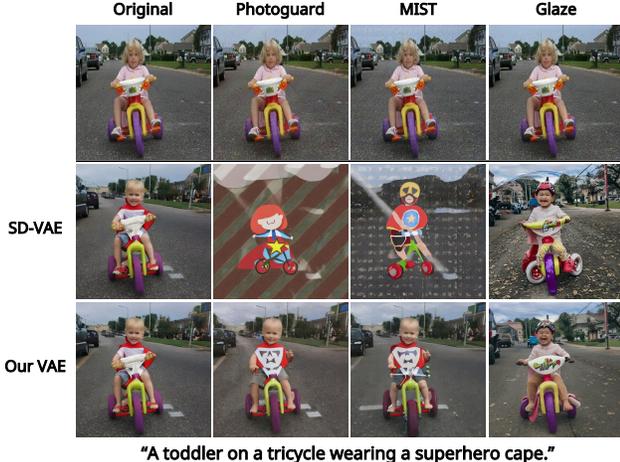
6

| Original | Photoguard | MIST | Glaze |

Figure 6. **Comparison of image-to-image editing results under defensive perturbations.** SRL-VAE produces valid and high-quality edited outputs based on the given prompts, while the baseline VAE fails to preserve the original semantics.

ability to reconstruct perturbed images processed by PhotoGuard [26], MIST [16], and Glaze [28]. These methods were originally devised to protect images via adversarial perturbations, we repurpose these methods to measure the extent to which these protections are neutralized, thus evaluating our VAE's latent space resilience. As shown in 5, the base SD-VAE struggles to reconstruct images with various perturbations, whereas our SRL-VAE successfully encodes both clean and perturbed images into its latent space, enabling accurate reconstructions. Subsequently, we further demonstrate SRL-VAE's robustness by showcasing its image-to-image editing performance on these same protected images. For this experiment, we constructed an editing dataset of 100 images from ImageNet, each resized to 512×512. As shown in Figure 6, the protection methods successfully disrupt the editing results of the original VAE, producing outputs that are significantly different from the source images. In contrast, SRL-VAE generates valid edited images based on the provided prompt "A toddler on a tricycle wearing a superhero cape", demonstrating strong robustness. Specifically, we measure the FID and CLIP similarity between the image-to-image results of the original and protected images. As shown in Table 2, SRL-VAE consistently outperforms the baseline VAE across all protection methods, achieving lower FID scores and higher CLIP similarity. These results demonstrate that SRL-VAE preserves better visual quality and semantic consistency, even when editing images protected by strong perturbation defenses.

4.4. Analysis

Ablation studies on loss component To understand the contributions of each loss component, we perform ablation studies employing different loss functions. Specifically, we compare three loss types using adversarial loss without

VAE Variant	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	rFID $\downarrow$
SD-VAE	23.68	0.74	0.14	8.79
Full SRL-VAE	24.46	0.76	0.15	7.92
+ w/o originality loss	26.55	0.78	0.23	15.46

Table 3. **Ablation study of SRL-VAE loss components.**

VAE	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	rFID $\downarrow$
SD-VAE	19.67	0.5906	0.6691	109.5
$\alpha = 0.1$	21.79	0.6562	0.5520	75.85
$\alpha = 0.01$	27.23	0.7635	0.3244	28.47
$\alpha = 0.001$	**28.38**	**0.7834**	**0.3064**	**19.90**

Table 4. **Ablation study of an α hyper-parameter.** Smaller α values improve the reconstruction quality of perturbed Photoguard [26] images, enhancing robustness.

originality regularization (Equation 8), and our proposed SRL-VAE.

$$\begin{aligned} \mathcal{L}_{\texttt{wo-originality}} = \ &\mathcal{L}_{\text{MSE}}(D_\phi(E_\theta(x_{\text{adv}})), x) \\ &+ \lambda \mathcal{L}_{\text{LPIPS}}(D_\phi(E_\theta(x_{\text{adv}})), x) \end{aligned} \quad (8)$$

The experimental results in Table 3 indicate that originality regularization significantly contributes to leverage the original performance of SD-VAE.

Ablation studies on hyper-parameter α In Equation 6, we regularize the originality loss using the hyperparameter α. α acts as a controller, regulating the influence of robustness during optimization. As α increases, the impact of robustness decreases in the overall objective function, leading to decreasing robustness, as shown in Table 4. However, to preserve generation performance on clean images, we set α to 0.01, achieving an optimal balance between high fidelity on clean images and robustness against perturbations. Moreover, originality loss plays a critical role in maintaining compatibility with pre-trained diffusion models, making it an important component of our approach.

Latent space analysis We analyze the latent space learned by our SRL-VAE using two analysis approaches, which are loss surface visualization and t-SNE visualization of latent distributions. First, we visualize the loss surfaces of SD-VAE and SRL-VAE by applying perturbations along two random directions on the input image. For each perturbed input, we compute the mean squared error (MSE) between the latent representations of the perturbed and clean inputs and normalize the loss values for comparison. As shown in Figure 7, the SRL-VAE exhibits a smoother loss landscape compared to SD-VAE, indicating enhanced robustness and improved latent space smoothness through adversarial training. In other words, a smoothness
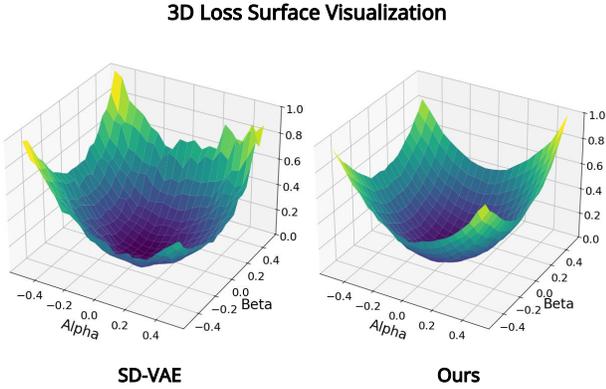
3D Loss Surface Visualization

Figure 7. **Comparison of the loss surfaces of SD-VAE and SRL-VAE under encoder input perturbations.** SRL-VAE shows a smoother and more stable loss landscape, highlighting its improved robustness and smooth latent representation.

of latent space ensures that perturbed examples are adequately mapped to the similar region as their corresponding original examples in the latent space.

Additionally, we use t-SNE visualization to analyze latent samples from both original inputs and Gaussian-noise-added inputs, allowing us to directly assess the robustness of distributions in the VAE's latent space. As shown in Figure 8, the visualization reveals tighter clusters for the latent distributions of original and Gaussian-noise-added inputs in our SRL-VAE while the original SD-VAE demonstrates random scatters, highlighting that our model's latent space is more robustly constructed as intended.

Perturbed image compression in latent space To further analyze the structure of the latent space learned by our SRL-VAE, we perform a Principal Component Analysis (PCA) on latent vectors, following Kouzelis et al. [13]. Specifically, we apply PCA on latent representation vectors obtained from the original SD-VAE and our SRL-VAE, derived from adversarially perturbed examples, using three types of perturbations, Photoguard, Mist, and Glaze. As shown in Figure 9, our SRL-VAE produces more structured and significantly smoother latent vector distributions compared to the original VAE on perturbed inputs. Furthermore, our SRL-VAE consistently generates robust latent vectors regardless of the type of perturbation. This suggests that adversarial training plays a crucial role in constructing better latent representations, enabling our SRL-VAE to encode features more reliably even under adversarial perturbations.

5. Conclusion

In this work, we first introduce an adversarial training framework for Variational Autoencoders (VAEs) that enhances both generation quality and robustness by encoding a smooth latent space. Unlike conventional adversarial training in classification models, which has a clear trade-off

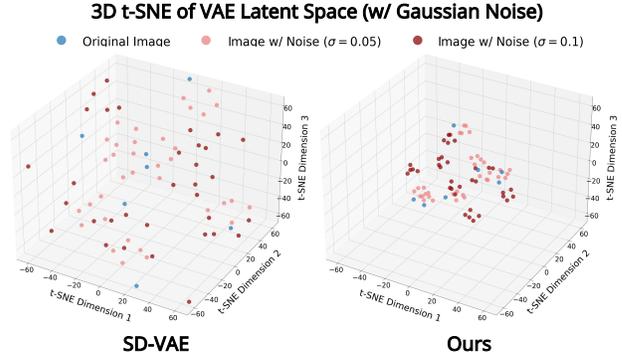## 3D t-SNE of VAE Latent Space (w/ Gaussian Noise)

Figure 8. **3D t-SNE visualization of latent representations under Gaussian noise.** SRL-VAE exhibits tighter and more consistent clusters than the baseline VAE, demonstrating improved robustness in latent space.
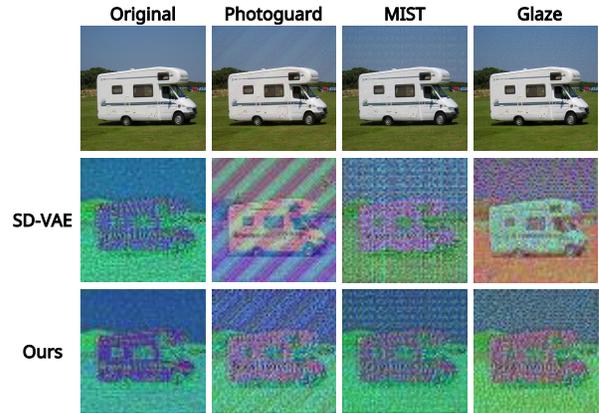

Figure 9. **PCA visualization of latent representations under adversarial perturbations.** Compared to SD-VAE, which exhibits distorted latent structures, SRL-VAE produces smoother, more organized, and well-separated distributions. This highlights its superior robustness and stability against diverse perturbations.

between performance and robustness, our approach leverages adversarial perturbations with an originality regularization term to preserve the learned latent space from the pre-trained model, ensuring smooth latent encodings in VAEs and enhancing both fidelity and robustness. Our method is a post-training step, requiring minimal computational resources, making it an efficient and practical solution for recent diffusion-based generative models. Extensive experiments demonstrate that our approach not only improves image quality, but also significantly enhances robustness against diverse types of adversarial attacks, such as poisoning and perturbation attacks. By bridging adversarial training and generative modeling, our work highlights the importance of obtaining a robust, high-quality latent space in VAEs, opening new directions for future research in robust generative modeling.

References

[1] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36:10657–10677, 2023. 6

[2] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 2019. 2, 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 3

[6] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 2021. 3

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[8] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024. 6

[9] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In *Advances in Neural Information Processing Systems*, 2020. 3

[10] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020. 2

[11] Minseon Kim, Hyeonjeong Ha, Sooel Son, and Sung Ju Hwang. Effective targeted attacks for adversarial self-supervised learning. *Advances in Neural Information Processing Systems*, 2023. 3

[12] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 2019. 1, 2

[13] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509*, 2025. 3, 8

[14] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, 2018. 2

[15] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 2024. 1

[16] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023. 5, 7, 12

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 5

[18] Yiwei Lu, Matthew YR Yang, Zuoqiu Liu, Gautam Kamath, and Yaoliang Yu. Disguised copyright infringement of latent diffusion models. In *International Conference on Machine Learning*, 2024. 5

[19] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024. 1

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 3

[21] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. 6

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 5, 11

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PmLR, 2021. 5

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 11

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 2016. 5

[26] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning*, 2023. 5, 7, 12

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022. 4

[28] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists

from style mimicry by {Text-to-Image} models. In *USENIX Security Symposium (USENIX Security 23)*, 2023. 5, 7, 12

[29] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 807–825. IEEE, 2024. 5

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[31] Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025. 3

[32] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[33] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 2017. 2

[34] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019. 3

[35] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 5

[36] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 2020. 2

[37] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[38] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *International Conference on Learning Representations*, 2025. 1

[39] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019. 2, 3

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[41] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 6

Enhancing Variational Autoencoders with Smooth Robust Latent Encoding

Supplementary Material

A. Experimental details

A.1. Implementation details

Since there is no official fine-tuning script available for SD-VAE[1], we implemented our own fine-tuning script using the `diffusers` library. The SD-VAE is based on the Latent Diffusion Model (LDM) [24], and we analyzed its structure to derive meaningful insights into the training procedure. This implementation provided the flexibility needed to adapt the training process to our specific objectives.

To better understand fine-tuning methods, we examined the `sd-ft-mse` and `sd-ft-ema` models released by Stability AI via Hugging Face, which are fine-tuned versions of the SD-VAE decoder. The models are trained to enhance the detail of the image, with a particular focus on human facial features. SDXL retains a structure largely similar to SD-VAE but is trained with a larger batch size on an internal dataset, further demonstrating the scalability of VAE-based architectures.

As most related works are based on SD-VAE, we adopted it as the primary baseline for our experiments. Furthermore, we confirmed that the proposed adversarial training method generalizes effectively to other VAE architectures, including SDXL-VAE, indicating its broad applicability. Our VAE encoder fine-tuning is relatively lightweight and was conducted using four A5000 GPUs with 24GB of VRAM each, taking approximately 7 hours to complete. Our training configuration is in Table 5:

Hyperparameter	Value
Batch size	20
Total training steps	5000
Learning rate	1×10^{-4}
Optimizer	AdamW
ϵ-bound (ℓ_∞ norm)	8/255
PGD iterations	10
PGD attack step size	0.02
Originality loss weight (α)	0.01

Table 5. **Training Configuration.**

A.2. Evaluation details

Image reconstruction quality We adopted evaluation metrics consistent with prior studies to measure the recon-

struction quality of VAEs. Specifically, we utilized the MS-COCO validation set (5,000 images) and ImageNet validation set (50,000 images), resizing all images to 256×256 pixels, as commonly done in prior studies. Minor numerical discrepancies with prior results may occur due to differences in code implementations. For evaluation, we employed LPIPS with a VGG backbone to measure perceptual similarity. Torchmetrics was used to compute Fréchet Inception Distance (FID), and Scikit-image was utilized for calculating Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Pre-trained VAE models were loaded and processed using the Diffusers library.

Diffusion generation quality To evaluate the diffusion generation quality, we utilized the official DiT implementation [22] and trained the model on ImageNet-1000K for 10 epochs using four A100 GPUs. For the computation of standard evaluation metrics such as Inception Score (IS) and Fréchet Inception Distance (FID), we used the `sample_ddp.py` script provided by the official repository. This script supports parallel sampling of a large number of images and automatically generates a `.npz` file that is compatible with ADM's TensorFlow-based evaluation suite. We followed the standard protocol of generating 50,000 samples to ensure comparability with prior work.

Robustness in Nightshade malicious attack In our Nightshade attack experiments, we utilized the official implementation code. Instead of the original LIPIPS perturbation, we employed L_∞ perturbation, which is more visually noticeable to the human eye yet provides more stable results. We tested eight concept pairs (Poisoned concept C, Destination concept P), namely (cake, car), (cat, dog), (hat, horse), and (boat, bird), including their reversed pairs, and reported the average attack success rate. All models were trained for 10 epochs, altering only the poisoning ratio, with a batch size of 32 and a learning rate of 1×10^{-5}. For evaluation, we generated 100 images for each trained model using the prompt "A photo of [C]." We then set the threshold $\tau = 0.25$ for the CLIP classifier to a reasonable value based on human inspection.

Robustness against various type of perturbations For image-to-image experiments against various perturbations, we used the Stable Diffusion v1.5 model with a strenght value of 0.5. Editing prompts were extracted using BLIP and appropriately modified. We observed that higher strength values lead to more extensive image modifications,

---

[1] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

as they correspond to starting the reverse diffusion process from noisier latents (i.e., later timesteps), which can diminish the effectiveness of adversarial perturbations. To balance the degree of modification and the preservation of the original content, we set the strength value to 0.5. A brief description of the perturbation-based defense methods can be found in Appendix A.3.

A.3. Details of perturbation methods

PhotoGuard [26] protects images by adding imperceptible perturbations that disrupt the latent diffusion pipeline. It introduces two types of attacks: the *encoder attack* and the *diffusion attack*. The encoder attack perturbs the input image so that the encoder E maps it to a misleading latent representation. Formally, the perturbation is computed as:

$$\delta_{\text{enc}} = \arg \min_{\|\delta\|_\infty \leq \epsilon} \|E(x + \delta) - z_{\text{targ}}\|_2^2. \quad (9)$$

This causes the diffusion model to generate irrelevant outputs, effectively preventing inpainting and style imitation. The diffusion attack, on the other hand, directly targets the entire generation process, aiming to produce a specific target image x_{targ} as output:

$$\delta_{\text{diff}} = \arg \min_{\|\delta\|_\infty \leq \epsilon} \|f(x + \delta) - x_{\text{targ}}\|_2^2. \quad (10)$$

While more powerful, the diffusion attack requires backpropagation through the full denoising process and is computationally expensive. In our work, we adopt only the encoder attack, since the diffusion attack in PhotoGuard is tailored to a specific inpainting model.

MIST [16] extends the idea of adversarial perturbations by aiming for broader transferability across diffusion-based image imitation pipelines. While PhotoGuard applies encoder or diffusion attacks separately, MIST combines both approaches through a joint loss function. Specifically, it introduces two loss terms: a *textural loss*, which maximizes the distance between latent representations of clean and perturbed images, and a *semantic loss*, which increases the diffusion model's denoising error. The combined objective is optimized via projected gradient descent:

$$\delta = \arg \max_{\|\delta\|_\infty \leq \epsilon} \left[w \cdot \mathbb{E}_{t,\varepsilon} \|\varepsilon - \epsilon_\theta(x'_t, t)\|_2^2 \right.$$
$$\left. - \|E(y) - E(x + \delta)\|_2 \right] \quad (11)$$

where E is the encoder, x'_t is the perturbed latent at step t, and y is a target image. This joint formulation improves transferability, making MIST effective against a range of downstream applications, including style transfer, textual inversion, and DreamBooth. In contrast to PhotoGuard's

model-specific attacks, MIST focuses on general-purpose protection. Our experimental results show that SRL-VAE maintains robustness against MIST perturbations, highlighting the importance of a robust VAE bottleneck in defending against advanced attacks.

Glaze [28] defends against style mimicry by applying imperceptible perturbations. Specifically, it computes a perturbation δ_x that shifts the feature representation of an original artwork x toward that of a style-transferred version $\Omega(x, T)$ in the feature space of a pretrained encoder Φ, where T denotes a visually distinct target style:

$$\delta_x = \arg \min_{\delta_x} \text{Dist}(\Phi(x + \delta_x), \Phi(\Omega(x, T)))$$
$$\text{s.t.} \quad |\delta_x| < p \quad (12)$$

where p is a perceptual distortion bound measured by LPIPS. This ensures that the cloaked image remains visually similar to the original, while altering its representation in the model's latent space. When a model is trained on such cloaked images, it learns distorted style representations that blend the artist's original style with the target style, leading to degraded mimicry performance. Glaze is released as a utility tool, so its perturbation logic is a black box. Nonetheless, our experiments show that SRL-VAE is robust against Glaze, indicating that our method generalizes well even to black-box defenses.

A.4. Perturbation Strength Selection

For practical deployment, perturbations should remain imperceptible to human observers while effectively disrupting model performance. To ensure realistic scenarios, we fix the perturbation magnitude for each method as follows: $\epsilon = 16/255$ for **PhotoGuard**, $\epsilon = 8/255$ for **MIST**, and the strongest available setting provided by the official utility tool for **Glaze**. This configuration balances perceptual quality and defensive efficacy, preventing diffusion models from successfully learning and replicating protected image styles.

B. More experimental results

We provide additional qualitative results in our experiments. Figure 10 shows results under the Nightshade attack. While the baseline model produces manipulated outputs, SRL-VAE preserves the original prompt concept, demonstrating strong robustness. Figure 11, 12 and 13 show image-to-image editing results on adversarially protected images, comparing the outputs of SRL-VAE and the baseline SD-VAE. These examples further demonstrate the robustness of SRL-VAE in reconstructing and editing perturbed inputs while preserving semantic consistency and visual quality.
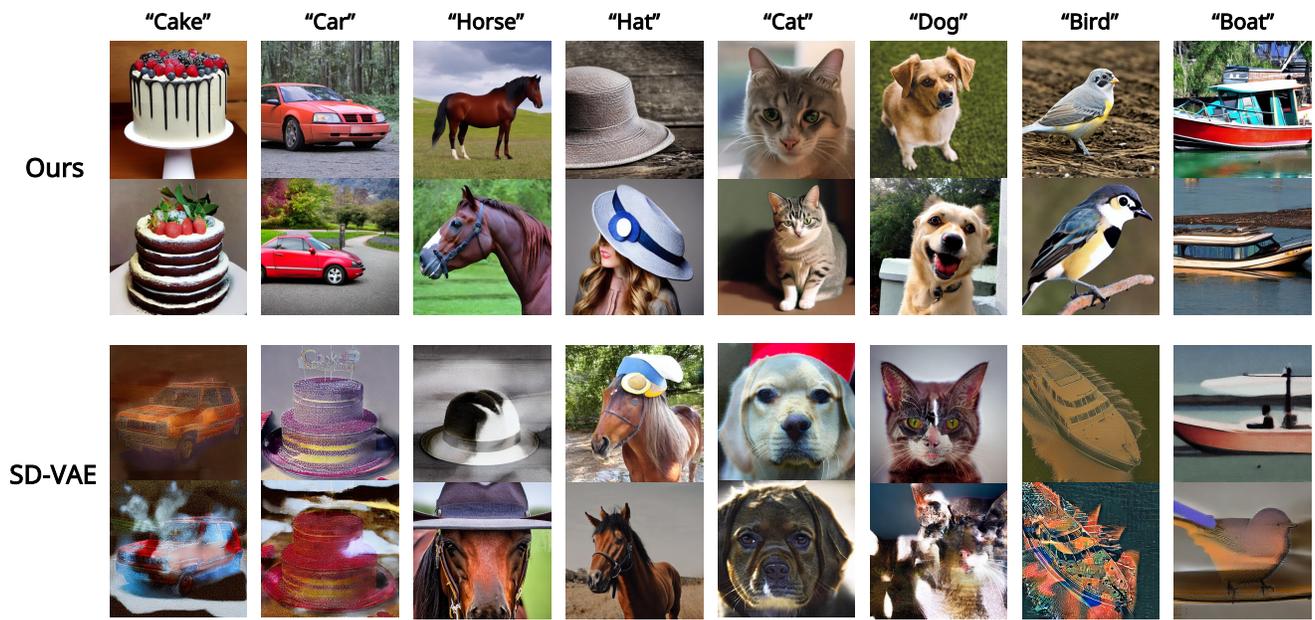
Figure 10. **Examples generated from models that were attacked using Nightshade.**
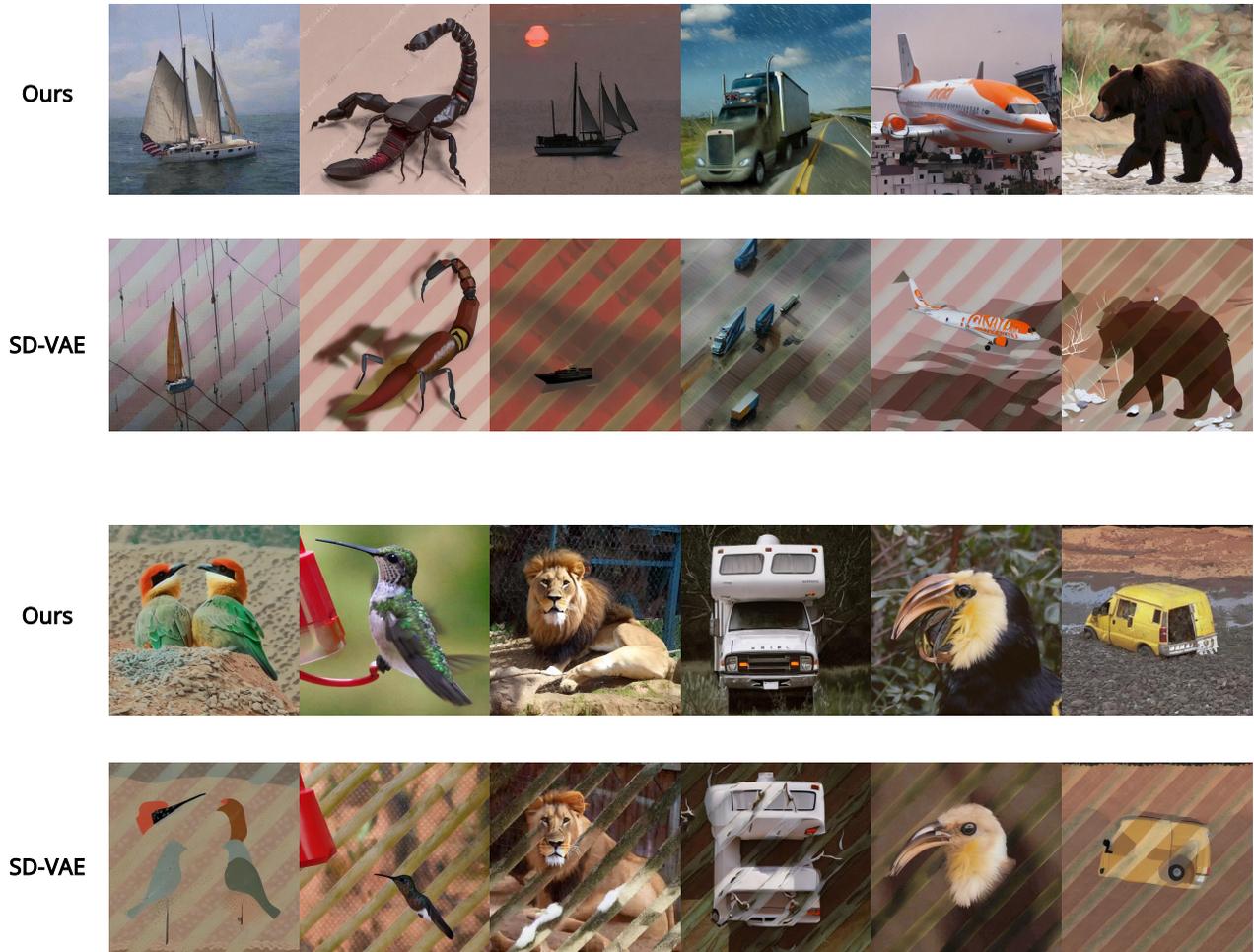
Ours

SD-VAE

Ours

SD-VAE

Figure 11. **Comparison of image-to-image editing results on protected images (Photoguard).**
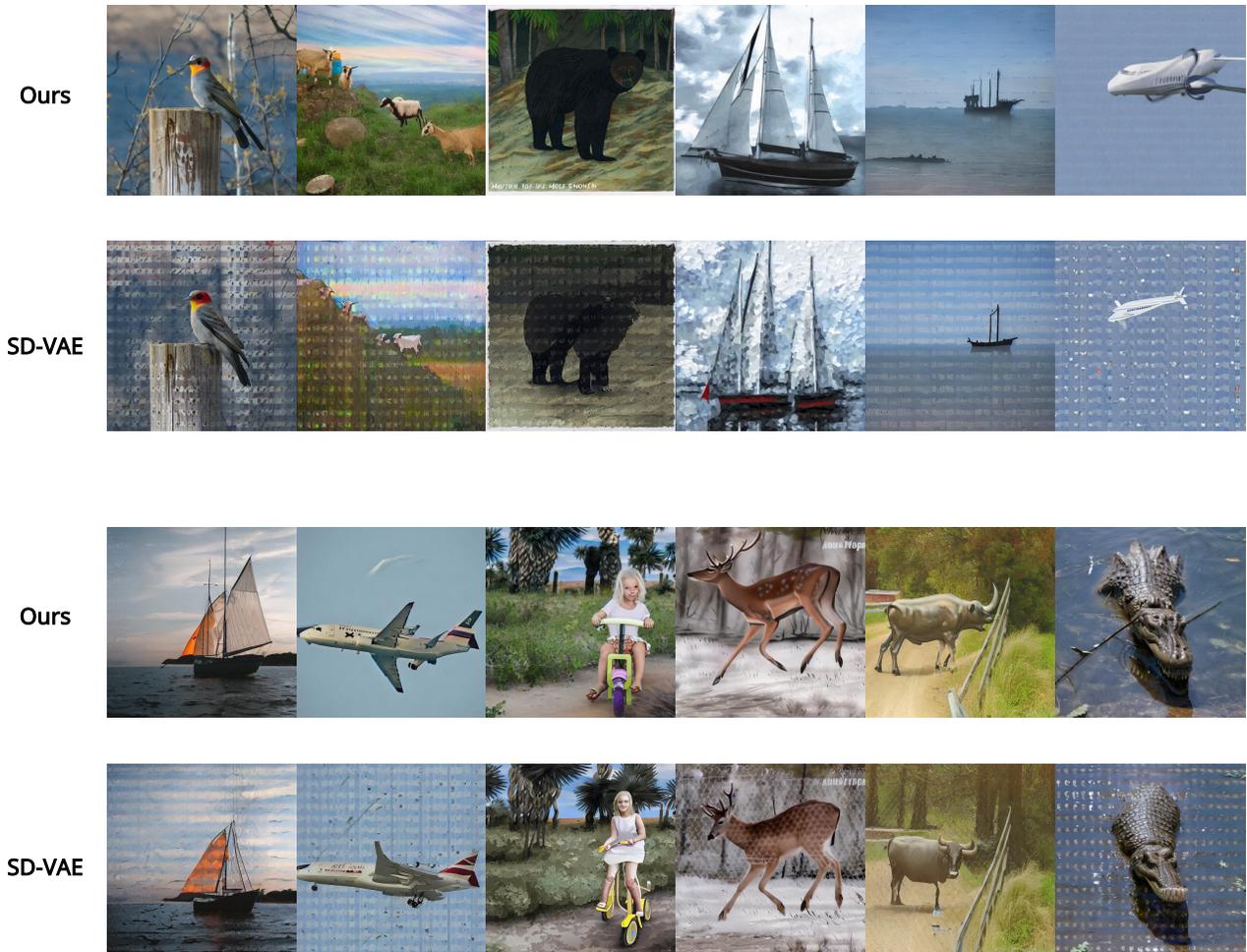
Figure 12. **Comparison of image-to-image editing results on protected images (MIST).**
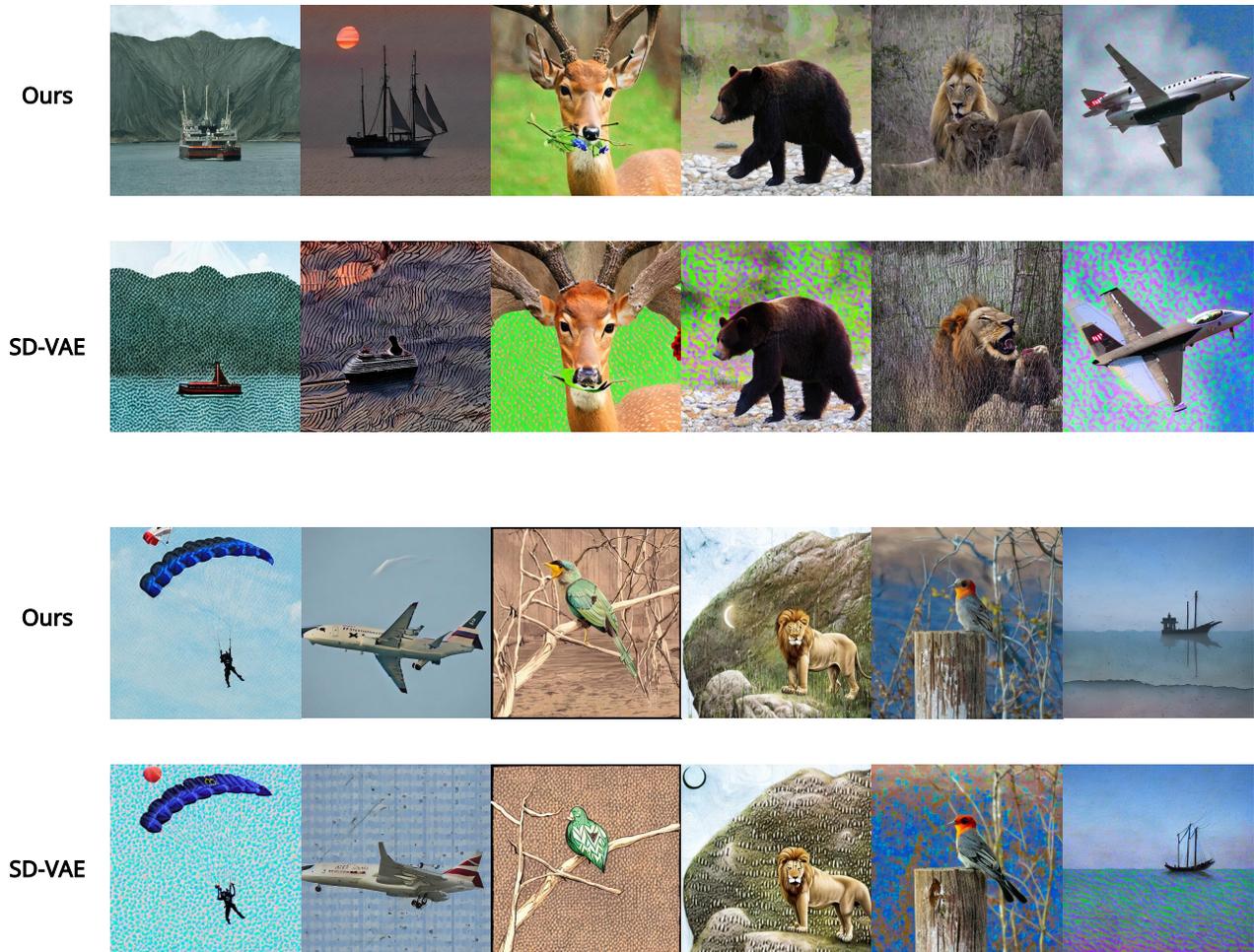
Figure 13. **Comparison of image-to-image editing results on protected images (Glaze).**