Preventing Latent Diffusion Model-Based Image Mimicry via Angle Shifting and Ensemble Learning

Minghao Li, **Rui Wang***, **Ming Sun** and **Lihua Jing**
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
{liminghao, wangrui, sunming, jinglihua}@iie.ac.cn

Abstract

The remarkable progress of Latent Diffusion Models (LDMs) in image generation has raised concerns about the potential for unauthorized image mimicry. To address these concerns, studies on adversarial attacks against LDMs have gained increasing attention in recent years. However, existing methods face bottlenecks when attacking the denoising module. In this work, we reveal that the robustness of the denoising module stems from two key factors: the cancellation effect between adversarial perturbations and estimated noise, and unstable gradients caused by randomly sampled timesteps and Gaussian noise. Based on these insights, we introduce a cosine similarity adversarial loss to prevent the generation of perturbations that are easily impaired and develop a more stable optimization strategy by ensembling gradients and fixing the noise in the latent space. Additionally, we propose an alternating iterative framework to reduce memory usage by mathematically dividing the optimization process into two spaces: latent space and pixel space. Compared to previous strategies, our proposed framework reduces video memory demands without sacrificing attack effectiveness. Extensive experiments demonstrate that the alternating iterative framework and the stable optimization strategy on cosine similarity loss are more efficient and more effective. Code is available at https://github.com/MinghaoLi01/cosattack.

1 Introduction

Latent Diffusion Models (LDMs) exhibit exceptional capabilities [Song *et al.*, 2021; Ho *et al.*, 2020; Sohl-Dickstein *et al.*, 2015; Rombach *et al.*, 2022a], achieving state-of-the-art performance in various image synthesis tasks [Meng *et al.*, 2022; Saharia *et al.*, 2023]. The remarkable progress of LDMs opens up new possibilities in content creation and art design. However, alongside these groundbreaking achievements, the power of LDMs also presents significant ethical and security challenges. Their capabilities can be maliciously exploited to generate forged human faces or fake artworks. These concerns highlight the urgent need for safeguards to prevent the potential misuse of LDMs.

Figure 1: The visualization of cancellation effect. The cancellation vector represents the projection of the estimated noise in the direction of the adversarial perturbation.

Currently, protections against unauthorized image mimicry are primarily based on adversarial attack methods. Most existing methods [Liang *et al.*, 2023; Xue *et al.*, 2023; Liang and Wu, 2023] add imperceptible perturbations to input images by maximizing ℓ_2 losses (e.g. the original training loss), thus introducing errors into the denoising module of LDMs. The perturbations cause LDMs to predict the ground-truth noise with bias, thereby preventing the generation of high-quality images. However, several issues remain unresolved. Firstly, although the adversarial robustness inherent in the denoising module is pointed out [Xue *et al.*, 2023], the underlying reasons for this robustness remain insufficiently explored, leading to the absence of more effective attacks on the denoising module. Secondly, to protect high-resolution images on devices with limited video memory resources, current approaches [Xue *et al.*, 2023; Liang and Wu, 2023] sacrifice protection performance more or less by altering the adversarial loss or the protected image.

To address these gaps, we investigate the factors contributing to the robustness of the denoising module and identify two key reasons. Firstly, the cancellation effect between adversarial perturbations and estimated noise diminishes the magnitude of perturbations. We refer to the projection of the estimated noise onto the adversarial perturbations as the cancellation vector, and the remaining part of the adversarial perturbations as the effective vector. The adversarial perturbations are severely disrupted by the denoising module during the reverse process as shown in Figure 1. Secondly, multiple attack objectives at multi-step Markov chains and oscillating gradient directions under random Gaussian noises undermine

*Corresponding author.

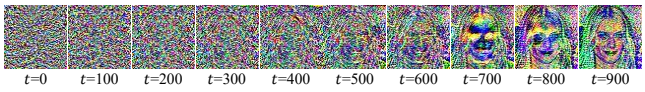


Figure 2: Gradients under different timesteps are complementary. At small timesteps, the gradient carries more about the noise, while at large timesteps, it contains more image details.

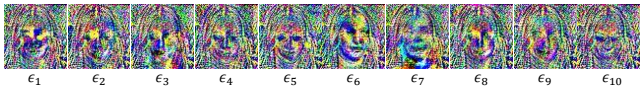


Figure 3: Gradients under random noises when $t = 700$. The irregular gradients oscillate the optimization direction.

the stability of optimization in latent space. We discover that the gradients under different timesteps are complementary as shown in Figure 2, and the gradients under randomly sampled noise are irregular as shown in Figure 3.

Based on these discoveries, we demonstrate that the amplified cancellation effect results from maximizing the scale-sensitive ℓ_2 loss, which increases the magnitude of estimated noise when optimizing adversarial perturbations. We conduct an angle-shifting attack by maximizing cosine similarity loss, which introduces bias to the noise estimator without increasing its magnitude. Furthermore, we propose a stable optimization strategy to attack the multi-step Markov chains effectively. Regarding adversarial attacks at different timesteps as a special ensemble attack [Liu *et al.*, 2017], we develop a grouped gradient ensemble strategy to efficiently leverage the complementary gradients at different timesteps. To eliminate the disruption to gradient stability caused by randomly sampled Gaussian noise, we fix the noise in latent space for a certain input image. In addition, we propose an alternating iterative framework. By mathematically decomposing the gradient computations into two steps, we load only the gradients of either the denoising module or the encoder at a time to reduce the VRAM usage. This framework is applicable to other adversarial attacks without compromising their effectiveness.

In summary, our contributions can be divided into the following points.

- Revisiting existing adversarial attacks on LDMs, we propose a plug-and-play alternating iterative framework to decrease VRAM demands without sacrificing attack effectiveness.
- We reveal the cancellation effect between the adversarial perturbations and the estimated noise. To avoid the side effect which is caused by the scale sensitivity of ℓ_2 loss, we conduct an angle-shifting attack by maximizing the cosine loss.
- Regarding attacks against the denoising module at different timesteps as an ensemble attack, we propose a gradient ensemble strategy and fix the noise to stabilize the adversarial optimization process.
- Extensive experiments conducted on the facial dataset CelebA-HQ and the artworks dataset WikiArt demonstrate that our approach outperforms existing protection methods across various scenarios.

2 Related Work

PhotoGuard [Salman *et al.*, 2023] proposes attacking the encoder or denoising process of LDMs to raise the cost of malicious AI-Powered image editing. However, attacking the denoising process is impractical due to its significant VRAM requirements. AdvDM [Liang *et al.*, 2023] successfully attacks the denoising module by employing Monte Carlo sampling across timesteps and maximizing the ℓ_2 training loss, marking the first effective adversarial attack against the denoising module of LDMs. Mist [Liang and Wu, 2023] builds upon PhotoGuard and AdvDM to implement a hybrid targeted attack. To further minimize video memory usage while protecting high-resolution images, Mist generates low-resolution adversarial patches, which are then assembled into a full image. SDS [Xue *et al.*, 2023] reduces computational complexity and VRAM requirements by discarding the Jacobian term in the loss function. Additionally, SDS demonstrates that minimizing the adversarial loss yields more natural image protection. Most existing methods rely on ℓ_2-norm loss and direct PGD optimization in pixel space. MFA [Yu *et al.*, 2024] proposes attacking the denoising module by mean fluctuation under a certain timestep, but calculating the vulnerability of timesteps relies on specific adversarial loss and numerous experiments. In this paper, we introduce an alternating iterative framework that significantly reduces video memory requirements without compromising attack performance, thereby enhancing the practicality of these methods. By revisiting the objective function and optimization strategies of current approaches, we propose an angle-shifting attack alongside a more stable optimization strategy.

3 Preliminary

One of the reasons why LDMs can achieve great success in image generation is that they combine Variational Auto Encoder (VAE) [Kingma and Welling, 2014] and diffusion models [Ho *et al.*, 2020] to improve the efficiency of generation. Firstly, images sampled from the real distribution are compressed through the VAE $\mathcal{E}(\cdot)$ encoder as:

$$z_0 = \mathcal{E}(x_0). \tag{1}$$

Then, the forward process and the reverse process (i.e. denoising process) are carried out on the latent space $\mathcal{Z}$. The forward diffusion process of LDMs is designed to progressively introduce noise into the latent representation z_0. The process unfolds over a fixed number of timesteps T. At each timestep t $(1 \leq t \leq T)$, the latent variable z_t is computed by adding Gaussian noise to the previous latent variable z_{t-1}. Mathematically, the forward process is formulated as:

$$z_t = \sqrt{\alpha_t} z_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \tag{2}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ is a standard Gaussian noise sampled independently at each timestep t. Continuously expand z_{t-1} in Eq.(2) until reaching z_0, and sum up the remaining ϵ_t. This way, a faster forward sampling Eq.(3) can be obtained.

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{3}$$

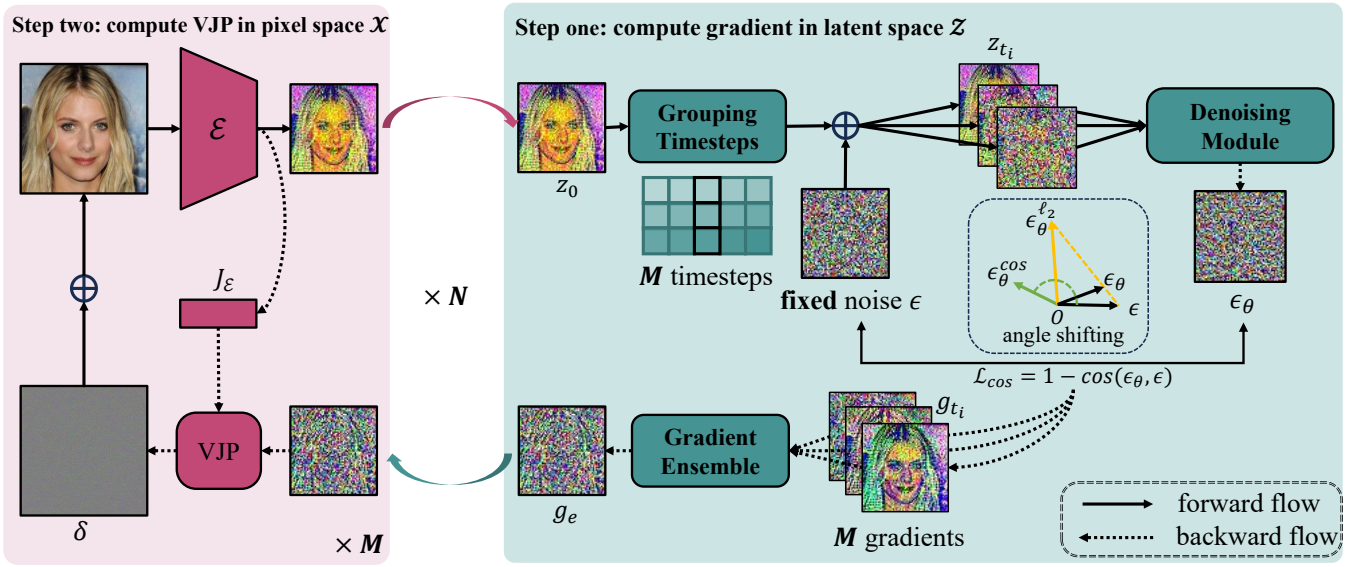


Figure 4: The pipeline of our methods. First, the gradient in latent space is computed using cosine loss and stabilized through gradient ensemble and fixed Gaussian noise. Then, the latent gradient is used to guide the optimization of adversarial perturbations in the pixel space by computing the VJP. At each step, only the gradients of the denoising module or encoder are loaded.

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. When $t \to T, \bar{\alpha}_t \to 0, z_T$ almost becomes Gaussian noise.

The reverse process is tasked with recovering the original latent representation z_0 from the noisy latent variable z_T. The transition probability is defined as $p_\theta(z_{t-1}|z_t)$, which is parameterized by a neural network θ. The network is designed to predict the mean $\mu_\theta(z_t, t)$ and variance $\Sigma_\theta(z_t, t)$ of the distribution from which z_{t-1} should be sampled given z_t. The predicted mean $\mu_\theta(z_t, t)$ is computed as

$$\mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(z_t, t) \right), \tag{4}$$

where $\epsilon_\theta(z_t, t)$ estimates the noise that is added at step t during the forward process. The neural network is trained to minimize the error in this noise estimation. This is accomplished by optimizing the following ℓ_2 loss function finally:

$$\mathcal{L}_{train} = \mathbb{E}_{t,\epsilon_t} \mathbb{E}_{z_t} \| \epsilon_\theta(z_t, t) - \epsilon_t \|_2. \tag{5}$$

4 Methods

To reduce VRAM consumption without sacrificing attack effectiveness, we propose an alternating iterative framework in Sec. 4.1. Building on this framework, we further explore more effective adversarial loss functions in Sec. 4.2 and propose stable optimization strategies in Sec. 4.3. The pipeline of our methods is illustrated in Figure 4.

4.1 Alternating Iterative Framework

In general, existing adversarial attacks on the denoising module consume substantial VRAM, severely limiting their deployment on devices with constrained resources. The existing strategies sacrifice attack performance to overcome this challenge. In this section, we propose an alternating iterative framework that not only enables execution on resource-limited devices without sacrificing attack effectiveness but also provides deeper insights into the mechanics of current methods.

Let $\mathcal{L}_{adv}$ denote the adversarial loss against the denoising module. The optimization process of previous approaches can be formulated as:

$$x^{i+1} = \Pi_\infty(x^i + \alpha \cdot sign(\frac{\partial \mathcal{L}_{adv}}{\partial x^i})). \tag{6}$$

where Π_∞ represents the projection operator and x^i represents the adversarial example at the i-th iteration. By calculating the gradient of $\mathcal{L}_{adv}$ with respect to the adversarial image x', a PGD-based optimization step is performed in the pixel space $\mathcal{X}$ to complete one step optimization.

The reason why previous methods consume more VRAM is that they simultaneously load the gradients of both the encoder and the denoiser. Inspired by SDS [Xue *et al.*, 2023], we view the adversarial loss $\mathcal{L}_{adv}$ on the denoising process as a function of the image x in pixel space and the variable z in latent space, where z is derived from x through the VAE encoder $\mathcal{E}$ as shown in Eq.(1). According to the chain rule, we can expand $\mathcal{L}_{adv}$, as:

$$\frac{\partial \mathcal{L}_{adv}}{\partial x^i} = \frac{\partial \mathcal{L}_{adv}}{\partial \mathcal{E}(x^i)} \frac{\partial \mathcal{E}(x^i)}{\partial x^i}. \tag{7}$$

Note that Eq.(7) is a Vector Jacobian Product (VJP). The first term $\frac{\partial \mathcal{L}_{adv}}{\partial z^i}$ signifies the gradients in latent space. The second term is the Jacobian matrix that maps the latent gradients to pixel space.

To reduce VRAM demands without altering input images or adversarial loss, we divided the optimization process into two alternating iterative steps as Eq.(8)-(9). Initially, we endeavored to determine an optimal value for z^i by solving Eq.(8). Subsequently, this optimal z^{i*} serves as guidance for

the computation of the VJP as described in Eq.(9).

$$z^{i*} = z^i + \frac{\partial \mathcal{L}_{adv}}{\partial z^i}, \tag{8}$$

$$x^{i+1} = \Pi_\infty(x^i + \alpha sign((z^{i*} - z^i)\frac{\partial z^i}{\partial x^i})). \tag{9}$$

When computing Eq. (8) or Eq. (9) respectively, only the parameters of the denoiser or encoder are loaded into VRAM, while direct computation of Eq. (6) requires loading both the denoiser and encoder parameters simultaneously. Furthermore, performing one round of computation for Eq.(8) and (9) yields identical results to Eq.(6). Consequently, adversarial attacks based on Eq.(6) can be executed on devices with lower VRAM without sacrificing effectiveness by decomposing the adversarial attack into two alternating iterative steps.

4.2 Angle-Shifting Attack

Most current methods build upon maximizing the ℓ_2 loss such as training loss as Eq.(10) to disrupt the denoising process of LDMs to protect the images.

$$\mathcal{L}_{adv} = \|\epsilon_\theta(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t) - \epsilon_t\|_2. \tag{10}$$

However, these methods overlook the fact that adversarial perturbations δ can be disrupted by the denoiser ϵ_θ. Based on the framework proposed by Sec. 4.1, we consider perturbations directly on the latent variable z. We revisit the reverse process on adversarial examples by perturbing the Eq.(4) as Eq.(11).

$$\begin{aligned}\mu_\theta(z'_t, t) &= \frac{1}{\sqrt{\alpha_t}}\left(z'_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z'_t, t)\right) \\ &= \frac{1}{\sqrt{\alpha_t}}\left(z_t + \sqrt{\bar{\alpha}_t}\delta - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z'_t, t)\right),\end{aligned} \tag{11}$$

where $z'_t = \sqrt{\bar{\alpha}_t}(z_0+\delta)+\sqrt{1-\bar{\alpha}_t}\epsilon_t$ represents the adversarial latent variable $z_0 + \delta$ at timestep t. We find that the adversarial perturbation δ in Eq. (11) is weakened by the denoiser $\epsilon_\theta(z'_t, t)$. To explore the relationship between δ and $\epsilon_\theta(z'_t, t)$ simply, we ignore the time-dependent hyper-parameter coefficients $\sqrt{\bar{\alpha}_t}$ and $\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}$ and decompose $\epsilon_\theta(z'_t, t)$ orthogonally into two parts: ϵ^p_θ parallel to δ and ϵ^o_θ orthogonal to δ as shown in Figure 5. We find that the directions of $-\epsilon^p_\theta$ and δ are opposite, and $-\epsilon^p_\theta$ significantly undermines the effect of the adversarial perturbation by weakening its magnitude. We quantify the cancellation effect between the adversarial perturbations and estimated noise via the magnitude of cancellation vector ϵ^p_θ. As shown in Figure 7, when optimizing the adversarial perturbations, the magnitude of cancellation vector ϵ^p_θ increases simultaneously.

The reason lies in the scale sensitivity of the ℓ_2 adversarial loss, which means it is highly sensitive to the magnitude of the vectors. When optimizing ℓ_2 losses such as Eq.(10, the magnitude of ϵ_θ is inevitably increased, which further amplifies the norm of the cancellation vector via projection. Based on the analysis above, we conduct an angle-shifting attack by maximizing the cosine similarity loss to disregard the impact of $-\epsilon^p_\theta$ as:

$$\mathcal{L}_{cos} = \mathbb{E}_{t,\epsilon_t}\mathbb{E}_{z_t}(1 - cos(\epsilon_\theta(z_t, t), \epsilon_t)). \tag{12}$$

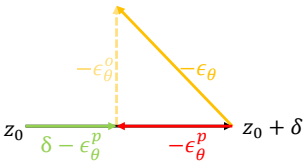


Figure 5: Cancellation effect

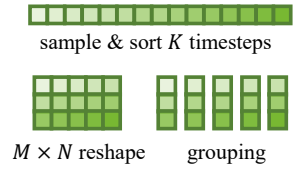


Figure 6: Grouping timesteps

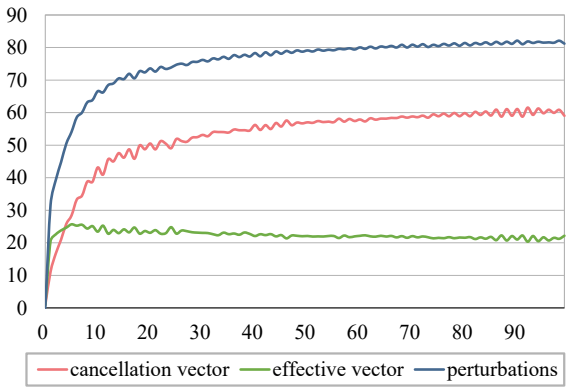


Figure 7: The magnitude of vectors during optimizing ℓ_2 loss when $t = 100$. The effective vector even decreases during optimization.

4.3 Stable Optimization Strategy

Previous works usually execute attacks by sampling a single timestep t from a uniform distribution at each optimization step. The shortcoming of this approach lies in the fact that one-step optimization in the pixel space $\mathcal{X}$ can only concentrate on the adversarial loss corresponding to a particular sampled t. We revisit adversarial optimization under different timesteps from the ensemble attack perspective and design a more efficient grouped gradient ensemble optimization strategy. Moreover, we observe that the random noise can lead to irregular oscillations in the gradient direction. We propose that applying a fixed Gaussian noise is conducive to the stable optimization of adversarial perturbations.

Gradient Ensemble in Latent Space

We perform attacks on the same image at different timesteps and visualize the gradients in Figure 2. It can be observed that the gradients at lower timesteps tend to disturb the noise ϵ_t while the gradients at higher timesteps tend to disturb z_0. To make a better use of the complementary information under different timesteps, we regard adversarial attacks against different timesteps t as those against different models and expect that optimization in the pixel domain $\mathcal{X}$ at each step can leverage the latent gradients under different t simultaneously and effectively.

Based on the alternating iterative framework, we propose a grouped gradient ensemble strategy in latent space. As illustrated in Eq.(13)-(14), the gradients under M timesteps t_i are computed for current z. Variance normalization is employed as Eq.(13) to eliminate the dimensional differences of the gradients when t varies. Subsequently, the average of these gradients is calculated to obtain an ensemble gradient in the la-

Edit Strength	0.1				0.2				0.3			
Metrics	LPIPS↑	SSIM↓	PSNR↓	FID↑	LPIPS↑	SSIM↓	PSNR↓	FID↑	LPIPS↑	SSIM↓	PSNR↓	FID↑
AdvDM	0.292	0.644	30.08	62.65	0.356	0.578	29.62	67.65	0.416	0.521	29.33	84.71
Mist	0.208	0.687	30.11	46.92	0.240	0.646	29.72	50.61	0.274	0.609	29.44	53.03
SDS	0.173	0.713	30.33	38.15	0.203	0.674	29.89	40.17	0.237	0.638	29.58	42.46
PhotoGuard	0.208	0.686	30.10	46.61	0.242	0.645	29.72	50.60	0.275	0.607	29.43	54.24
ours	**0.316**	**0.625**	**29.96**	**66.59**	**0.380**	**0.559**	**29.54**	**79.04**	**0.439**	**0.502**	**29.27**	**99.46**

Table 1: Quantitative results of different protection methods on CelebA-HQ

Edit Strength	0.1				0.2				0.3			
Metrics	LPIPS↑	SSIM↓	PSNR↓	FID↑	LPIPS↑	SSIM↓	PSNR↓	FID↑	LPIPS↑	SSIM↓	PSNR↓	FID↑
AdvDM	0.277	0.540	29.43	88.92	0.341	0.470	29.13	96.30	0.400	0.411	28.91	111.63
Mist	0.205	0.555	29.52	67.33	0.254	0.503	29.24	77.05	0.307	0.457	29.05	85.66
SDS	0.189	0.571	29.64	63.93	0.238	0.521	29.36	72.88	0.292	0.475	29.12	82.78
PhotoGuard	0.205	0.556	29.52	68.59	0.255	0.503	29.25	77.29	0.308	0.458	29.05	85.09
ours	**0.292**	**0.524**	**29.35**	**93.47**	**0.353**	**0.457**	**29.07**	**104.83**	**0.413**	**0.398**	**28.87**	**120.72**

Table 2: Quantitative results of different protection methods on WikiArt

tent space $\mathcal{Z}$. This ensemble gradient is then utilized to guide the optimization in $\mathcal{X}$. Thus, the adversarial perturbations in $\mathcal{X}$ at each optimization step can be regarded as an ensemble attack on the denoising estimators under M timesteps.

$$g_j^i = \frac{\partial \mathcal{L}_{adv}(z_{t_j}^i)}{\partial z^i}, \tag{13}$$

$$g_e^i = \frac{1}{M}\sum_{j=1}^{M}\frac{g_j^i}{\sigma_j^i}. \tag{14}$$

This process is iterated for N iterations. We name this optimization strategy as the $N \times M$ grouped optimization approach. Specially, the previous optimization methods can be seen as $K \times 1$ strategies. To ensure an equitable comparison, we set $N \times M = K$. Moreover, since M is considerably smaller than T , the sampling of M timesteps will introduce substantial randomness. To relieve the randomness, we initially sample and sort K timesteps, and then reshape them into an $M \times N$ matrix, which is then transposed to acquire the $N \times M$ timestep grouping as shown in Figure 6. Finally, the ensemble attack effectively leverages complementary information across M timesteps in latent space $\mathcal{Z}$ through one-step optimization in the pixel space $\mathcal{X}$, resulting in enhanced overall performance. It can be observed from the ablation experiment in Sec. 5.3 that the proposed optimization strategy effectively enhances the attack performance.

Fixing the Gaussian Noise

When training LDMs, distinct Gaussian noises ϵ_t are sampled at each t. This noise sampling strategy is widely adopted by previous adversarial attacks to perform the forward sampling of x_t as Eq.(2). However, we identify that this approach can induce gradient oscillations, further resulting in unstable optimization directions.

The practice of sampling diverse Gaussian noise during training LDMs is essential for augmenting the diversity of generated images. But in the context of adversarial example generation, the inherent randomness of Gaussian noise sampling introduces substantial instability into the optimization of adversarial perturbations. As illustrated in Figure 3, the gradient exhibits significant fluctuations under different Gaussian noise samples, even when the timestep remains constant. These irregular gradients counteract each other, resulting in unstable optimization directions.

5 Experiments

5.1 Setup

Datasets

We evaluate our methods on two datasets. Considering that infringement issues mainly occur on human faces and artworks, we use a subset of the CelebA-HQ [Karras *et al.*, 2018] and a subset of WikiArt [Nichol, 2016] respectively. We randomly select 500 face images from CelebA-HQ. The WikiArt dataset contains artworks from 27 different styles. We randomly selected 20 images from each style of artworks.

Baseline and Metrics

We compare our methods with four protection methods: AdvDM [Liang *et al.*, 2023], PhotoGuard [Salman *et al.*, 2023], Mist [Liang and Wu, 2023] and SDS [Xue *et al.*, 2023]. To assess the quality of the reconstructed images, we utilize PSNR, SSIM [Wang *et al.*, 2004], LPIPS [Zhang *et al.*, 2018] and FID [Heusel *et al.*, 2017] to evaluate the quality of the reconstructed images. Lower quality of the reconstructed images indicates better protection performance.

Experimental Settings

Following the existing research, we use the ℓ_∞-norm to constrain the generated adversarial examples, with the constraint range as 8/255 and the step size $\alpha = 1/255$. To facilitate the exploration of the impact of the grouping strategy, we set the number of iterations $K = 100$ for all the methods. For grouping strategy, we set $N \times M = 20 \times 5$.

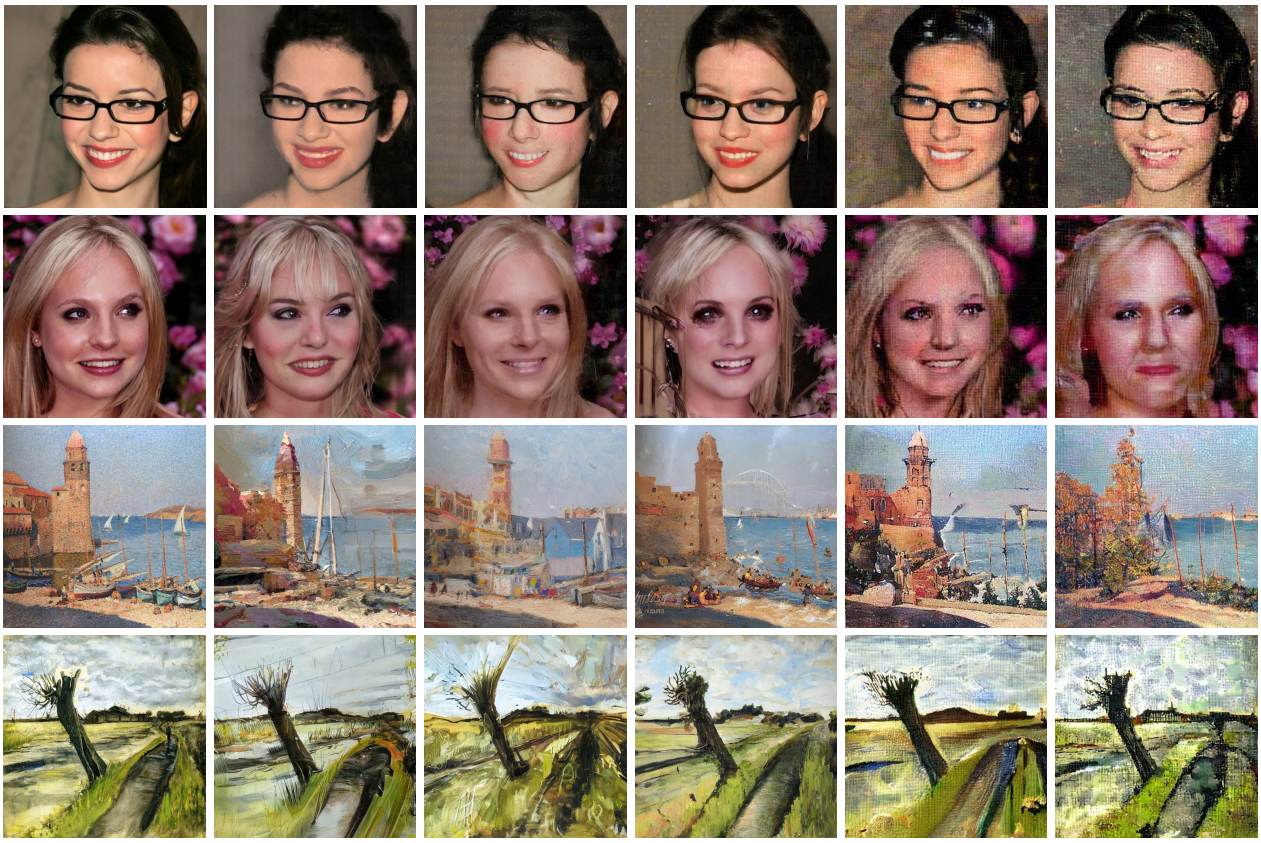


Figure 8: Qualitative results of different protection methods. Lower image quality indicates better protection effectiveness.

5.2 Protection Performance

We evaluate the performance of our methods on the SDEdit [Meng *et al.*, 2022] task. The input images are resized to 512×512. SDEdit conducted image reconstruction based on the backbone of SD-v1-4 [CompVis *et al.*, 2022] and DDIM100 [Song *et al.*, 2021] sampling. The edit strength of forward sampling is set to 0.1, 0.2, and 0.3 respectively.

Table 1 and Table 2 respectively display the quantitative results of our methods on CelebA-HQ and WikiArt. The best results are marked in bold, and the second-best results are underlined. It can be observed that when protecting high-resolution images, the protection performance of untargeted attacks (AdvDM,Ours) is superior to that of targeted attacks using textual loss (Mist, PhotoGuard, SDS). Our methods demonstrate SOTA protection performance under different edit strengths. Figure 8 shows the qualitative results of various protection methods when the edit strength is 0.5. The attack methods using texture loss against the encoder reconstruct clear and smooth images with disrupted textures, whereas untargeted attacks cause the reconstructed images to contain a significant amount of high-frequency noise, as these attacks are designed to interfere with the denoising module. Compared with AdvDM which maximizes the ℓ_2-norm training loss to optimize adversarial perturbations, the adversarial perturbations generated by our methods lead to a more solid detrimental effect on the denoiser ϵ_θ. See more experimental results on other tasks in the supplementary materials.

5.3 Ablation Study

Ablation Study on Different Modules

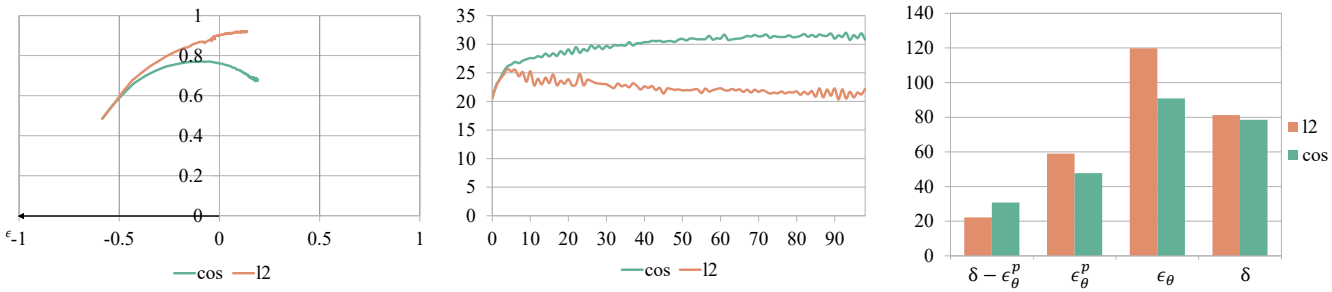
To verify the effectiveness of each module, we conduct ablation experiments on the SDEdit task and the CelebA-HQ dataset. Specifically, we examined the impact of the cosine similarity loss (cos) introduced in Sec. 4.2, the fixed noise (FN) and the Gradient Ensemble (GE) proposed in Sec. 4.3 on the attack performance. We chose AdvDM with ℓ_2 training loss as the baseline, which is also the second-best method under this experimental setup. Table. 3 presents the results of the ablation experiments. The results indicate that each of the modules we proposed significantly improves the attack performance. To evaluate the effect of the alternating iterative framework on the attack performance, we designed additional ablation experiments. In addition, we make a further exploration and analysis on the impact of different grouping strategies on the attack effectiveness. The experimental results in the supplementary materials further validate the improvement of protection performance by GE.

Ablation Study on Alternating Iterative Framework

To verify the improvement in VRAM resource usage efficiency and the impact on attack performance of the alternat-

Edit Strength	0.1				0.2				0.3			
Metrics	LPIPS↑	SSIM↓	PSNR↓	FID↑	LPIPS↑	SSIM↓	PSNR↓	FID↑	LPIPS↑	SSIM↓	PSNR↓	FID↑
no protect	0.080	0.789	31.37	15.48	0.109	0.747	30.72	19.06	0.138	0.710	30.25	22.95
AdvDM	0.292	0.644	30.08	62.65	0.356	0.578	29.62	67.65	0.416	0.521	29.33	84.71
+FN	0.303	0.628	30.03	63.46	0.367	0.563	29.60	74.01	0.429	0.505	29.31	94.36
+cos	0.300	0.639	30.03	62.94	0.366	0.571	29.59	70.04	0.427	0.513	29.30	86.98
+FN+cos	0.307	0.628	30.01	62.66	0.371	0.563	29.58	73.77	0.432	0.505	29.30	91.79
+FN+cos+GE	**0.316**	**0.625**	**29.96**	**66.59**	**0.380**	**0.559**	**29.54**	**79.04**	**0.439**	**0.502**	**29.27**	**99.46**

Table 3: Ablation Study of different modules on CelebA-HQ under different edit strengths



(a) the trajectory of ϵ_θ during optimizations (b) the curves of $\|\delta - \epsilon_\theta\|$ during optimizations (c) the norms of different vectors

Figure 9: Figure (a) illustrates the trajectory of ϵ during the optimization of different losses. Compared to the ℓ_2 loss, optimizing cosine loss does not increase the norm of ϵ. Figure (b) shows the curve of effective vector norms during optimization, where the cosine loss demonstrates a more effective increase in the norm of the effective vector. Figure (c) depicts the norms of various vectors. The cosine loss achieves the growth of the effective vector by suppressing ϵ.

ing iterative framework proposed in Sec. 4.1, we conduct experiments on NVIDIA GeForce GTX 1080Ti with 12G VRAM. In this scenario, AdvDM is unable to perform adversarial attacks on 512×512 images. We use three different methods to address this issue:

1. Image partitioning (IP): One solution is to tile the image and perform adversarial attacks on patches separately and then piece together the adversarial patches.
2. SDS: SDS reduces the VRAM occupancy of optimization by modifying the loss function.
3. Alternating iterative framework (AIF): The alternating iterative attack framework we proposed in Sec. 4.1. We apply the framework on AdvDM to ensure that the performance improvement comes from AIF but not other modules.

Table 4 shows the quantitative results of the three solutions when edit strength is 0.3. Alternating iterative framework reduces the VRAM demand while maintaining attack effect since it is mathematically equivalent to the attack process of AdvDM.

Metrics	LPIPS	SSIM	PSNR	FID
IP	0.188	0.666	29.92	29.65
SDS	0.200	0.671	29.89	33.66
AIF	**0.402**	**0.526**	**29.40**	**74.04**

Table 4: Ablation Study on Alternating Iterative Framework

Validation of Cosine Loss

Figure 9 shows how the cosine similarity loss mitigates the side effect. The timestep is fixed at $t = 100$ to provide stable visualization. Figure 9a visualizes the trajectory of ϵ_θ during optimizations, where the ground-truth ϵ_t is mapped to the coordinate $(-1, 0)$. Optimizing the cosine loss does not lead to the growth of ϵ_θ compared to ℓ_2 loss. Figure 9b shows the curve of the magnitude of effective vector $\delta - \epsilon_\theta^p$. Maximizing the ℓ_2 loss even shortens the effective vector. Figure 9c presents the magnitudes of vectors after 100 steps of optimizations. The growth of ϵ_θ is suppressed and the magnitude of the effective vector is increased by maximizing the cosine loss. It can be clearly observed that optimizing the cosine similarity loss reduces the side effects of ϵ_θ^p by suppressing ϵ_θ, thus ensuring the growth of effective vectors.

6 Conclusion

In this paper, we propose an alternating iterative attack framework that reduces VRAM demands without sacrificing performance. Based on the framework, we provide a mathematical analysis of the cancellation effect between perturbations and the denoiser. To mitigate this side effect and enhance attack effectiveness, we introduce a cosine similarity loss to address the limitations of the conventional ℓ_2 loss. Furthermore, by interpreting attacks on LDMs at multiple timesteps as an ensemble attack, we propose a grouped gradient ensemble strategy to better exploit complementary information across timesteps. In addition, we improve optimization stability by fixing the Gaussian noise during the attack process.

References

[CompVis *et al.*, 2022] CompVis, StabilityAI, and Runway. Stable diffusion v1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021.

[Gal *et al.*, 2023] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations,ICLR 2023, Kigali, Rwanda, May 1-5*, 2023.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Gur *et al.*, 2020] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Heusel *et al.*, 2017] 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.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[Kumari *et al.*, 2023] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1931–1941. IEEE, 2023.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[Le *et al.*, 2023] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2116–2127, 2023.

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

[Liang *et al.*, 2023] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pages 20763–20786, 2023.

[Liu *et al.*, 2017] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[Liu *et al.*, 2024] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24219–24228, 2024.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[Meng *et al.*, 2022] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[Nichol, 2016] K. Nichol. Painter by numbers. https://www.wikiart.org, 2016.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021.

[Rombach *et al.*, 2022a] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Rombach *et al.*, 2022b] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510, 2023.

[Ruiz *et al.*, 2024] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6527–6536, 2024.

[Saharia *et al.*, 2023] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4713–4726, 2023.

[Salman *et al.*, 2023] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

[Shan *et al.*, 2023] 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 *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 2187–2204, 2023.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021.

[van den Oord *et al.*, 2017] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315, 2017.

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

[Wang *et al.*, 2024] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 12047–12056, 2024.

[Xue *et al.*, 2023] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2023.

[Yu *et al.*, 2024] Hongwei Yu, Jiansheng Chen, Xinlong Ding, Yudong Zhang, Ting Tang, and Huimin Ma. Step vulnerability guided mean fluctuation adversarial attack against conditional diffusion models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 6791–6799, 2024.

[Zhang *et al.*, 2018] 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.