# QRShield: Exploiting Vulnerabilities of Latent Diffusion Models for Preventing AI Art Plagiarism

**Xunyue Mo, Weibin Wu*, Qingrui Tu, Hang Wang, Junxi He, Zibin Zheng**

School of Software Engineering, Zhuhai Key Laboratory of Trusted Large Language Models,
Sun Yat-sen University, Zhuhai 519082, China
{moxy53, tuqr, wangh856, hejx67}@mail2.sysu.edu.cn, {wuwb36, zhzibin}@mail.sysu.edu.cn

## Abstract

Latent Diffusion Models (LDMs) have achieved remarkable success in image generation tasks, yet their low barrier to customization poses severe threats related to art plagiarism. As a countermeasure, adversarial methods have been proposed to protect artworks from plagiarism. However, current methods suffer from limited effectiveness, high cost, and complex optimization. Moreover, their exploration and exploitation of LDM vulnerabilities remain limited, restricting effectiveness and applicability. To address this issue, we analyze the VAE and U-Net components of LDMs, revealing their vulnerabilities. Specifically, we study the response of U-Net to specific structural and frequency patterns in the latent space and find that it is susceptible to high-frequency and periodic latent features. Furthermore, we observe channel correlations during the VAE encoding process. Inspired by these, we propose QRShield, an efficient protection method that exploits the vulnerabilities of LDMs. By constructing high-frequency and periodic features consistent across latent channels and combining them with a momentum-based translation-invariant attack strategy, QRShield achieves stronger and more efficient protection. QRShield significantly improves protection performance in various fine-tuning settings, with over 10% gains in multiple metrics, a threefold increase in generation speed, and nearly 50% reduction in memory usage. Therefore, our work offers a more practical method to prevent AI art plagiarism.

**Code** — https://github.com/TAILab-W/QRShield

## 1 Introduction

Diffusion models (Sohl-Dickstein et al. 2015; Song and Ermon 2019; Ho, Jain, and Abbeel 2020; Song et al. 2021) have attracted remarkable attention for their strong performance in image generation. Among them, Latent Diffusion Models (LDMs) (Rombach et al. 2022; Podell et al. 2023) are widely adopted for their high efficiency and ability to produce high-quality results. However, their increasing popularity has also raised serious copyright concerns: malicious actors can plagiarize an artist's style using only a few samples and generate large amounts of unauthorized content, threatening the original artist's rights.

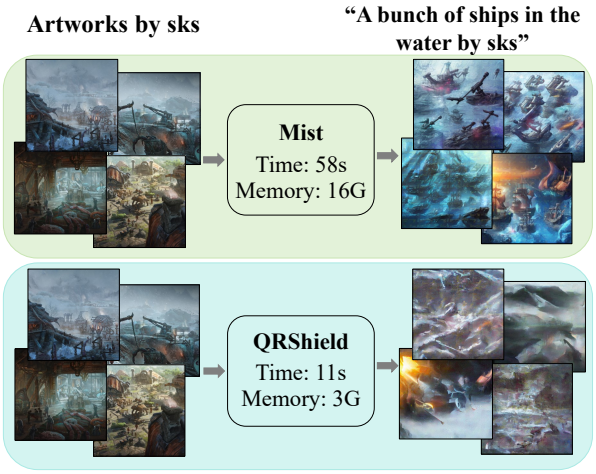


Figure 1: Comparison between Mist and our method: QRShield. Clean samples (left) are used to generate adversarial examples with both methods, which are then used to fine-tune the model to generate images (right).

As a countermeasure, researchers have proposed adversarial methods targeting these LDMs. These methods embed subtle perturbations into artworks to prevent LDMs from replicating artistic styles, protecting artists' rights. Among them, the Mist series (Zheng, Liang, and Wu 2023; Liang and Wu 2023; Zheng et al. 2023) stands out for its strong defense performance. These approaches apply adversarial perturbations (Wu et al. 2024; Deng et al. 2023) to bring the artworks closer in the feature space to a carefully designed anchor image, which has highly repetitive and high-contrast textures. If malicious users attempt to fine-tune models using these "poisoned" artworks, the resulting outputs will be visually disrupted textures rather than the original artistic styles, thus achieving protection.

However, despite such promising progress, existing methods still face several limitations, including limited effectiveness, high computational cost, and complex implementation processes, which limit their practical applicability. Moreover, current research focuses primarily on the end-to-end design of adversarial loss functions, while the incorporation of the properties of target LDMs remains limited.

*Corresponding author.

The potential vulnerabilities in the model architecture are largely underexplored, which limits further improvements in both computational efficiency and defensive effectiveness. As demonstrated in Figure 1, we compare our method with the state-of-the-art approach: Mist.

Motivated by these observations, we propose QRShield, an efficient protection method that uncovers and exploits vulnerabilities of LDMs. Drawing inspiration from the Mist series, our method incorporates anchors to enhance protection. However, Mist's anchors are manually selected based on experience. They focus on introducing visually unrelated textures to the protected images to interfere with the model. In contrast, QRShield focuses more on the relationship between anchor images and the LDM. By designing anchors that directly target the model's vulnerabilities, QRShield achieves more efficient and robust defenses.

Specifically, in this work, we focus on one of the most widely used types of LDMs: U-Net-based LDMs. Inspired by the frequency response analysis of LDMs in SimAC (Wang et al. 2024), we further study the sensitivity of the U-Net in LDMs to frequency components in the latent space. Our experiments show that U-Net maintains strong responsiveness to latent high-frequency features in all timesteps. Based on this finding, we construct anchors dominated by high-frequency components in the latent space, aiming to interfere with the model's perception of protected artworks maximally. Furthermore, we discover that U-Net has a preference for periodic textures, which motivates us to impose periodic constraints on our anchors.

Beyond the vulnerabilities of U-Net, we also find that VAE (Kingma and Welling 2013) of LDMs shows interesting properties that can be exploited. We observe consistency among multiple latent channels in the VAE: When drawing a protected artwork closer to a specific single-channel latent anchor in the latent space by manipulating only one channel, most other channels also shift toward the same anchor to varying degrees. To leverage this phenomenon, we propose a Channel Consistency Strategy that applies the same anchor across all latent channels, thereby enhancing the convergence of protected artworks toward the anchor under limited perturbation budgets.

After constructing the anchor, we apply adversarial perturbations to the artworks by minimizing the distance between them and the anchor in the VAE latent space, which is simple but effective. Meanwhile, to further improve protection effectiveness and robustness, we incorporate the Momentum Iterative FGSM (MI-FGSM) strategy (Dong et al. 2018) and a Translation-Invariant (TI) mechanism (Dong et al. 2019) during optimization.

Our main contributions are as follows:

- We demonstrate through experiments that the structural bias of the U-Net in LDMs toward latent high-frequency and periodic features, and identify channel consistency in the VAE latent space, providing guidance for anchor design.
- We propose a novel approach to construct anchors based on these LDMs' vulnerabilities. Unlike previous works that focus on designing anchors in the image space, we introduce anchors directly in the VAE latent space, focusing on full exploitation of LDM vulnerabilities.
- We develop QRShield, a more practical method for preventing AI art plagiarism. QRShield achieves stronger protection performance while significantly reducing computational cost, outperforming existing baselines in multiple metrics.

# 2 Related Work

## 2.1 Art Plagiarism by LDMs

With the rapid development of AI across diverse domains (Wu et al. 2025a,b,c), even non-expert individuals can now create stunning artworks within seconds using powerful AI generation tools. However, this technological advancement has also brought new challenges: AI art plagiarism. Some malicious users copy artworks using AI generators without the original artists' authorization or attribution to obtain illicit gains. These acts result in grave violations of the original artists' rights. Among these tools, latent diffusion models (LDMs) (Rombach et al. 2022; Podell et al. 2023) have become a primary choice for style plagiarism due to their outstanding performance and low cost.

At the same time, with the rapid development of LDM fine-tuning techniques, the threshold for replicating styles has continuously decreased. Early full fine-tuning of the U-Net, while highly effective, incurred high costs. Recently, lightweight fine-tuning methods such as LoRA (Hu et al. 2022) with Dreambooth (Ruiz et al. 2023), which can quickly replicate styles with only a few samples, have become the mainstream approach for personalized style generation. Meanwhile, Textual Inversion (Gal et al. 2022) has also shown great potential by achieving low-cost style transfer through updating only a few pseudo-token embeddings.

The broad availability of high-quality models and the rapid progress of fine-tuning techniques have increased the risk of AI plagiarism. Therefore, beyond enhancing copyright awareness and legal regulations, it is urgently necessary to design targeted technical methods that can greatly raise the technical costs of style imitation, effectively safeguarding artists' rights.

## 2.2 Adversarial Attacks against LDMs

To mitigate the risks above, a variety of adversarial methods have been proposed. These methods introduce carefully designed perturbations into artworks, causing models trained on these artworks to generate low-quality images, thereby preventing style theft. Notably, the perturbations are incredibly subtle, ensuring that the visual quality of the artworks remains unaffected.

Existing protection methods for LDMs are diverse, each offering unique strengths but also exhibiting notable limitations. As an early representative, AdvDM (Liang et al. 2023) first introduced adversarial examples into LDMs to protect images. While it performs well under the Textual Inversion setting, its effectiveness declines in more complex fine-tuning scenarios and requires expensive gradient computations on the denoising module. Glaze (Shan et al. 2023) brings the artworks closer to a content-similar

but style-different anchor in the VAE latent space, offering lightweight protection. Although Glaze is computationally efficient, its protective strength remains limited.

Mist (Liang and Wu 2023) and its enhanced version ITA (Zheng et al. 2023) inject misleading textures into the original image using manually designed high-contrast, highly repetitive anchors, confusing the model's understanding of style. Although these methods achieve strong protection performance, they rely on complex loss functions and handcrafted anchors without in-depth analysis. Diff-Protect (Xue et al. 2023) improves protection efficiency by leveraging Score Distillation Sampling (SDS) (Poole et al. 2022) into the adversarial optimization process, but its protection effectiveness has seen slight improvement.

Anti-DreamBooth (Van Le et al. 2023) aims to prevent misuse of DreamBooth-style personalization through an alternating training strategy, and its improved variant, SimAC (Wang et al. 2024), further enhances protection performance by optimizing timestep selection and introducing feature interference loss. Both approaches suffer from substantial memory consumption and computational costs, making them challenging to deploy in resource-limited environments.

In summary, current approaches struggle to balance between efficiency and lightweight design. Designing an effective and lightweight protection method is an important challenge in defending against AI art plagiarism.

# 3 Method

## 3.1 Preliminaries

**Diffusion Model (DM).** Diffusion Model (DM) is a class of powerful generative models with great success in image generation tasks. Latent Diffusion Model (LDM) is an efficient variant of diffusion models that transfers the diffusion process to the latent space.

Suppose $x_0 \sim q(x)$ denotes the real data distribution. LDM first maps $x_0$ to a latent representation $z_0 = E_\phi(x_0)$ using an encoder $E_\phi$, and then performs the diffusion process in the latent, including forward and reverse processes.

In the forward process, LDM gradually adds Gaussian noise to $z_0$ at each timestep $t \in \{1, \ldots, T\}$, producing a sequence of noisy latent variables $\{z_1, \ldots, z_T\}$. During the reverse process, the model $\epsilon_\theta(z_t, t, c)$ is trained to predict the noise added to $z_t$, typically optimized with an $\ell_2$ loss. The classic loss function is given by:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,t,c,\epsilon\sim\mathcal{N}(0,I)} \left[ \|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 \right]. \quad (1)$$

Finally, the denoised latent variable $\hat{z}_0$ is decoded back to the image space via the decoder $D_\psi$, generating a new image sample $\hat{x}_0 = D_\psi(\hat{z}_0)$.

Our work focuses on one of the most widely used classes of LDMs: U-Net-based LDMs. By analyzing vulnerabilities in their VAE and U-Net components, we design a more efficient approach for preventing AI art plagiarism.

**Adversarial Attack.** In preventing AI art plagiarism, adversarial attacks aim to introduce subtle perturbations to artworks $\{x\}$, resulting in adversarial examples $\{x_{\text{adv}}\}$ that cause the trained model to produce incorrect or severely degraded outputs. Such attacks typically constrain the perturbation within a small budget $\eta$, striving to minimize the impact on the artworks. For diffusion models, adversarial methods focus on attacking their key components.

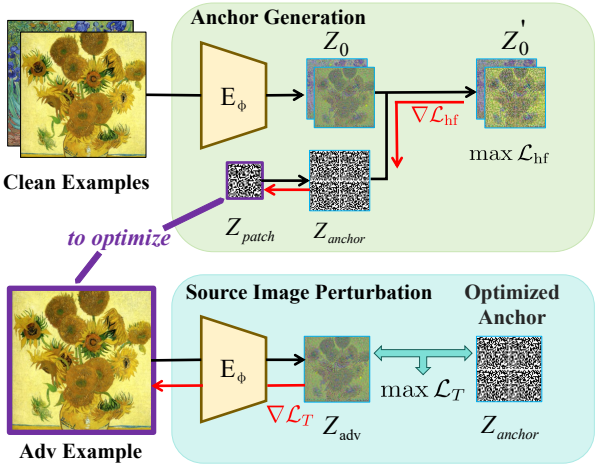


Figure 2: The workflow of QRShield. We first optimize the anchor image in the anchor generation stage. We then use the optimized anchor to perturb the source image.

Currently, most adversarial attack strategies against LDMs can be broadly categorized into three types: attacks targeting the VAE encoder, attacks targeting the noise predictor (U-Net), and hybrid attacks combining both. Our work focuses on anchor-guided attacks on the VAE encoder.

Specifically, given an input image $x$ and an anchor image $y$, we optimize the adversarial example $x_{\text{adv}} = x + \delta$ to maximize the similarity between the latent representations $E_\phi(x_{\text{adv}})$ and $E_\phi(y)$. The loss is defined as:

$$\mathcal{L}_{\text{T}}(x_{\text{adv}}) = -\|E_\phi(x_{\text{adv}}) - E_\phi(y)\|_2^2. \quad (2)$$

To control the perturbation magnitude while ensuring attack effectiveness, a common approach is to apply Projected Gradient Descent (PGD) with the following update rule:

$$x^{t+1} = \mathcal{P}_{B_\infty(x,\eta)} \left( x^t + \alpha \cdot \text{sign} \left( \nabla_{x^t} \mathcal{L}_{\text{T}}(x^t) \right) \right), \quad (3)$$

where $\mathcal{P}_{B_\infty(x,\eta)}(\cdot)$ denotes the projection operation onto the $\ell_\infty$-norm ball centered at $x$ with radius $\eta$, $\alpha$ is the step size, and the superscript $t$ indicates the PGD iteration step.

## 3.2 Overview

In this section, we analyze the vulnerabilities of LDMs and develop our method accordingly. Section 3.3 explains the critical role of the anchor and presents our novel perspective on anchor design. Section 3.4 focuses on the vulnerabilities of LDMs, based on which we design the anchor to guide the attack effectively. Finally, Section 3.5 presents the overall framework of QRShield, which includes two core components: anchor generation and source image perturbation. The workflow of QRShield is illustrated in Figure 2.

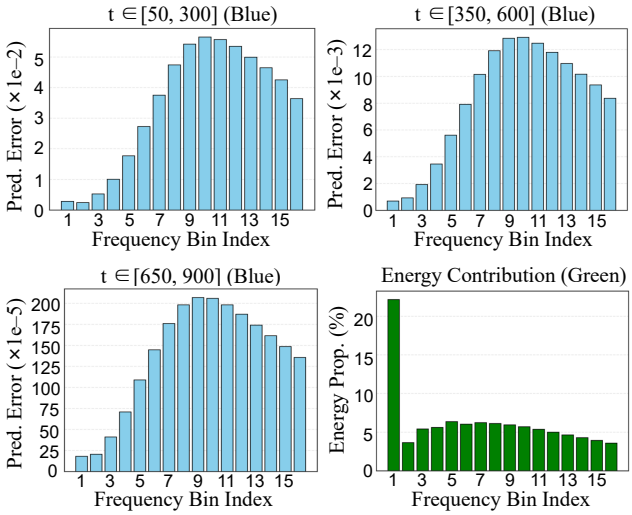


Figure 3: Frequency analysis results. Blue bars indicate the prediction error caused by removing the corresponding frequency components, while green bars represent the energy proportion of different frequency components. Higher index means higher frequency.

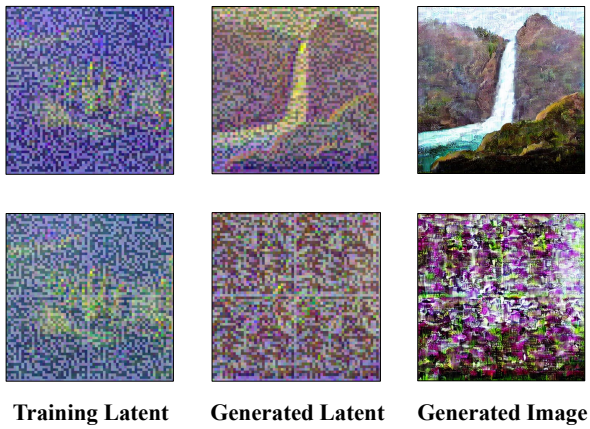


Figure 4: The first and second rows list results without and with periodicity, respectively. Latents are visualized using the first three channels as RGB for display, and the fourth channel is omitted. LoRA + DreamBooth is used to fine-tune model with adversarial examples.

### 3.3 Revisiting Anchor Design for LDMs

The Mist series (Zheng, Liang, and Wu 2023; Liang and Wu 2023; Zheng et al. 2023) has shown that the anchor plays a crucial role in adversarial attacks against LDMs. A carefully designed anchor can enhance the effectiveness of adversarial attacks. They explain the role of anchors from two perspectives: (i) target consistency, where all adversarial samples are pulled toward the same anchor in feature space, guiding the model to learn a unified adversarial direction with synergistic effects; and (ii) image-level disruption: suitable anchors introduce complex textures unrelated to the original image in human perception, causing the fine-tuned model to generate images with irrelevant textures, which significantly degrade the quality.

However, anchor design should not be limited to human perceptual disruption. While Mist's anchor performs well, we believe there remains room to further explore. That is, how can the anchor exploit specific properties of LDMs? We revisit anchor design and propose a deeper design strategy: an effective anchor should have strong perceptual disruptiveness and align well with the structural properties of LDMs to enable more efficient and powerful attacks.

### 3.4 Understanding the Vulnerabilities of LDMs

**Differences at Different Frequencies.** Inspired by SimAC (Wang et al. 2024), we analyze how latent-space frequency components affect the U-Net's behavior. For multiple artworks, we extract their initial latent representations $\{z_0\}$ and apply FFT-based frequency filtering to remove specific bands, generating the corresponding filtered latents $\{z_0^f\}$. We then add the same noise to both $\{z_0\}$ and $\{z_0^f\}$ at various timesteps $\{t\}$, resulting in noisy latents $\{z_t\}$ and $\{z_t^f\}$, which are fed into a pre-trained LDM to predict noise. We quantify the influence of each frequency component by computing the $\ell_2$ distance between the predicted noise from each $z_t^f$ and that from $z_t$. Notably, to ensure a fair comparison, we try to make the removed frequency bands contain approximately equal energy.

As shown in Figure 3, across all timesteps, although the energy of the high-frequency components is even lower than that of the low-frequency components, removing the high-frequency components still results in significantly higher prediction errors. This indicates that the U-Net relies more heavily on high-frequency signals for denoising. Based on this finding, we deliberately enhance the high-frequency components when designing the anchor, aiming to disrupt the original high-frequency features in the latent space while embedding erroneous high-frequency features.

**Differences at Different Periodicities.** Previous work (Liang and Wu 2023) observed that anchor images with periodic patterns in the pixel space substantially improve protection performance. Inspired by this, we introduced periodicity into the anchor within the latent space, improving the protection performance. As illustrated in Figure 4, we compare the effectiveness of the anchor with and without periodic structures (specifically, $2 \times 2$ tiling). Introducing periodicity results in significantly stronger attacks, indicating that U-Net is vulnerable to periodic patterns in the latent space.

**Similarities across Different Latent Channels.** While studying anchor design, we observed that the VAE latent typically consists of multiple channels. This raises a question: should each channel have its own anchor, or is one unified anchor better? To explore this, we tested the interaction among channels. Specifically, we selected multiple artworks and optimized their latent representations to approach a specific anchor, while applying loss to only one channel.

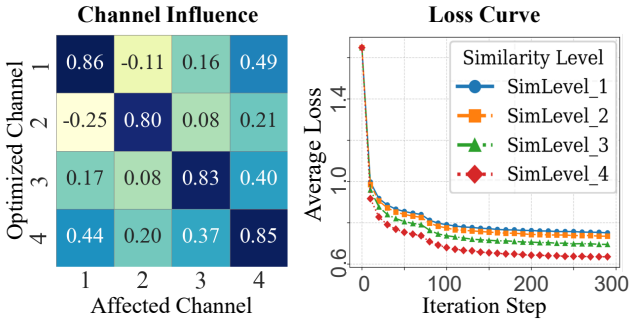


Figure 5: Channel consistency analysis results. Left: the heatmap showing the influence between channels. Right: the convergence curves.

The results are interesting. Although optimization was performed on one channel, most other channels also moved toward the same anchor to varying degrees, indicating channel synergy. To quantify this effect, we show a heatmap on the left of Figure 5, where each entry measures how much a non-optimized channel $j$ is influenced when only one channel $i$ is optimized. The value is defined as:

$$\text{Influence}_{i\to j} = 1 - \frac{\text{MSE}(z_j^{\text{after}}, z_{\text{anchor}})}{\text{MSE}(z_j^{\text{before}}, z_{\text{anchor}})}. \tag{4}$$

The $z_{\text{anchor}}$ denotes the single-channel anchor applied to the channel $i$. Higher values mean stronger alignment of channel $j$ toward the anchor. Our experiments were conducted on the commonly used 4-channel latent representation. As shown in the heatmap, most channels exhibit consistency, with only the channel pair $(1, 2)$ showing a weak opposite trend.

To validate whether this synergy can enhance protection performance, we designed a follow-up experiment to examine the impact of anchor consistency across channels on convergence. Under a limited perturbation budget, we adopted the optimization objective proposed in Equation (7), under a limited perturbation budget, to draw the original latent closer to the anchor. We defined four levels of anchor consistency (SimLevel 1-4). SimLevel 1 indicates that all four channels use distinct anchors; SimLevel 4 means all channels share the same anchor; SimLevel 2 and 3 represent cases where two and three channels share the same anchor, respectively. Experimental results show that higher levels of anchor consistency lead to improved convergence during optimization. The average convergence curves for each level are shown on the right of Figure 5. To ensure a fair comparison, all anchors were generated using our method in Section 3.5. We also ran multiple trials with different seeds to ensure reliability.

Based on these findings, we choose to use the same anchor for all channels to utilize channel synergy better while keeping the method efficient.

### 3.5 QRShield: Our Protection Method

Based on our analysis, we propose QRShield, a lightweight and practical protection framework with two stages: anchor generation and source image perturbation. The workflow is illustrated in Figure 2.

**Anchor Generation.** We aim to construct a latent anchor that is high-frequency and periodic. Specifically, we initialize a small learnable block $z_{\text{patch}}$ and tile it across the latent space to form a periodic structure $z_{\text{anchor}}$. To enhance sharp transitions and limit the complexity of the anchor representation, we apply a hard mask that constrains its values to $\{-1, +1\}$. This discretization reinforces high-frequency features and prevents the anchor from drifting into regions inaccessible to adversarial samples.

Given an artwork $x_0$, we encode it into the latent representation $z_0$. We consider not only the anchor's high-frequency components but, more importantly, those of the adversarial samples after alignment. So we pull $z_0$ toward the anchor using a convex combination:

$$z_0' = (1 - \lambda) z_0 + \lambda \cdot z_{\text{anchor}}. \tag{5}$$

We then optimize $z_{\text{anchor}}$ to maximize the high-frequency content of the perturbed latent $z_0'$. This is achieved by comparing $z_0'$ with its Gaussian-blurred version and defining the high-frequency loss as:

$$\mathcal{L}_{\text{hf}} = \|z_0' - \text{Blur}(z_0')\|_2^2. \tag{6}$$

By maximizing this loss, we maximize the high-frequency energy of $z_0'$.

It is worth noting that we optimize only one anchor for each artist, shared by all adversarial samples of their artworks. Specifically, the high-frequency loss is computed across the batch of artworks. This design is motivated by the idea in Section 3.3.

**Source Image Perturbation.** Due to the core idea of our method being the injection of specific anchor features into the latent space, we do not need the U-Net component. Instead, we perform our perturbations using only the VAE encoder. Specifically, we perturb the input by maximizing the similarity in the latent space between the perturbed sample and the anchor. This is implemented by the following loss:

$$\mathcal{L}_T(x_{\text{adv}}) = -\|E_\phi(x_{\text{adv}}) - z_{\text{anchor}}\|_2^2. \tag{7}$$

To enhance the transferability and robustness of the perturbations, inspired by translation-invariant attacks (Dong et al. 2019), we adopt a momentum-based translation-invariant PGD optimization strategy. The formula is as follows:

$$g = \mu \cdot g_{\text{prev}} + \frac{W * \nabla_{x_{\text{adv}}} \mathcal{L}_T(x_{\text{adv}})}{\mathbb{E}\left[|W * \nabla_{x_{\text{adv}}} \mathcal{L}_T(x_{\text{adv}})|\right] + \epsilon}, \tag{8}$$

$$x_{\text{adv}} \leftarrow \text{clip}_{x,\eta}\left(x_{\text{adv}} + \alpha \cdot \text{sign}(g)\right). \tag{9}$$

Here $W$ is a Gaussian kernel, $\alpha$ is the step size, and $\eta$ denotes the maximum perturbation budget.

## 4 Experiments

In this section, we conduct a comprehensive analysis of QRShield protection performance across multiple mainstream mimicry pipelines. The experiments are organized into three parts: (i) we comprehensively compare QRShield's protection effectiveness with that of various mainstream baselines; (ii) we analyze the individual contributions of each key module via ablation studies to verify the design rationale; (iii) we test the transferability of QRShield across different LDMs to demonstrate its generalization capability.

## 4.1 Experimental Setup

**Threat Model.** We evaluate QRShield under three representative generation pipelines widely adopted for style mimicry. (i) Full U-Net fine-tuning. It adjusts all U-Net parameters in an end-to-end manner. Though computationally expensive, it achieves strong style mimicry performance. (ii) LoRA (Hu et al. 2022) with DreamBooth (Ruiz et al. 2023). It is a widely used practical method that combines efficient parameter injection with personalized style mimicry. It enables high-fidelity results with low resource cost. (iii) Textual Inversion (Gal et al. 2022). It is a lightweight and easily deployable method that learns pseudo-token embeddings from a small number of images without requiring access to the model's internal weights, posing an increasing threat to art plagiarism. These pipelines span a wide range of accessibility, efficiency, and threat intensity, allowing for a comprehensive evaluation of QRShield's robustness and generalization in realistic and challenging mimicry scenarios.

**Datasets and Backbone Model.** We construct an evaluation dataset comprising 12 artists, including five contemporary artists selected from ArtStation (ArtStation 2025) and seven historical artists curated from WikiArt (Saleh and Elgammal 2015) to cover both low- and high-exposure scenarios. This setup enables us to evaluate the performance of our method in scenarios where the model has likely been exposed to an artist's style and where it has not. For each artist, we collect a total of 26 images, 16 used for training and 10 for testing. Our main experiments are conducted on Stable Diffusion 2.1, one of the most widely adopted LDMs.

**Baselines and Metrics.** We compare our method with several existing protection methods targeting LDMs, including classical methods AdvDM (Liang et al. 2023) and Glaze (Shan et al. 2023), strong methods Mist (Liang and Wu 2023) and ITA (Zheng et al. 2023), and recently emerging methods SimAC (Wang et al. 2024) and Anti-Diffusion (Zheng et al. 2025). For each artist, 16 images are selected for fine-tuning, followed by the generation of 60 test images using six fixed random seeds. Each seed produces 10 test images from 10 different prompts. To evaluate the protection performance, we compared images generated by clean models fine-tuned on clean data with those produced by models fine-tuned on adversarially perturbed data. Evaluation metrics include LPIPS (Zhang et al. 2018), FID (Heusel et al. 2017), MS-SSIM (Wang, Simoncelli, and Bovik 2003), and CLIP-SIM (Radford et al. 2021). In the tables, we use MS to denote MS-SSIM and CS to denote CLIP-SIM.

## 4.2 Main Results

We compare QRShield with other protection methods. The results are in Table 1. Experiments show that QRShield consistently achieves superior protection performance in three fine-tuning settings, outperforming all baselines. While some baselines perform well under specific fine-tuning settings, their effectiveness drops in other configurations. In contrast, QRShield consistently shows stable protection performance in all tested scenarios.

In addition, we evaluate the average generation time per sample and memory usage of several baselines on an NVIDIA RTX 4090 GPU. The results are shown in Table 2, which indicate that our method has better efficiency compared to other baselines. SimAC and Anti-Diffusion require more memory than the 24GB available on the RTX 4090 used in our experiments, and their overall resource consumption is high. As a result, they are not included in the efficiency comparison.

| Method | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| AdvDM | 0.54 | 141.64 | 0.48 | 0.91 |
| Glaze | 0.55 | 149.46 | 0.46 | 0.90 |
| Mist | 0.63 | 186.09 | 0.35 | 0.86 |
| ITA | 0.62 | 189.65 | 0.35 | 0.85 |
| Sim-AC | 0.52 | 142.51 | 0.49 | 0.91 |
| Anti-Diffusion | 0.52 | 138.02 | 0.49 | 0.91 |
| **QRShield (Ours)** | **0.69** | **306.64** | **0.28** | **0.70** |

(a) Full U-Net Fine-tuning

| Method | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| AdvDM | 0.61 | 222.77 | 0.41 | 0.83 |
| Glaze | 0.55 | 174.78 | 0.49 | 0.88 |
| Mist | 0.70 | 327.94 | 0.23 | 0.69 |
| ITA | 0.70 | 307.71 | 0.24 | 0.70 |
| Sim-AC | 0.60 | 237.38 | 0.42 | 0.82 |
| Anti-Diffusion | 0.61 | 220.22 | 0.41 | 0.83 |
| **QRShield (Ours)** | **0.76** | **382.69** | **0.18** | **0.62** |

(b) LoRA with Dreambooth

| Method | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| AdvDM | 0.66 | 246.92 | 0.31 | 0.75 |
| Glaze | 0.53 | 179.20 | 0.50 | 0.84 |
| Mist | 0.64 | 249.01 | 0.32 | 0.75 |
| ITA | 0.59 | 201.18 | 0.42 | 0.80 |
| Sim-AC | 0.67 | 269.48 | 0.29 | 0.70 |
| Anti-Diffusion | 0.62 | 218.63 | 0.36 | 0.78 |
| **QRShield (Ours)** | **0.69** | **306.49** | **0.23** | **0.68** |

(c) Textual Inversion

Table 1: Comparison with other baselines on different fine-tuning settings.

## 4.3 Ablation Study

To further validate the effectiveness of QRShield's design, we conducted ablation studies focusing on three key factors of QRShield: the anchor's frequency, its periodic structure, and whether the same anchor is shared in different latent channels. In each experiment, only one key factor was modified, while all other settings were kept at their default values unless otherwise specified. The experimental results under full U-Net fine-tuning are summarized in Table 3.

In the study on frequency, we observed that the anchor constrained by the hard mask (detailed in Section 3.5) converged very quickly during optimization, making it difficult to control its frequency distribution effectively. Therefore, we replaced the hard mask with a smoother $\tanh$ constraint

| Metric | ITA | Glaze | Mist | AdvDM | **QRShield (Ours)** |
|---|---|---|---|---|---|
| VRAM↓ | ∼6G | ∼8G | ∼16G | ∼16G | **∼3G** |
| TIME↓ | ∼30s | ∼33s | ∼58s | ∼56s | **∼11s** |

Table 2: Comparison with other baselines on protection efficiency.

| Variant | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| No Freq. | 0.58 | 153.11 | 0.44 | 0.90 |
| Low Freq. | 0.66 | 280.91 | 0.31 | 0.74 |
| Mid Freq. | 0.68 | 304.34 | 0.30 | 0.71 |
| High Freq. | **0.69** | **306.64** | **0.28** | **0.69** |

(a) Results of Different Frequencies

| Variant | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| $1 \times 1$ | 0.60 | 178.52 | 0.41 | 0.86 |
| $2 \times 2$ | **0.69** | **306.64** | **0.28** | **0.69** |
| $4 \times 4$ | 0.67 | 226.90 | 0.32 | 0.81 |
| $8 \times 8$ | 0.65 | 188.82 | 0.34 | 0.83 |

(b) Results of Different Periodicities

| Variant | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| Consistency | **0.69** | **306.64** | **0.28** | **0.69** |
| Inconsistency | 0.66 | 274.83 | 0.29 | 0.75 |

(c) Results of Anchor Consistency vs. Inconsistency

Table 3: Ablation results under the full U-Net fine-tuning setting. Top to bottom: frequency ablation, periodicity ablation, and anchor consistency ablation.

for this experiment and initialized the anchor as a zero vector to regulate its frequency characteristics better. We evaluated the results at 0, 50, and 100 iterations of anchor optimization, corresponding to no, low, and middle frequency levels, respectively. The High Freq. setting corresponds to the anchor generated under the original hard mask constraint, which exhibits a higher frequency than the other settings. This setup allowed us to analyze the effect of frequency variation on protection capability. The experimental results show that increasing the anchor frequency can enhance protection performance, which validates our analysis of U-Net's frequency sensitivity.

In the study on periodicity, we investigated the presence and intensity of periodic structures in the anchor and their effects on protection performance. We tested four periodic configurations: $1\times1$, $2\times2$, $4\times4$, and $8\times8$, where $n\times n$ indicates that the anchor consists of $n\times n$ repeatedly tiled blocks. Notably, the $1 \times 1$ setting corresponds to a non-periodic configuration where the entire anchor consists of a single block. The results show that periodicity has a strong impact on protection performance. The $2 \times 2$ configuration achieves the best performance and is thus adopted as our final setting.

In the study on anchor consistency, we examined the effect of whether different latent channels share the same anchor on protection performance. We designed two comparative experiments. Specifically, in the Consistency setting, all channels share the same anchor. By contrast, in the Inconsistency setting, each channel's anchor is optimized independently without enforcing consistency. The results indicate that leveraging consistency among VAE channels can further improve protection performance.

In summary, the ablation studies validate the critical roles of frequency, periodicity, and consistency in QRShield and demonstrate that these three factors jointly contribute to enhanced protection performance.

| Backbone → Target | LPIPS↑ | FID↑ | MS↓ | CS↓ |
|---|---|---|---|---|
| SD2.1 → SD2.1 | 0.69 | 306.64 | 0.28 | 0.69 |
| SD2.1 → SD1.5 | 0.68 | 286.06 | 0.27 | 0.75 |
| SD2.1 → SD1.4 | 0.67 | 262.55 | 0.29 | 0.78 |
| SD1.5 → SD2.1 | 0.69 | 306.68 | 0.28 | 0.69 |
| SD1.5 → SD1.5 | 0.68 | 284.69 | 0.27 | 0.75 |
| SD1.5 → SD1.4 | 0.67 | 268.43 | 0.28 | 0.77 |
| SD1.4 → SD2.1 | 0.69 | 313.34 | 0.28 | 0.69 |
| SD1.4 → SD1.5 | 0.68 | 294.31 | 0.27 | 0.75 |
| SD1.4 → SD1.4 | 0.67 | 263.50 | 0.29 | 0.77 |

Table 4: Cross-model transferability of QRShield under the full U-Net fine-tuning setting. Each row shows the performance when adversarial examples are generated with the "Backbone" model and then used to attack the "Target" model.

### 4.4 Cross-Model Transferability Evaluation

Our method relies on the VAE parameters of a specific LDM, which makes it a white-box method. To evaluate its effectiveness in black-box settings, we examine the cross-model transferability of our method in this section. Specifically, we conduct experiments on three versions of Stable Diffusion: Stable Diffusion 1.4, 1.5, and 2.1. These models are based on the LDM framework, but differ in their training data, architecture designs, and key component parameters. As widely used open-source models for text-to-image generation, they are strong representatives for testing cross-model transferability. The results in Table 4 show that our method remains effective even across different models.

## 5 Conclusion

This paper reveals the channel consistency of the VAE and the sensitivity of the U-Net to high-frequency and periodic patterns within LDMs. Motivated by these findings, we propose an efficient protection method: QRShield. Experimental results show that QRShield is more efficient and effective than existing baselines. It provides a viable and effective tool for preventing AI art plagiarism. While our investigation centers on U-Net-based LDMs, future research may explore extending QRShield's core ideas to non-U-Net-based LDMs or other generative models, thereby broadening its applicability for protecting a wider range of artistic and intellectual works.

## Acknowledgments

We are very appreciative of the anonymous reviewers' time and effort in offering thoughtful comments and valuable suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62206318), the Guangzhou Intelligent Educational Technology Collaborative Innovation Center Construction Project (Grant No. 2023B04J0004), and the Research Project on Large Model-Based Higher Education Institution Development Data Governance Technology (Grant No. 2025B04J0037).

## References

ArtStation. 2025. ArtStation. https://www.artstation.com. Accessed: 2025-5-16.

Deng, Y.; Wu, W.; Zhang, J.; and Zheng, Z. 2023. Blurred-Dilated Method for Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 36, 58613–58624.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4312–4321.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. ArXiv preprint arXiv:1312.6114, arXiv:1312.6114.

Liang, C.; and Wu, X. 2023. Mist: Towards Improved Adversarial Examples for Diffusion Models. arXiv:2305.12683.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial example does good: preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the 40th International Conference on Machine Learning*, 20763–20786.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. arXiv:2209.14988.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Saleh, B.; and Elgammal, A. 2015. Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. arXiv:1505.00855.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2187–2204.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.

Wang, F.; Tan, Z.; Wei, T.; Wu, Y.; and Huang, Q. 2024. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12047–12056.

Wang, Z.; Simoncelli, E.; and Bovik, A. 2003. The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. *IEEE, Piscataway, NJ*, 2: 1398.

Wu, H.; Ou, G.; Wu, W.; and Zheng, Z. 2024. Improving Transferable Targeted Adversarial Attacks with Model Self-Enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24615–24624.

Wu, W.; Cao, Y.; Yi, N.; Ou, R.; and Zheng, Z. 2025a. Detecting and Reducing the Factual Hallucinations of Large Language Models with Metamorphic Testing. *Proceedings of the ACM on Software Engineering*, 2(FSE): 1432–1453.

Wu, W.; Hu, H.; Fan, Z.; Qiao, Y.; Huang, Y.; Li, Y.; Zheng, Z.; and Lyu, M. 2025b. An Empirical Study of Code Clones from Commercial AI Code Generators. *Proceedings of the ACM on Software Engineering*, 2(FSE): 2874–2896.

Wu, W.; Wang, Z.; Luo, Z.; Chen, W.; and Zheng, Z. 2025c. Detecting Violations of Physical Common Sense in Images: A Challenge Dataset and Effective Model. In *ACM International Conference on Multimedia*, 10945–10954.

Xue, H.; Liang, C.; Wu, X.; and Chen, Y. 2023. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, B.; Liang, C.; and Wu, X. 2023. Targeted Attack Improves Protection against Unauthorized Diffusion Customization. ArXiv preprint arXiv:2310.04687, arXiv:2310.04687.

Zheng, B.; Liang, C.; Wu, X.; and Liu, Y. 2023. Understanding and improving adversarial attacks on latent diffusion model. https://openreview.net/pdf?id=yNJEyP4Jv2. OpenReview preprint.

Zheng, L.; Xie, L.; Zhou, J.; Wang, X.; Wu, H.; and Tian, J. 2025. Anti-Diffusion: Preventing Abuse of Modifications of Diffusion-Based Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10582–10590.