Rethinking the Invisible Protection against Unauthorized Image Usage in Stable Diffusion

Shengwei An*, Lu Yan*, Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Xiangyu Zhang
Purdue University

Abstract
Advancements in generative AI models like Stable Diffusion, DALL·E 2, and Midjourney have revolutionized digital creativity, enabling the generation of authentic-looking images from text and altering existing images with ease. Yet, their capacity poses significant ethical challenges, including replicating an artist’s style without consent, the creation of counterfeit images, and potential reputational damage through manipulated content. Protection techniques have emerged to combat misuse by injecting imperceptible noises into images. This paper introduces INSIGHT, a novel approach that challenges the robustness of these protections by aligning protected image features with human visual perception. By using a photo as a reference, approximating the human eye’s perspective, INSIGHT effectively neutralizes protective perturbations, enabling the generative model to recapture authentic features. Our extensive evaluation across 3 datasets and 10 protection techniques demonstrates its superiority over existing methods in overcoming protective measures, emphasizing the need for stronger safeguards in digital content generation.

1 Introduction
Generative AIs such as Stable Diffusion [26], DALL·E 2 [25], and Midjourney [9] have taken the world by storm because of their superb capability of generating authentic-looking images from only a few words. A stunning “photograph” can be captured without venturing into the wilderness to search for the perfect scene. Instead, it can be achieved by describing all the essential elements required within the scene. If one takes a photo with some flaws but is not a master of image editing tools such as Photoshop, they can simply demand the generative AIs to repair it. Generative AIs also exhibit a level of painting expertise that rivals that of the finest human artists. A breathtaking “artwork” can be made by naming the genre (e.g., the brushstroke and color palette) and describing the content to the AIs.

*Equal contribution.
attacker downloaded those paintings to fine-tune a Stable Diffusion model to produce counterfeits. The image at the top left corner of Figure 1 (annotated with “Unprotected” and “Training”) denotes a real painting from Vincent used to fine-tune a Stable Diffusion model and the image below is a counterfeit generated by the model. Looking at the shape of the flowers and the vase as well as the impasto brushstrokes, one has enough confidence in believing they are both from Vincent. To make a profit, the attacker sold them at a lower price. This could completely destroy Vincent’s life and career.

Realizing these attacks, researchers make every effort to devise protection methods to prevent images from being misused. One of the most famous tools is Glaze [36] that added invisible perturbations to the painting to prevent the Stable Diffusion model from learning the correct painting style. Vincent decided to use it to protect his artwork. But shortly, he found there were new counterfeits because the recently proposed countermeasure Impress [3] can purify the protected images to disable the protective effect. Before he gave up, he found another tool Mist whose protection cannot be removed by Impress as demonstrated by the second and third columns in Figure 1. In particular, the attacker fine-tuned a Stable Diffusion model on the Mist-protected paintings but found the generated image (in the second column) was of low quality. Then the attacker tried to use Impress to purify the protected paintings before training a model on them, but the generated image (in the third column) is still chaotic. Now, Vincent is satisfied with the protection and publishes the Mist-protected paintings online.

However, is this invisible-perturbation-based protection really sufficient? To better answer this question, we study the inherent properties of existing protections. All of these methods are based on adding invisible noises [14, 15, 36, 44, 46, 49, 53], with which the models can no longer learn the correct features. The invisibility requirement is because they don’t want humans to perceive the difference between unprotected and protected images. Otherwise, the market values of the paintings may greatly degrade. In other words, these noises can only fool AI models but not humans. If we can transfer this human invisibility to the models by aligning the features of the protected images in the lens of models and the eyes of humans, the added noises would be neutralized, and the model should be able to pick up the correct features (as humans). With this insight, we propose a new approach called INSIGHT to evaluate the robustness of existing protections.

More specifically, we propose to use the photo of a protected image as the alignment reference. The photo can be considered a good approximation of human visual perception because of the similarity between the structures of the eyes and the design of a camera [23, 39, 45]. We observe the architecture of a Stable Diffusion model contains two components. Therefore we align the features of the protected image with its reference via a contrastive loss in each component. This is achieved by finding a new perturbation to counteract the protective perturbation. The rightmost column in Figure 1 shows our approach can invalidate the protection. The attacker first uses our approach to purify the protected paintings and then fine-tune a Stable Diffusion model on them. The generated image in the second row resembles Vincent’s style again.

Besides the style mimicry attack, attackers can also create fake images of certain identities using Stable Diffusion models, including subject recontextualization and image manipulation attacks. These fake images can damage the subject’s reputation (e.g., a fake photo of the person in prison) or even cause significant financial loss. For example, fake generations made a multinational company lose more than 25 million dollars in this February [4]. Details of different attacks are in Section 2.2. We want to emphasize that our goal is not to facilitate attacks, but to provide a tool to evaluate protections and help build stronger protections. Our contributions are summarized as follows:

- We analyze the design space of existing protections and countermeasures (that aim to remove protections). We point out the invisible perturbation is not robust. We observe that the ineffectiveness of existing countermeasures is because they fail to constrain both components of the Stable Diffusion models and lack a good reference (to facilitate removing protective noises).

- We design a new countermeasure against protections. It exploits the invisibility of the added perturbation. It invalidates protection by aligning the features of the protected images with the human visual reference (e.g., a photo) in the two components of a Stable Diffusion model.

- We build and open-source a tool INSIGHT [21] (Rethinking the Invisible Protection against Unauthorized Image Usage in Stable Diffusion) and extensively evaluate it against 10 existing protections on three datasets. We compare our approach against 4 baselines including the SOTA Impress in 3 types of attacks (i.e., style mimicry, subject recontextualization, and image manipulation). Experimental results show our approach can outperform all baselines against all protections in all attacks. For style mimicry, our approach has 1.4x the effectiveness of the best baseline. For the other two attacks, our approach is also the most effective and can help generate images of the best quality. We also conduct human studies and GPT4-based studies. On average, users/GPT4 prefer our approach to the best baseline in 93.9%/94.2% cases. We also show the effectiveness of our approach in the commercial service.

**Threat Model.** We use the same threat model as the existing countermeasure [3]. The defenders want to use invisible
perturbations to protect the images from being misused. The defender can utilize arbitrary image transformation methods and pre-trained diffusion models. The goal of the attackers is to diminish the protection to misuse images. They have no knowledge of the exact protection method or the corresponding clean (unprotected) images. But they have complete control over the protected images and thus can utilize countermeasures to remove protection. They can exploit pre-trained diffusion models.

Ethics. Our human study has received approval from our institutional review board (IRB). It is collected anonymously via Amazon Mechanical Turk. We don’t collect or store any personally identifiable information.

2 Latent Diffusion Model and Image Misuse

2.1 Latent Diffusion Model

Figure 2 shows the architecture of the Latent Diffusion Model (LDM) or Stable Diffusion (SD) [26]. Different from existing Diffusion Models (DM), such as the Denoising Diffusion Probabilistic Model [7] and Noise Conditional Score Network [41], that operate in the pixel domain, LDM runs the diffusion process in a compressed latent space created by a Variational Autoencoder (VAE) consisting of an encoder E and a decoder D [5]. E encodes an image x from the pixel space into the latent space as z^0 = E(x) while D decodes a latent vector to recover an image to the pixel space. The encoder and decoder compose the so-called first stage, while the diffusion process in the latent space is the second stage.

The forward process (e.g., z^0, z^1,..., z^T) in LDM iteratively adds Gaussian noises to z^0 until it becomes a Gaussian noise, that is, z^T ∼ N(0, I). The training goal of LDM is to learn a network Mθ (usually a UNet) to form a reverse process to iteratively denoise the Gaussian noise z^T to recover z^0 according to the defined distribution:

\[ z^{-1} \sim N(z^{-1}; \frac{1}{\alpha^t} \left( z^{-1} - \frac{1 - \alpha^t}{\sqrt{1 - \alpha^t}} M_\theta(z^{-1}, t) \right), \sigma'^t I) \]  

where α^t is the transitional content schedulers mathematically defined in DM and \( \sigma'^t = \prod_{i=1}^t \alpha^t \) [7, 26].

One advantage of LDM is that it can utilize an additional condition c such as text prompts to guide the reverse process to generate the desired images\(^2\). For text-to-image generation (Stable Diffusion Models), each training sample is a pair of an image x and a text prompt c. This conditional generation is usually implemented as extra attention layers in Mθ. Intuitively, the conditional training goal is to approximate the added noise ε\^t ∼ N(0, I) at the t-th step, and the simplified training loss is as follows.

\[ L_{LDM} = E_{\epsilon_1, \epsilon_2, \epsilon_3} || \epsilon - M_\theta(\sqrt{\alpha^t} \epsilon + \sqrt{1 - \alpha^t} \epsilon(c, t), t) \|^2_2 \]  

where t is sampled from \([1, T]\), \( \epsilon = E(x) \), \( \epsilon(c, t) \sim N(0, I) \).

After training, to generate an image for a given text prompt c, a latent vector z^T is first sampled from the Gaussian distribution and progressively denoised to get z^0. z^0 is further decoded by D to synthesize the final image \( \tilde{x} = D(z^0) \).

2.2 LDM-based Image Misuse

The high capability of LDM in few-shot image generation and manipulation empowers non-experts to create amazing artwork with only a few lines of text. These fantastic models and techniques contribute to the blossom of generative AIs. However, once used by malicious attackers, they also brought a lot of concerns about image misuse. There are three main misuse scenarios. Each of them can cause severe damage.

Style Mimicry. The first type of misuse is style mimicry which has concerned a lot of researchers and artists [36]. The malicious attackers want to generate synthetic paintings with the same style as certain artists to make profits. Figure 3 provides an illustration. The victim artist (e.g., Vincent van Gogh) draws some paintings and advertises them online. After downloading them, the attackers use DreamBooth [28] to fine-tune a pre-trained Stable Diffusion model. For example, they can force the model to associate the text prompt “a [V] painting” with the victim’s painting style. They can then use this model to generate “new” paintings of Vincent as long as they insert “a [V] painting” into the prompt. Comparing the

\(^2\)Another network will extract embedding of the text prompt and feed it to UNet. We omit this step here for simplicity.
Figure 4: Subject recontextualization. Attackers use the victim’s (e.g., Adam Michnik) photos to train a Stable Diffusion model to create the victim in new contexts (e.g., Eiffel Tower).

Figure 5: Image manipulation. Attackers use a Stable Diffusion model to manipulate the unmasked (dark) area of the victim’s (e.g., Roger Federer and Trevor Noah) photo.

images on the right and left in Figure 3, it looks like they were painted by the same artist. Attackers don’t need to spend years learning and practicing painting. What it takes is just a few clicks and several minutes. This can have negative impacts on artists, students, and even the whole society.

Subject Recontextualization. The Stable Diffusion model is not only able to mimic painting styles but also able to memorize subjects. This enables the second type of image misuse known as subject recontextualization as demonstrated in Figure 4. The goal of this attack is to create the subject in a new context. The created subject is usually in new poses, different facial expressions, and unseen clothes. The attackers first obtain a set of the victim’s photos. They are either taken by the attackers or retrieved from online resources. The attackers then use DreamBooth [28] to make the model connect the text prompt “a [V] person” with the victim. As a result, the attackers can generate photos of the victim in front of the Eiffel Tower using the prompt “a [V] person in front of the Eiffel Tower”. When the victim is put into some wicked context such as being in prison or arrested by the police, news with such fake high-fidelity photos can produce a catastrophe.

Single Image Manipulation. Different from the previous two, the attackers want to manipulate a single photo, as shown by Figure 5. They mask the photo to denote the unchanged area and ask the Stable Diffusion to repaint the other area according to the provided text [6, 32]. In this case, they don’t use a set of photos to fine-tune a Stable Diffusion model. Instead, they use a pre-trained model to complete the manipulation. This makes the attacks easier to conduct and cause damage. Different from the previous attack, it usually preserves the same subject (e.g., the face pixels). The two fake photos in Figure 5 are crafted using the prompts “men in casino” and “men in prison”. The high fidelity of the photos can persuade people to believe the two people are indeed in prison.

There are multiple differences between the last two types of attacks. First, subject recontextualization requires training a model using several images while single image manipulation utilizes a pre-trained model. Second, the former needs to memorize the subject and then recreate it in a new context. The created subject usually has different poses, facial expressions, and clothes from the training images. For example, Adam in the fake photo of Figure 4 is smiling, wearing a backpack, and has a smaller head size. All of these features are different from the training photos. However, image manipulation does not memorize anything and directly modifies the unmasked area. That is, the pixel values in the masked area are copied from the real image. For example, the two images on the right side have the same heads as the real image on the left in Figure 5.

3 Existing Protection Methods

Attacks being proposed, researchers devise different protection methods to mitigate them. They add protective perturbations $\delta$ to the original photos and publish the protected ones instead. We have seen that LDM consists of two stages/components: the VAE ($E$ and $D$) stage, and the diffusion stage with $M_0$. In the following, we briefly explain existing protection methods based on which component they explicitly target (summarized in Table 1). There are mainly two categories: 1) constraining VAE only, and 2) constraining both VAE and UNet. There is no existing protection only constraining UNet because one has to go through VAE to constrain UNet in the latent space.

Constraining VAE Only. This type of protection tries to mislead the VAE to regard the protected image $x^\text{protected} = x + \delta$ as a carefully selected target image $x_{\text{target}}$ [30, 36] by minimizing the distance between the latent vector of $x$ and that of $x_{\text{target}}$: $\delta = \arg\min_\delta ||E(x + \delta) - E(x_{\text{target}})||_2^2$. Glaze [36] first transforms the painting into a genre different from the ground truth one (e.g., from Post Impressionism to Cubism) and uses
that as $x_{\text{target}}$, while the encoder protection of PhotoGuard [32] (denoted by PG_E) uses a gray image.

**Constraining both VAE and UNet.** We can also achieve protection by misleading the UNet. Because UNet in an LDM is defined in the latent space, constraining UNet cannot avoid considering VAE. Based on the space where the optimization loss is defined, protection of this type can be further subdivided into three categories. The first one is the diffusion protection in PhotoGuard [32] (denoted by PG_D). It defines the loss in the pixel space and thus implicitly considers the VAE and UNet. It wants the image generated by LDM based on the protected one $x_{\text{protected}} = x + \delta$ to be similar to a target image (e.g., a gray image): $\delta = \text{argmin}_\delta \| f_{\text{LDM}}(x + \delta) - x_{\text{target}} \|_2$, where $f_{\text{LDM}}$ denotes the whole LDM generation process.

The second subdivision defines the optimization loss in the UNet component, including AdvDM+ [15], SDS± and AdvDM- [49], and Anti-DreamBooth [44]. AdvDM+ forces the UNet’s output for $x_{\text{protected}}$ to deviate from the ground truth: $\text{argmax}_\delta E_{x,t} \| \epsilon' - M_0(\sqrt{\delta} e_0 + \sqrt{1-\delta} \epsilon', c, t) \|_2$. Anti-DreamBooth [44] also updates the LDM to improve the protection effect against DreamBooth [28]. SDS+ reduces the time cost and GPU memory required by AdvDM+ via approximating the gradients propagated from the UNet. SDS- and AdvDM- use a similar loss with modified directions.

The third subdivision explicitly defines the loss in the VAE component in addition to the UNet, including Mist [14], ITA [53], and SDS- [49]. They want to mislead both the VAE and UNet at the same time. Mist can be considered as AdvDM+ with PG_E. ITA also updates LDM as Anti-DreamBooth. SDST is Mist with the gradient approximation.

**Perturbation Bound.** To avoid destroying the original image, the protection methods bound the perturbation to ensure it’s invisible. Glaze [36] uses a learned perceptual image patch similarity (LPIPS) [51] regularization term to penalize large perturbation while the others clip the perturbation according to predefined distance $\ell_\infty$ or $\ell_2$ [14, 15, 32, 44, 49, 53].

## 4 Limitations of Existing Countermeasures (Against Protection Techniques)

This section discusses four existing countermeasures to the aforementioned protection methods: JPEG compression, Gaussian noise, Crop+resize and Impress [3]. They aim to nullify the protective perturbations such that the adversary can succeed. The first three methods are based on transformations evaluated in existing literature [3, 14, 33, 44, 49]. They don’t utilize any information from LDM, they constrain neither VAE nor UNet. Impress is the state-of-the-art LDM-based method against Glaze and PhotoGuard. It implicitly constrains VAE as explained later. Table 2 summarizes the used strategy of each countermeasure.

Figure 6 visualizes their performance against Mist in style mimicry and indicates that improvement is still needed. The first row shows one training sample for each case. Each column shows one type of training data to fine-tune an SD. The “Clean” column fine-tunes an SD using unprotected paintings (here they are painted by Vincent). The images generated by the fine-tuned SD model are displayed in the second and third rows. We can see the generated images and the clean paintings both feature similar choppy and expressive brushstrokes and lines. Thus without any protection, style mimicry is easy. The second column shows the result when the training paintings are protected by Mist. Because it only adds small perturbations, there is no significant visual difference between the first two images in the first row. However, the generated images (in the second column of the second and third rows) are completely destroyed and have obvious Mist patterns. This is a successful protection.

We now examine the performance of the first three transformation-based methods. JPEG compresses an image in a lossy way and has been shown to be able to remove some protection effects by researchers [3, 33]. We follow their setting and use 15 as the quality factor. We first use JPEG to compress the protected paintings and use the compressed ones to fine-tune an SD model to generate new paintings. As shown by the third column in Figure 6, it removes some protection so that the generated images start to show the similar contents as the clean ones. However, the brushstrokes and color palettes are different. Another baseline is Gaussian noise, which adds Gaussian noises to the protected image. It does not work and sometimes even strengthens the protection as indicated by the fourth column.

Crop+resize has been empirically proven by researchers to be the most effective post-processing (non-LDM-based) method of removing protection compared with JPEG compression and Gaussian perturbation [3, 14, 49]. We follow the literature to run Crop+resize by first cropping 64 pixels in all directions around the protected 512×512 image and then resizing it back to 512×512 [14, 49]. An SD model is fine-tuned on the cropped and resized images. The results of Crop+resize are listed in the fifth column. There is some improvement as the semantic contents and some wavy brushstrokes are mostly recovered. However, we can see that the painting of flowers has obvious strange patterns and the painting in the last row lacks fine brushstrokes and uses a different color scheme. Therefore, it cannot completely remove the protection.

<table>
<thead>
<tr>
<th>Comp. JPEG</th>
<th>Gaussian</th>
<th>Crop+resize</th>
<th>Impress</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>UNet</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

## Table 2: Strategies of countermeasures. The first three are post-processing methods independent of diffusion models.
Figure 6: Examples of existing countermeasures for Mist. The first row shows samples used to train the Stable Diffusion model. The second and third rows show generated samples. Each column denotes a different setting.

Impress [3] is the SOTA countermeasure against Glaze and PhotoGuard. It observed that the image reconstructed by VAE \( D(E(x_{\text{protected}})) \) deviates from \( x_{\text{protected}} \) noticeably whereas a clean image (used in training) is close to its reconstruction, i.e., \( x_{\text{clean}} \approx D(E(x_{\text{clean}})) \). Therefore, it proposes to remove the protection effect (in \( x_{\text{protected}} \)) by finding another perturbation \( \delta \) such that \( x_{\text{protected}} + \delta \) is close to \( D(E(x_{\text{protected}} + \delta)) \). From its optimization goal, we can see it only implicitly considers VAE but does not explicitly enforce anything inside of the VAE or UNet. The sixth column in Figure 6 lists the \( x_{\text{protected}} + \delta \) in the first row. An SD is fine-tuned on the Impress-purified images. As shown by the generated images, instead of removing the protection noises, it undesirably fortifies them and the MIST pattern is more obvious now. Actually, the image processed by Impress in the first row already starts to have that pattern. Impress is effective on Glaze and PhotoGuard because the larger inconsistency produced by them can only be mitigated by removing the protection noises. However, the protection method Mist leverages a smaller inconsistency. Hence, without any other guidance, instead of removing the protection noises from \( x_{\text{protected}} \), the optimization in Impress tends to inject similar noises into \( x_{\text{protected}} + \delta \) to make the reconstruction consistent. As a result, Impress enhances the Mist protection instead of removing it.

As indicated by Table 1 and Table 2, different protection methods can exploit different components while existing countermeasures do not fully cover both VAE and UNet. Recall that the protection methods add invisible noises to make the protected image divert from the original one from LDM’s perspective. Because of this invisibility, we propose to craft a visual reference in the eyes of a human and use it to explicitly guide resolving the divergence produced by protection perturbation in both VAE and UNet to overcome the shortcomings of existing countermeasures. The rightmost column in Figure 6 lists our results. The styles of the two synthetic images are very similar to the two synthetic ones in the first column since the brushstrokes and color schemes are similar, indicating our method successfully removes the protections.

5 Our Design

Our goal is to remove the protection added to \( x_{\text{protected}} \) by explicitly resolving the divergence brought by the protection perturbation in both VAE and UNet. It takes a protected image \( x_{\text{protected}} \) and its visual reference \( x_{\text{visual}} \) (e.g., by taking a picture of the protected image, disrupting some of the well crafted perturbations) and produces an aligned image \( x_{\text{aligned}} \) (i.e., with protection disabled to some extent). It derives the output by aligning with \( x_{\text{visual}} \) in the latent space created by VAE and the intermediate diffusion steps. Figure 7 presents an overview of the whole attack pipeline. The left part of the figure shows our optimization framework to remove the protection. The right part shows an example of style mimicry. If attackers directly train a Stable Diffusion model using \( x_{\text{protected}} \), the model can only generate useless images. If attackers first use our method to remove the protection and train a model on \( x_{\text{aligned}} \), they can fulfill the style mimicry. We will elaborate on the five crucial elements in our framework in this section: 1) visual reference \( x_{\text{visual}} \) providing guidance, 2) LPIPS loss constraining the allowed amount of changes, 3) UNet contrastive alignment explicitly utilizing the guidance as the positive direction and the protected sample as the negative direction in the UNet space, 4) VAE contrastive alignment similar to the previous one but being applied in
Therefore, we would like to involve such guidance to diminish methods.

Figure 7: Overview of our pipeline. The left part shows our optimization framework that takes in a protected image and outputs an aligned (i.e., purified) one with the protection removed. The right part shows the image misuse attack (e.g., style mimicry).

Figure 8: Cases where Photo is effective. The first/second row shows the training/generated images. The odd/even columns denote the settings of protections/photos.

Figure 9: Effects of the UNet and VAE alignment. The first/second row shows the training/generated images. Each column denotes a different setting.

the VAE stage, and 5) reconstruction alignment reducing the inconsistent features recognized by LDM.

5.1 Visual Reference

Our intuition is that the added invisible noises by various protection methods only confuse the LDM but not humans. Therefore, we would like to involve such guidance to diminish the protection effect by aligning the LDM’s and human’s cognition on the protected image. To approximate what the painting looks like in human eyes, we choose to use a photo captured by the camera, since researchers consider camera can faithfully reflects humans’ visual perception [23, 39, 45].

A naive solution is to directly use this visual reference to replace the protected image in SD fine-tuning. Figure 8 shows its effectiveness. The first row shows the training images used to fine-tune the SD model. The second row shows the generated images. The first two columns show its effectiveness on Mist-protected images. The attacker wants to generate a new flower painting by mimicking Vincent’s style. The first column fine-tunes the SD model on Mist-protected images and generates images with chaotic content covered by the “MIST” pattern. In the second column, we first take photos of Mist-protected images and use them to fine-tune the SD model. The generated image draws the flowers in Vincent’s style and thus is a good mimic. The third and fourth columns show a similar case when the protection method is AdvDM+. The attackers try to generate a painting of a person in a suit. The AdvDM+ results in very strange color clusters and the Photo’s result is much better, suggesting the transformations through the physical world by taking pictures are effective in removing the digital protective noises.
Training

<table>
<thead>
<tr>
<th>Clean</th>
<th>Glaze</th>
<th>Clean</th>
<th>Glaze</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_\infty = 141/255, \ell_{\text{LPIPS}} = 0.0540$</td>
<td>$\ell_\infty = 180/255, \ell_{\text{LPIPS}} = 0.0504$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 10: $\ell_\infty$ and $\ell_{\text{LPIPS}}$ between the clean training image and the corresponding protected training one. The average $\ell_\infty$ and $\ell_{\text{LPIPS}}$ on 24 random pairs are 174/255 and 0.0555.

However, it’s not sufficient in some cases as shown by the second column in Figure 9. The text prompt is “a river with a bridge and a statue”. When using Mist-protected paintings to fine-tune an SD model to generate images, there is no meaningful content. When photos are used as the training data, the general scene starts to appear. We can see the blue sky, green trees, the river, and even a statue in the middle. However, some Mist patterns are still there. This necessitates our further optimization using the reference. Note that our optimization method needs a lossy positive reference that can preserve the semantic content. We choose to use photos, but our method is not limited to working only with photos.

5.2 Visual Bound

Before starting the optimization, we need to define a way to confine the search space so that the final result is still visually similar to $x_{\text{protected}}$. This requirement is because of the invisible property of the added protective noises. Many existing works bound the perturbation in the pixel space by up to 16/255 [14, 35] or even only 4/255 [53]. This bound is not sufficient for removing the protective noises. Figure 10 shows the two pairs of a clean training image and the corresponding Glaze-protected one. We list the $\ell_\infty$ and $\ell_{\text{LPIPS}}$ distances between them. The $\ell_\infty$ of the first pair is 141/255. This bound is too large to constrain meaningful optimization in the pixel space. Because LPIPS is based on the distance between the internal features of a pre-trained network and has been shown to best match human perception, we choose to use it to constrain the visual change. More specifically, we add a regularization term during the optimization:

$$L_{\text{bound}} = \max(\ell_{\text{LPIPS}}(x_{\text{protected}} + \delta, x_{\text{protected}}) - \Delta, 0)$$

where $\Delta$ denotes the bound. If the optimization stays within the $\Delta$ distance, it’s 0 and adds no penalty. Otherwise, this loss penalizes the overlarge visual difference. We use the same budget 0.1 as in the literature [3].

5.3 UNet Contrastive Alignment

Note that although the photo usually cannot completely corrupt the protective noises, it must lose some of it. As such, it can be used as a guidance for the direction of optimization, instead of the ground truth. As existing work shows the diffusion space (i.e., UNet component) corresponds to the generation of the semantic content [14, 15], we propose to use the contrastive alignment in the UNet component to pull the $x_{\text{aligned}}$ (i.e., $x_{\text{protected}} + \delta$) in the same direction of $x_{\text{visual}}$ and push it in the opposite direction of $x_{\text{protected}}$. More specifically, at the $t$-th step, we want the transition of $z_{\text{aligned}}^{t} \rightarrow z_{\text{aligned}}^{t+1}$ to be similar to $z_{\text{visual}}^{t} \rightarrow z_{\text{visual}}^{t+1}$ and different from $z_{\text{protected}}^{t} \rightarrow z_{\text{protected}}^{t+1}$. To how the DM training loss is derived, we can encode the contrastive alignment as follows:

$$\min_\delta ||M_\theta(z^{t}_{\text{aligned}}, c, t) - M_\theta(z^{t}_{\text{visual}}, c, t)||^2_2$$

$$\max_\delta ||M_\theta(z^{t}_{\text{aligned}}, c, t) - M_\theta(z^{t}_{\text{protected}}, c, t)||^2_2$$

which can be combined into a loss:

$$L_{\text{UNet}} = ||M_\theta(z^{t}_{\text{aligned}}, c, t) - M_\theta(z^{t}_{\text{visual}}, c, t)||^2_2 - ||M_\theta(z^{t}_{\text{aligned}}, c, t) - M_\theta(z^{t}_{\text{protected}}, c, t)||^2_2$$

The image aligned with UNet contrastive loss is shown in the third column of the first row in Figure 9. This is still visually similar to the protected image because of the visual bound. After attackers fine-tune an SD model using these aligned images, they use it to generate images with a similar painting style. The third image of the second row shows the generated results. As expected, more semantic contents are restored, and we can see the trees, the river, and the statue in a clearer way. However, the textural “MIST” pattern has not been thoroughly removed. That is, the optimized result is a local optimum (i.e., an image with protective noises partially removed) instead of the global optimum. The root cause is that the guidance is only directional instead of pointy. As such, protective noises cannot be completely removed in training images and the leftover is still picked up during the training.

5.4 VAE Contrastive Alignment

The previous UNet alignment mainly covers the UNet component, while the textural pattern is usually injected via the VAE component and the resulting divergence will produce the watermark-like patterns [14, 49, 53]. More specifically, the textural protection effect is achieved by pulling the encoded feature of $x_{\text{protected}}$ to that of $x_{\text{target}}$ by minimizing the distance between them: $||\mathcal{E}(x_{\text{protected}}) - \mathcal{E}(x_{\text{target}})||^2_2$. To counteract this, we propose to use the contrastive alignment in the textural space (the VAE component) to pull the $\mathcal{E}(x_{\text{aligned}})$ in the same direction of $\mathcal{E}(x_{\text{visual}})$ and push it in the opposite direction of $\mathcal{E}(x_{\text{protected}})$. Similar to the UNet contrastive alignment, our goal is:

$$\min_\delta ||\mathcal{E}(x_{\text{aligned}}) - \mathcal{E}(x_{\text{visual}})||^2_2$$

$$\max_\delta ||\mathcal{E}(x_{\text{aligned}}) - \mathcal{E}(x_{\text{protected}})||^2_2$$
Fine-tune Stable Diffusion which can be combined into a loss:

\[
\mathcal{L}_{\text{reconst.}} = \| \mathcal{D}(\mathcal{E}(x_{\text{aligned}})) - x_{\text{aligned}} \|^2_2. \tag{12}
\]

The complete optimization objective of our framework is:

\[
\mathcal{L}_{\text{INSIGHT}} = \lambda_1 \cdot \mathcal{L}_{\text{bound}} + \lambda_2 \cdot \mathcal{L}_{\text{UNet}} + \lambda_3 \cdot \mathcal{L}_{\text{VAE}} + \lambda_4 \cdot \mathcal{L}_{\text{reconst.}},
\]

where \(\lambda\)’s adjust the strength of each loss element. By default, we set them as \(\lambda_{\text{bound}} = 0.1, \lambda_{\text{UNet}} = 0.1, \lambda_{\text{VAE}} = 0.1, \lambda_{\text{reconst.}} = 1\).

6 Evaluation

6.1 Experimental Setup

We implemented our optimization framework in PyTorch [22]. We evaluate our approach against 10 existing protection methods and reveal their vulnerability. Our experiments run on a server with Intel Xeon Silver 4214 2.20GHz 12-core CPUs with 256 GB RAM and 8 NVIDIA Quadro RTX 6000 GPUs.

Datasets and Models. We evaluate \(\mathcal{L}_{\text{INSIGHT}}\) on three different datasets WikiArt [31], CelebA-HQ [10], and Helen [11], following the literature [3, 14, 15, 36, 44, 49, 53]. The WikiArt dataset contains 42k+ artworks from 129 artists. Each artwork is categorized into a genre (e.g., Cubism, Impressionism). We filter out artists whose paintings can be not well classified (i.e., classification accuracy < 80%) by a CLIP genre classifier [24] as the existing work [3, 36]. We then randomly select 7 artists for evaluation on style mimicry. The CelebA-HQ dataset contains 30k celebrity images. We use the same strategy as Anti-DreamBooth [44] and randomly select 10 identities with 12 images for each individual for evaluation on subject recontextualization. The Helen dataset contains 2k+
images from Flickr. Each image comes with a mask annotating the face area. We randomly select 48 pairs of images and prompts for evaluation on image manipulation. We use the same Stable Diffusion models [26] with resolution 512 × 512 as existing protections and countermeasures.

**Generation Methods.** For style mimicry and subject recontextualization, we mainly use Dreambooth. For image misuse, we use mask-based inpainting.

**Protection Methods.** We evaluate our approach against Glaze, AdvDM±, Mist, ITA, SDS±, Anti-DreamBooth and SDST for style mimicry, Anti-DreamBooth for subject contextualization, PhotoGuard for image manipulation. For each protection, we use its official implementation.

**Baselines.** As we have introduced in Section 4, we compare our approach against four baselines: JPEG compression, Gaussian noise, Crop+resize, and Impress. For the first three baselines, we use the same settings as existing works [3, 14, 32, 36, 44, 53]. For Impress, we use the official implementation.

**Evaluation Metrics.** We use similar metrics as existing works [3, 14, 32, 36, 44, 53]. For style mimicry, we use Clip Accuracy [3, 24, 36] and Diffusion Accuracy [3, 12]. For the other two attacks, we use image quality metrics including Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [47], Feature Similarity Index Metric (FSIM) [50], Visual Information Fidelity (VIFp) [37], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [18], and LPIPS [51]. In our scenario, for all metrics except the last two, it’s the higher the better.

**Clip Accuracy.** A pre-trained CLIP model is used to classify the genre of a given painting. A clip model measures the text-image similarity. Given a painting, the CLIP model computes its similarity to a set of pre-defined genres and returns the most similar one.

**Diffusion Accuracy.** Given a painting and a potential genre, it uses the genre as the text condition and computes the reconstruction error for the painting with a Stable Diffusion model. From the set of pre-defined genres, it chooses the one with the smallest error as the label.

**Image Quality Metrics.** PSNR quantifies the ratio between the maximum possible signal power and the power of corrupting noise. SSIM calculates the structural similarity between two images. FSIM measures the similarity of two images. VIFp measures the quality of an image based on the premise of information fidelity. BRISQUE is a no-reference image quality measurement for individual images.

**Anonymous Human Study.** Because the audience of artworks is human, it’s necessary to conduct human studies to evaluate whether our approach outperforms baselines. Figure 12 shows a question sample for style mimicry. We provide 3We mainly report Clip Accuracy because it’s very time-consuming to compute Diffusion Accuracy. For one artist with one protection and one baseline, it takes about 30 minutes. Results for Diffusion Accuracy can be found in Appendix A.

Do you think (B) or (C) has the more similar painting styles (e.g., brushstrokes, color schemes, etc.) as the painting (A)?

![Figure 12: A human study example for style mimicry.](image)

For experiments on style mimicry, we fine-tune Stable Diffusion models separately on clean data, protected data, protected data transformed by JPEG compression, Gaussian noise, or Crop+resize, protected data purified by Impress or ours. We then use the models to generate images and calculate the Clip accuracy on each set.

**Overall Results.** Figure 13 shows the average Clip accuracy of different settings. The x-axis denotes different protection methods except that the first “Average” group shows the average result over all protection methods. The y-axis shows the Clip accuracy. The red dashed line means the average Clip accuracy in the clean setting is about 90%. The blue bars show the accuracy of protections. On average the protection can decrease the accuracy from 90% to 35% and Mist
exhibits the strongest effect on decreasing it to below 10%. Among the three transformation-based methods, Crop+resize indeed is very effective in diminishing the protection. It improves the average accuracy from 35% to 59% much more than JPEG and Gaussian. As for Impress, although its average performance is slightly lower than Crop+resize, it outperforms Crop+resize by a lot of margin on Glaze, AdvDM− and SDS−. Especially on Glaze, Impress almost doubles the accuracy of the Glaze-protected one, consistent with its reported performance [3]. When trained on images purified by our approach, the Clip accuracy of the generated images gets improved by about 138%. It almost completely removes the protection effect of Glaze, AdvDM−, and ITA. On Mist-protect images, the accuracy restored by our approach is about 7.3x of Impress’s and 2.5x of Crop+resize’s. All these results show our approach outperforms existing baselines. Figure 14 visualizes some results. Each row shows the generated images by Stable Diffusion models fine-tuned on different settings. Among all the countermeasures, ours leads to the best mimicry. More visualized results are in Appendix D.

Results of Human and GPT4-based Studies. Figure 15 shows a radar chart of our human study results. Each spoke means a protection method and the length denotes how many users prefer ours to the baseline (Crop+resize in light red and Impress in light purple). A length larger than 50% means they consider ours to be better. For example, on ITA-protected

---

### Figure 13: Clip Accuracy for images generated in different settings. The protections intend to reduce it while baselines and ours try to improve it. The red dashed line shows that the unprotected setting has an accuracy of 90%.

### Figure 14: Examples of existing countermeasures for the protection method SDST-. Each row shows synthetic samples generated by the Stable Diffusion model using the same prompt. Each column denotes a different setting.

### Figure 15: Human study results. Results of Human and GPT4-based Studies. Figure 15 shows a radar chart of our human study results. Each spoke means a protection method and the length denotes how many users prefer ours to the baseline (Crop+resize in light red and Impress in light purple). A length larger than 50% means they consider ours to be better. For example, on ITA-protected...
images, 100% users prefer ours to Impress and 84% prefer ours to Crop+resize. Figure 16 shows the preferences of the GPT4-Vision model. On average, it prefers ours to Impress in about 94.2% cases. Table 3 shows the pair-wise similarity results. A score $\geq 3$ means GPT4-Vision considers the two images to have a similar style. Ours achieves the highest average score of 3.5 while baselines’ are lower than 3.

**Results on Different Genres.** Figure 17 shows the results grouped by different genres averaged over different protections. The x-axis denotes four different genres: Post Impressionism, Art Nouveau Modern, Impressionism, and Northern Renaissance. The gray bar denotes the clean accuracy while bars in other colors mean the same settings as Figure 13. Among the four genres, we can see existing protections perform best for Post Impressionism. Crop+resize removes the protections more on Post Impressionism and Northern Renaissance than the other two genres. Impress has a better effect on Impressionism than other genres. Our approach outperforms baselines in all genres and is slightly more effective in Northern Renaissance than other genres.

### 6.3 Results on Subject Recontextualization

Table 4 list image quality metrics in subject recontextualization. For FSIM/SSIM/PSNR/VIFp/LPIPS, we compute the average pair-wise scores between a set of images generated with clean training data and those with protected or purified ones. A larger value means the generated images are similar to those in the unprotected setting. For BRISQUE, we compute the average score directly on the generated images in each setting. A smaller value means a better image quality. Our approach achieves the best scores except for VIFp (the third best) and LPIPS (the second best) where ours still outperforms the SOTA DM-based Impress. Visualized results can be found in our online material [21].

### 6.4 Results on Image Manipulation

Table 5 presents the quantitative results. We compute the image quality metrics FSIM/SSIM/PSNR/VIFp/LPIPS between the image manipulated from the clean one and that from the other setting (i.e., either protected or purified). A larger value means the image is more similar to its clean counterpart. Our approach achieves the best scores on FSIM/PSNR/VIFp/LPIPS and the second best on SSIM. This means the image edited from our aligned image is more similar to the clean version than the baselines. The effectiveness of simple transformations against PhotoGuard is consistent with existing papers [3, 32, 33]. We also report the BRISQUE. A smaller BRISQUE value means a better image quality. Thus, images generated using our purified data have better quality. Figure 18 visualizes some manipulated images.

### 6.5 Results on the Commercial Service

Scenario [34] is a commercial service on generative AI. Users can upload a set of paintings and build a diffusion model to generate new paintings in the same style (i.e., style mimicry). We upload Mist-protected images and the corresponding images purified by our methods to obtain two diffusion models.
We then generate new images using the two models. Figure 19 shows the results. The first row shows the images generated on the model trained on Mist-protected images are meaningless. The result in the second row demonstrates our approach can remove the protection.

6.6 Other Experiments

Because of the space limit, we put other experimental results in the appendix. Appendix C studies the adaptive protections where the defenders know our countermeasure. The effect of different fine-tuning strategies, the ablation study of our approach, our performance against GAN-based methods, and the overhead are presented in our online material [21].

7 Related Work

Diffusion Models for Image Generation

Diffusion models [7, 26, 41–43] have made notable advances in image synthesis, including unconditional image generation [20, 40], text-to-image synthesis [27, 30, 38], image-to-image translation and editing [13, 29, 46, 48, 52], image inpainting [2, 16] and editing [1, 17]. Many techniques have been proposed to personalize (i.e., fine-tune) LDMs to generate desired images [8, 19, 28].
Acknowledgments

We thank the anonymous reviewers for their constructive comments. We are grateful to the Center for AI Safety for providing computational resources. This research was supported, in part by IARPA TrojAI W911NF-19-S-0012, NSF 1901242 and 1910300, ONR N000141712045, N000141410468 and N000141712947. Any opinions, findings, and conclusions in this paper are those of the authors only and do not necessarily reflect the views of our sponsors.

References


Appendix

A Diffusion Accuracy

Because computing Diffusion accuracy is very time-consuming, here we only display the results of different countermeasures against Glaze in Figure 20 and show the comparison is similar to Clip Accuracy and ours is still the best.

B Setup for GPT4-based Study

Figure 21 shows an example of our GPT4-based study. The left part lists the instructions and three images we provided for the GPT4-Vision model. In this case, the first image is generated from a model trained on unprotected paintings while the second and third images are generated from a model trained on images purified by ours and Crop+resize respectively. The right part presents its output. It considers the second image (ours) to have the same style as the first image. It also provides a step-by-step analysis. We highlight some sentences we think are the key reasons for this decision. In particular, it correctly infers “The second painting’s style seems to share the Post-Impressionist traits evident in the first painting.” For the example of the pair-wise comparison, please refer to our online material [21].
I will provide you three painting images. You need to justify which of the second and third paintings has the more similar painting styles to the first painting. The styles can be brushstrokes, color schemes, etc.

The output choice is as follows:
1. The first and second paintings both exhibit visible, expressive brushstrokes that contribute to a textured and somewhat abstract appearance. Each stroke seems to add individual color and form within the painting, rather than blending smoothly.
2. The second painting has more similar painting styles to the first painting.
3. The third painting has more similar painting styles to the first painting.

Your response should be in the following format:

#thechoice: your choice here.
[note: provide your choice in based on your analysis.]
#thereason: your analysis here

C Adaptive Protection

In this section, we aim to demonstrate the robustness of our purification method against adaptive protection. We hypothesize that attackers attempting to bypass our purification must contend with several regularization terms imposed by our proposed losses.

Firstly, the perturbations introduced by attackers must be subtle enough that the perceptual similarity (as measured by LPIPS) between the protected and original images remains within a threshold. Secondly, images generated from the protected versions should exhibit a close resemblance to those produced using visual references, indicating that perturbations must preserve the efficacy within the visual references throughout the diffusion process. Thirdly, the latent features of the protected image align with those of the visual reference (i.e., the reference’s effectiveness mirrors that of the protected image), while simultaneously presenting sufficient distinction from the original image. Lastly, the protected images should be capable of reconstruction after the diffusion process.

Take Mist as an example, now the attacker’s optimization goal becomes:

$$\delta^* = \arg\min_\delta -\omega E_{\text{protected}} \cdot L_{DM}(x_{\text{protected}}, \theta)$$

$$+ \left\| E(x_{\text{target}}) - E(x + \delta) \right\|_2^2$$

$$+ \lambda_3 \cdot \max(x + \delta, x - \Delta, 0)$$

$$+ \lambda_4 \cdot \max(x + \delta, x - \Delta, 0)$$

$$+ \lambda_3 \cdot \max(x + \delta, x - \Delta, 0)$$

$$+ \lambda_4 \cdot \max(x + \delta, x - \Delta, 0)$$

$$+ \lambda_3 \cdot \max(x + \delta, x - \Delta, 0)$$

$$+ \lambda_4 \cdot \max(x + \delta, x - \Delta, 0)$$

Since it is challenging to create a visual reference, e.g., a photo, in every optimization step, we add Gaussian noise to the original image as the surrogate visual reference. Following Mist, we adopt the kernel size $1 \times 1$ and sigma $8 \times 8$.

In our experimental setup, we operate under the strong assumption that attackers have full knowledge of our hyper-parameters. Table 6 shows the results of the adaptive protection of Glaze on Vincent van Gogh and Raphael Kirchner’s works.

We also examine the adaptive protection under the single image manipulation setting using PhotoGuard. The results are shown in Table 7, suggesting the efficacy of INSIGHT against PhotoGuard’s adaptive protection.

We can therefore conclude INSIGHT remain effective in purifying adaptive protections.

D More Visualized Results

Figures 22 and 23 present more visualized results for different protections. More results are in our online material [21].
Figure 22: Examples of existing countermeasures for the protection method Mist. Each row shows synthetic samples generated by the Stable Diffusion model using the same prompt. Each column denotes a different setting.

Figure 23: Examples of existing countermeasures for the protection method Glaze. Each row shows synthetic samples generated by the Stable Diffusion model using the same prompt. Each column denotes a different setting.