

Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion?

Supplementary Material

A. Preliminary and Implementations

A.1. Fine-tuning Stable Diffusion

Text-to-Image. To directly fine-tune a Stable Diffusion model, users need to optimize the following loss function:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_{\theta}(z_t, t, \tau_{\theta}(\mathbf{y}))\|_2^2 \right] \quad (4)$$

Where z_t is the latent vector generated by the image in pixel space x_0 and an image encoder $\mathcal{E}(\cdot)$. \mathbf{y} is the text embedding and $\tau_{\theta}(\cdot)$ is the layers in Stable Diffusion which align the text embedding with the latent image vector.

LoRA. Low-Rank Adaptation (LoRA) [12] is a light fine-tuning method designed for large language models, which introduces rank decomposition matrices of Transformer layers to make the fine-tuning process more efficient, as shown in Eq. 5. W_0 is a pre-trained weight matrix and B and A are low rank decomposition matrices of ΔW .

$$h = W_0 \mathbf{x} + \Delta W \mathbf{x} = W_0 \mathbf{x} + B A \mathbf{x} \quad (5)$$

Ryu et al. introduce LoRA into Stable Diffusion for fast text-to-image diffusion fine-tuning², providing an efficient training and small size outputs for Stable Diffusion fine-tuning.

DreamBooth. DreamBooth [26] combines the reconstruction loss of diffusion training with a class-specific prior preservation loss to better avoid overfitting when fine-tuning Stable Diffusion with just several images.

$$\mathbb{E}_{x, c, \epsilon, \epsilon', t} [\omega_t \|\hat{\mathbf{x}}_{\theta}(\alpha_t \mathbf{x} + \sigma_t \epsilon, c)\|_2^2 + \lambda \omega_{t'} \|\hat{\mathbf{x}}_{\theta}(\alpha_{t'} \mathbf{x}_{pr} + \sigma_{t'} \epsilon', c_{pr})\|_2^2] \quad (6)$$

The second term of the training loss in Eq. 6 is the prior-preservation loss which supervises the Stable Diffusion with its class-specific generated images.

Custom Diffusion. Custom Diffusion [16] updates weights in Key and Value matrices of cross-attention layers while freezing other layers in the Stable Diffusion model, which are more influential during the text-to-image fine-tuning. Besides, Custom Diffusion also uses a regularization set of real images to prevent overfitting and use text

encoding to better inject the new concept. In our implementation, we only train the cross-attention layer and keep the weights of the text encoder during fine-tuning to highlight the features of Custom Diffusion.

Textual Inversion. Textual Inversion [8] finds a target token v_* to match the personal concept by directly optimizing the LDM object as shown in Eq. 7.

$$v_* = \arg \min_v \mathbb{E}_{z, \mathbf{y}, \epsilon, t} [\|\epsilon - \epsilon_{\theta}(z_t, t, c_{\theta}(\mathbf{y}))\|_2^2] \quad (7)$$

The advantage of Textual Inversion compared with other fine-tuning methods is that it only changes the text encoder of the Stable Diffusion model and keeps all parameters in the UNet while fine-tuning, which enables users to inject personal concepts with much smaller computational and spatial overhead.

A.2. Protective Perturbations

AdvDM. AdvDM [19] introduces a simple yet effective pipeline to add l_{∞} adversarial perturbations into images. The basic motivation of AdvDM is to make generative images be out-of-distribution examples, which leads to maximizing the following object in Eq. 8:

$$\max_{\|\delta_a\|_{\infty} < \rho} \mathbb{E}_{\epsilon, t} \|\epsilon - \epsilon_{\theta}(\mathcal{E}(\mathbf{x} + \delta_a), t)\|_2^2 \quad (8)$$

Where ρ is the l_{∞} bound of the perturbations. Its results show that images protected by AdvDM can be prevented from being used for style transfer and Stable Diffusion fine-tuning.

Anti-DreamBooth. Anti-DreamBooth [31] proposes another strong method to optimize the protective perturbation. Specifically, Anti-DreamBooth alternatively optimizes perturbations by maximizing the training loss of LDM and minimizing the training loss of DreamBooth to change the parameters of the model, as shown in Eq. 9:

$$\begin{aligned} \delta &= \arg \max_{\|\delta\|_p < \rho} \mathcal{L}_{\text{LDM}}(\theta, \mathbf{x}) \\ \theta &= \arg \min_{\theta} \mathcal{L}_{\text{DreamBooth}}(\theta, \mathbf{x} + \delta) \end{aligned} \quad (9)$$

Although it seems that Anti-DreamBooth is designed for the DreamBooth fine-tuning method, our experiments indicate that Anti-DreamBooth is also effective in other fine-tuning methods both in face and style learning.

² <https://github.com/cloneofsimon/lora>

Glaze. Glaze [28] focuses on the copyright concerns of style mimicry of text-to-image models. Different from the full-model attack in AdvDM and Anti-DreamBooth, Glaze designs a targeted optimization object toward the image feature-extracting process, which corresponds to the VAE in the Stable Diffusion model.

$$\min_{\delta} \|\Phi(\Omega(x, T)), \Phi(x + \delta)\|_2^2 + \alpha \cdot \max(\text{LPIPS}(\delta) - p, 0) \quad (10)$$

As shown in Eq.10, $\Phi(\cdot)$ is the image feature extractor, $\Omega(\cdot)$ refers to the style transfer, while T is the targeted style. Following this pipeline, Glaze aims to make Stable Diffusion learn the targeted style instead of the real style of training images during the fine-tuning process.

A.3. Expectation over Transformation

Expectation over Transformation (EoT) [1] is firstly proposed to synthesize physical adversarial examples in the real-world environment, which is an effective robust-ascending method towards tons of physical transformation such as cropping, rotation and color transformations. To evaluate the robustness of protective perturbation on Stable Diffusion models, we adopt the EoT methods to AdvDM (as shown in Eq. 2) to assess whether EoT helps to defend natural image transformations such as compression and blur. Specifically, we sample transformations of EoT including regular color transformations and Gaussian blur, which are usually used in the traditional EoT on classification tasks.

A.4. Adaptive Attack against DiffPure

To further evaluate the robustness of adversarial perturbations when facing the state-of-the-art adversarial purification method, DiffPure [23], we design an adaptive attack pipeline following recent diffusion-based attack methods for classification tasks [14, 38]. More specifically, we adopt the implementation of [38] which alternates the reverse sampling of SDEdit into DDIM to speed the back-propagation while also preventing memory overflow of GPUs, which is represented in Eq. 3. However, we find that

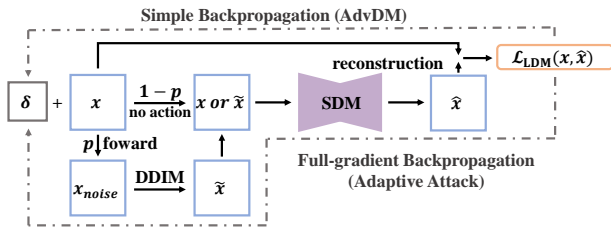


Figure 11. Framework of adaptive attack against DiffPure.

directly applying the full-gradient back-propagation to the optimization of perturbations will lead to failed protection even without purification, which may mainly be due to the Monte Carlo sampling of timestep t and noise ϵ . Thus, we

make a trade-off between the simple and the full-gradient back-propagation under the control of a probability p , as shown in Figure 11. By searching for the value of p , we find that setting p to 0.2 can get the best results in both with and without purification scenarios.

A.5. GrIDPure

Our GrIDPure is an iterative purification method as demonstrated in Figure 8 and Algorithm 1, which contains several GDP iterations (Figure 9 and Algorithm 2). There are two crucial parameters to control the purification pipeline, including the number of iterations M and the number of purification steps T of each iteration. Larger M and T can help to purify adversarial perturbations more completely but also lead to less similarity compared with the original clean image.

Algorithm 1 GrIDPure

Input: Perturbated Image x^a , Blend Weights γ , Purification steps T , Iterations M , the number of grids K

Output: Purified Image x^p

- 1: Initialize $x_0 \leftarrow x^a$
 - 2: **for** $m = 0 \rightarrow M - 1$ **do**
 - 3: $\tilde{x}_m \leftarrow \mathbf{GDP}(x_m, T, K)$
 - 4: $x_{m+1} \leftarrow (1 - \gamma) \cdot \tilde{x}_m + \gamma \cdot x_m$
 - 5: $x^p \leftarrow x_M$
 - 6: **return** x^p
-

Algorithm 2 Grid Diffusion-based Purification (GDP)

Input: Perturbated Image x , the number of grids K , Purification steps T

Output: Purified Image \tilde{x}

- 1: $x^0, x^1, \dots, x^{K-1} \leftarrow \mathbf{Crop}(x, K)$
 - 2: **for** $k = 0 \rightarrow K - 1$ **do**
 - 3: $x_n^k \leftarrow \text{diffusion}(x^k, T)$
 - 4: $\tilde{x}^k \leftarrow \text{denoise}(x_n^k, T)$
 - 5: $\tilde{x} \leftarrow \mathbf{Merge}(\tilde{x}^0, \dots, \tilde{x}^{K-1})$
 - 6: **return** \tilde{x}
-

B. Experiments Settings

B.1. Datasets and Metrics

We run the experiments on two main datasets: CelebA-HQ and WikiArt. CelebA-HQ is a high-quality dataset with a resolution of 1024×1024 that contains over 15 images for each attribute. WikiArt is an open-source painting dataset that contains artworks of different artists, we choose 6 to 10 images per artist to simulate usual practical fine-tuning. We resize the resolution of all these images to 512×512 to match the images with the base Stable Diffusion model

(*stable-diffusion-v1.5*³). To assess the generative quality, we generate 100 images for each concept and calculate two full-reference indexes, FID and precision score, and a non-reference quality metric, CLIP-Score. The FID and precision are based on the evaluation of *guided diffusion*⁴, while the CLIP-Score is based on the CLIP-IQA of the *piq* library⁵. In the experiment of purification quality which needs to compare the similarity between the purified image and the original clean image, we use SSIM and PSNR to evaluate the purification. Both of the indexes are based on the *piq* library.

B.2. Settings of Fine-tuning Methods

We choose LoRA as the default fine-tuning method in all the experiments, which is one of the most popular methods in the AIGC community. Besides, LoRA with training Text Encoder is also one of the most vulnerable fine-tuning methods in our results, which can help us further explore whether the protection is valid or not. For experiments in evaluating different fine-tuning methods (in Section 4), we apply all 8 different fine-tuning methods as shown in Table 2. For detailed parameters of fine-tuning, the learning rates of Text-to-Image, DreamBooth and Custom Diffusion are fixed at 3×10^{-5} and the training steps of these methods are fixed to 500. The learning rates of LoRA and Textual Inversion are fixed at 5×10^{-5} and 1×10^{-4} respectively. The training steps of LoRA and Textual Inversion are fixed at 300 and 3000 respectively. We make sure that with such fine-tuning settings, Stable Diffusion can successfully learn the concept from clean datasets. All these fine-tuning methods are based on the *diffusers* library⁶. The prompts used for fine-tuning are "a photo of a S^* person" and "a painting in the style of S^* " for CelebA-HQ and WikiArt datasets respectively.

B.3. Settings of Perturbations

All implementations of the protection methods are based on their official code and websites⁷⁸⁹. We maintain a fixed perturbation scale of 8/255 for ℓ_∞ noise (AdvDM, Anti-DreamBooth, and Improved-AdvDM). We set the optimizing rate of perturbations to 2/255 and the number of steps to 100 for AdvDM and the number of iterations to 10 for Anti-DreamBooth to ensure that the perturbations can provide enough protection. Additionally, we apply an adequate amount of Glaze perturbations to images following the recommended settings of its official application.

³ <https://huggingface.co/runwayml/stable-diffusion-v1-5>

⁴ <https://github.com/openai/guided-diffusion>

⁵ <https://github.com/photosynthesis-team/piq>

⁶ <https://github.com/huggingface/diffusers>

⁷ <https://github.com/VinAIRResearch/Anti-DreamBooth>

⁸ <https://github.com/CaradryanLiang/ImprovedAdvDM>

⁹ <https://glaze.cs.uchicago.edu/>

B.4. Settings of Natural Transformation

We apply two simple natural transformations to the protected image, including Gaussian blur and JPEG compression. For Gaussian blur, the kernel size is set to 7×7 and σ is set to 1.5. For JPEG compression, we use the implementation of *opencv2* library¹⁰ and the compression ratio is set to 40.

B.5. Settings of Purification

We follow the official code of DiffPure¹¹ with the off-the-shelf unconditional diffusion model trained on ImageNet to purify images and maintain most of the parameters but only change the number of purification steps to 50 or 100 (the total step of UDM is 1000). Experiments in Section 5.1 apply 100 steps DiffPure to ensure that the perturbations are successfully removed. For GrIDPure, we fix the number of iterations at 10 and the purification steps in each GDP iteration at 10 and γ at 0.1, which are sufficient to remove the protective perturbations.

C. Additional Results

In this section, we demonstrate more visualization results of Section 4 and Section 5.

C.1. Different Fine-tuning Methods

Results in Figure 15, Figure 16, Figure 17, Figure 18 and Figure 19 show more visualization of the effectiveness of different protective perturbations and different fine-tuning methods. The first, third, fifth and seventh lines are the results of without fine-tuning the text encoder and the other lines are the results of fine-tuning the text encoder. These indicate that the performance of protective perturbations is highly related to the chosen fine-tuning methods of image exploiters, especially the methods that train the text encoder.

C.2. Purification

Iterative DiffPure with Small Steps. The example in Figure 12 shows that protective perturbation can be removed by iterating a small-step DiffPure multiple times. Considering that a small-step purification changes less structure of the original clean images, we design our GrIDPure based on this insight.

Quality of Purification. Results in Figure 20, Figure 21, Figure 22 and Figure 23 compare the quality of purification between DiffPure and our GrIDPure. We set the purification steps of DiffPure to 100, and the purification steps and the number of iterations of GrIDPure to 10 and 10 respectively

¹⁰ <https://github.com/opencv/opencv>

¹¹ <https://github.com/NVlabs/DiffPure>

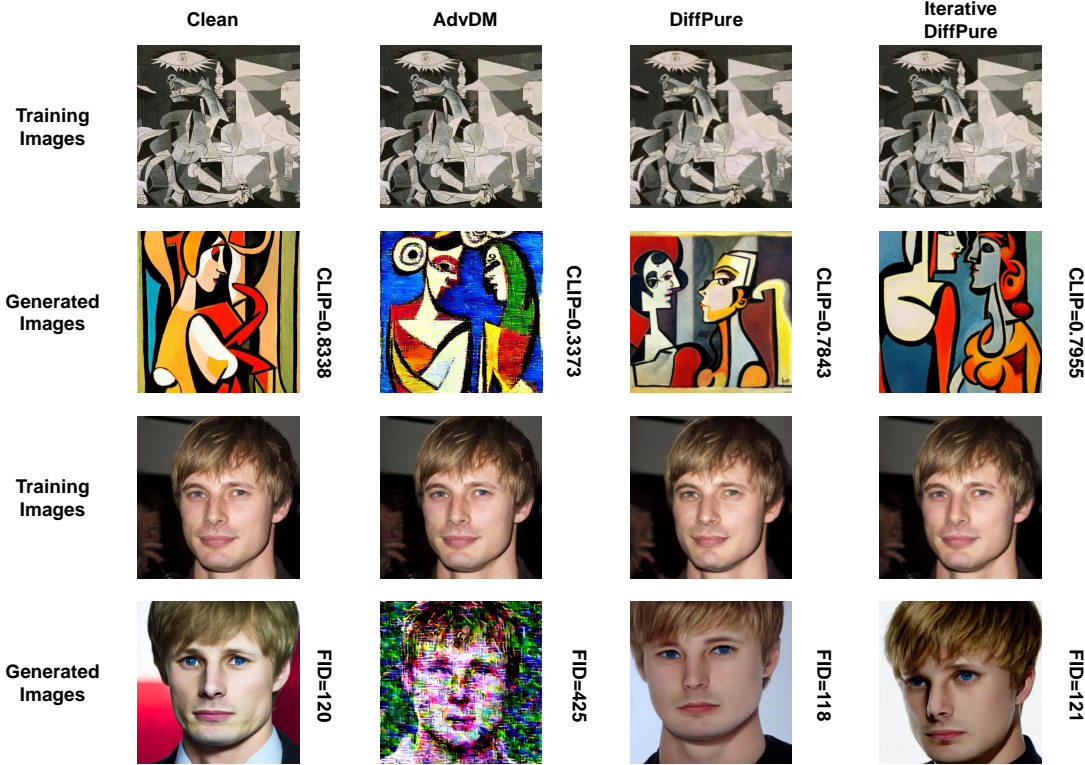


Figure 12. Iterative DiffPure with small steps can also successfully bypass the protective perturbations. The 100-step DiffPure is broken down into 10 iterations of 10-step DiffPure.

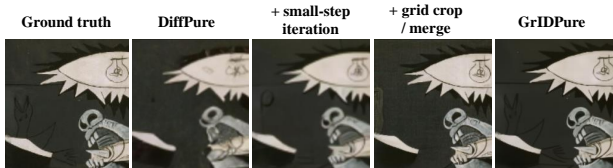


Figure 13. Visualization examples of DiffPure, small-step iteration, grid crop/merge and GrIDPure.

in Figure 20, Figure 21 and Figure 22. We set the purification steps of DiffPure to 200, and the purification steps and the number of iterations of GrIDPure to 10 and 20 respectively in Figure 23 to ensure the complete purification.

Effectiveness of Purification. Results in Figure 24, Figure 25, Figure 26, Figure 27, Figure 28 and Figure 29 demonstrate that our GrIDPure can successfully bypass all SOTA protective perturbations which remove the perturbation on the training images and recover these images into learnable images. The scales of perturbations are set to 8/255 for AdvDM, Anti-DreamBooth and ImprovedAdvDM, and 16/255 for AdvDM16. For Glaze, we apply the strongest settings provided by its official application.

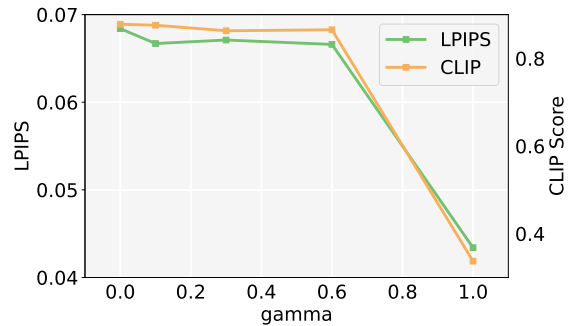


Figure 14. Ablation study on blending parameter γ .

C.3. Ablation Study

Adaptive Attack against GrIDPure Considering the theoretical consistency between GrIDPure and DiffPure and the significant GPU memory requirements, it's almost infeasible to conduct Adaptive Attacks on the GrIDPure framework. We make our best effort to design a white-box adaptive attack specifically targeting the core component of GrIDPure, *Small-step Iteration*. Results in Table 7 indicate that such adaptive attacks do not diminish the effectiveness of GrIDPure.

Training Method	Metrics	Clean	AdvDM	AdvDM +DiffPure	AdvDM +GrIDPure	AntiDB	AntiDB +DiffPure	AntiDB +GrIDPure
LORA w\te	FDFR	0.04	1.0	0.06	0.04	0.96	0.04	0.0
	ISM	0.59	0.0	0.55	0.56	0.03	0.56	0.56
LORA w\o te	FDFR	0.07	0.26	0.16	0.10	0.18	0.16	0.06
	ISM	0.58	0.49	0.53	0.60	0.52	0.57	0.61

Table 6. More metrics on face-generation tasks.

Dataset	Metrcs	Clean	AdvDM	Ada.	AdvDM +GrIDPure	Ada. +GrIDPure
CelebA	FID	119.8	424.7	253.1	121.4	114.8
	CLIP	0.7378	0.2316	0.5473	0.8526	0.7406
WikiArt	FID	201.9	251.1	240.8	203.4	206.6
	CLIP	0.8338	0.3373	0.5124	0.8758	0.8415

Table 7. Adaptive attack against GrIDPure.

Assessment Metrics	Quality			Effectiveness	
	PSNR	SSIM	LPIPS	FID	CLIP
<i>AdvDM (No Pure.)</i>	<i>37.46</i>	<i>0.9496</i>	<i>0.0434</i>	<i>251.1</i>	<i>0.3373</i>
DiffPure	22.24	0.6378	0.4425	214.2	0.7843
+ small-step iter.	23.42	0.7175	0.3904	211.2	0.7955
+ grid crop/ merge	27.14	0.8052	0.0757	214.7	0.7577
GrIDPure	30.60	0.9199	0.0672	203.4	0.8758

Table 8. Ablation study on small-step iteration and grid crop/merge.

Ablation Studies on GrIDPure We do ablation studies for mechanisms (shown in Figure 13 and Table 8) and blending parameter γ (shown in Figure 14) in GrIDPure. Small-step Iteration aids in better preserving the details of the images, while grid crop/merge helps in retaining the resolution of the images. As shown in Figure 14, by appropriately blending images from different iterations, we can mitigate the loss of details during the SDEdit processes, striking a balance between preserving image details (smaller LPIPS [42]) and removing protective perturbation (higher CLIP-Score).

C.4. Additional Metrics

To further demonstrate the influence of protective perturbations on the face-generation task, we refer to AntiDreamBooth [31] to calculate FDFR and ISM for the face generation in Table 6, where the lower FDFR and higher ISM represent better generative quality. The results from these metrics align with the conclusions that fine-tuning the text encoder greatly enhances the protection efficacy diffusion-based purification can successfully remove these protections.

D. Broader Impact

This paper evaluates methods that use protective perturbations to prevent generative models from exploiting personal

data, thereby addressing concerns such as privacy breaches and copyright infringement. Additionally, the paper proposes approaches to bypass these protections, potentially exposing protected data to risks and providing opportunities for unauthorized exploiters to bypass existing protective measures. Despite these challenges, we believe that assessing the effectiveness of such protections is crucial. In the long run, our work holds positive implications for safeguarding personal privacy and copyright in images.

Acknowledgements

This work is partially supported by the NSF of China (under Grants U22A2028, 61925208, 62222214, 62102399, 62102398, U20A20227, 62372436, 62302478, 62302482, 62302483, 62302480), CAS Project for Young Scientists in Basic Research (YSBR-029), Youth Innovation Promotion Association CAS and Xplore Prize.

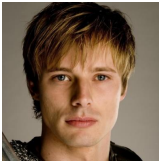


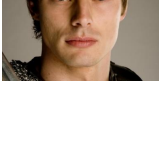





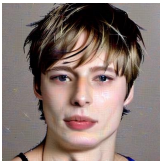

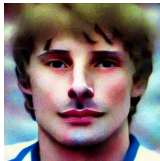



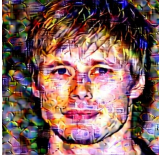
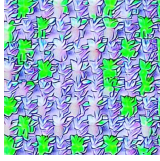
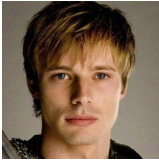




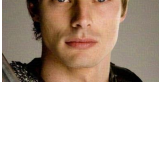


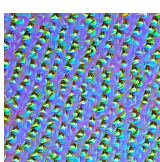
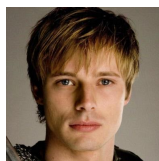



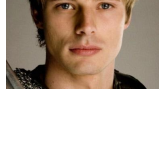
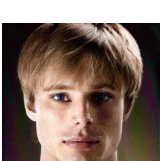
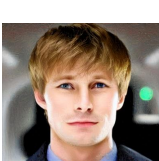

An Example of Training images		Generated Images			
		Text-to-Image w/o te w\ te	LoRA w/o te w\ te	DreamBooth w/o te w\ te	Custom Diffusion Textual Inversion
Clean					
					
AdvDM					
					
Anti-DB					
					
IAdvDM					
					

Figure 15. Generated images of Stable Diffusion training with different fine-tuning methods and different protective datasets. The prompt of generating is "a photo of a sks person".






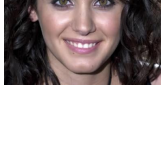
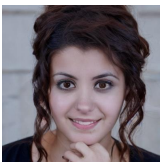
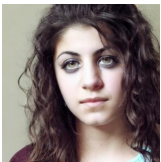
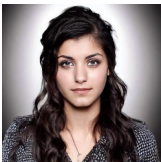
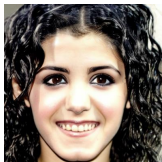




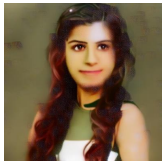
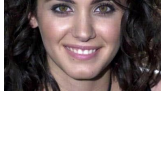



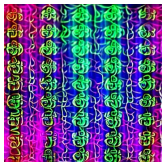




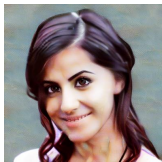
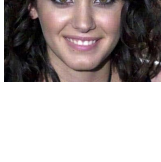



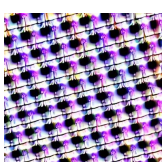

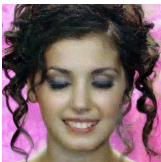

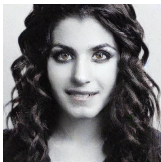

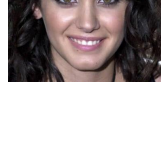



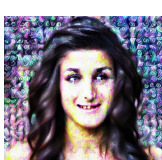
An Example of Training images		Generated Images			
		Text-to-Image w/o te w\ te	LoRA w/o te w\ te	DreamBooth w/o te w\ te	Custom Diffusion Textual Inversion
Clean					
					
AdvDM					
					
Anti-DB					
					
IAdvDM					
					

Figure 16. Generated images of Stable Diffusion training with different fine-tuning methods and different protective datasets. The prompt of generating is "a photo of a sks person".

	An Example of Training images	Generated Images			
		Text-to-Image w/o te w\ te	LoRA w/o te w\ te	DreamBooth w/o te w\ te	Custom Diffusion Textual Inversion
Clean					
					
AdvDM					
					
Anti-DB					
					
IAdvDM					
					

Figure 17. Generated images of Stable Diffusion training with different fine-tuning methods and different protective datasets. The prompt of generating is "a photo of a sks person".










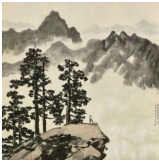

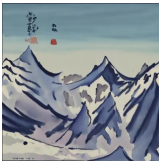

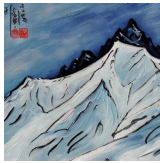
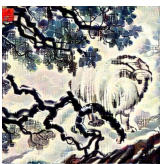
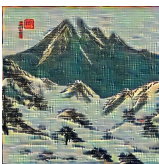







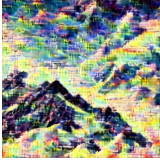
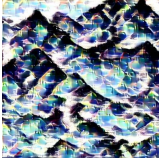
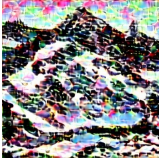









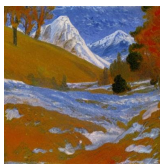
	An Example of Training images	Generated Images			
		Text-to-Image w/o te w\ te	LoRA w/o te w\ te	DreamBooth w/o te w\ te	Custom Diffusion Textual Inversion
Clean					
					
AdvDM					
					
Anti-DB					
					
Glaze					
					

Figure 18. Generated images of Stable Diffusion training with different fine-tuning methods and different protective datasets. The prompt of generating is "a painting of snow mountain in the style of Xu Bei-hong".

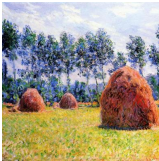






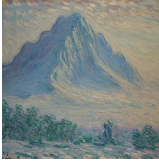


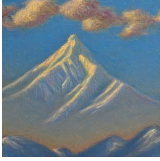




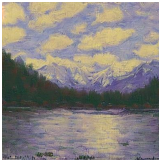

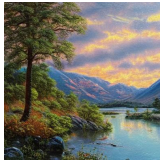
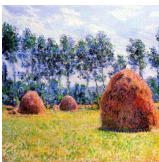





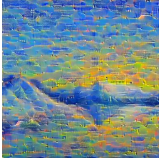
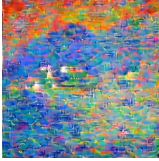



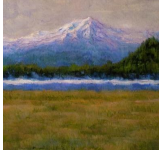


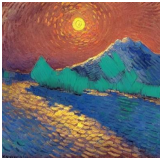



	An Example of Training images	Generated Images			
		Text-to-Image w/o te w\ te	LoRA w/o te w\ te	DreamBooth w/o te w\ te	Custom Diffusion Textual Inversion
Clean					
					
AdvDM					
					
Anti-DB					
					
Glaze					
					

Figure 19. Generated images of Stable Diffusion training with different fine-tuning methods and different protective datasets. The prompt of generating is "a painting of snow mountain in the style of Monet".

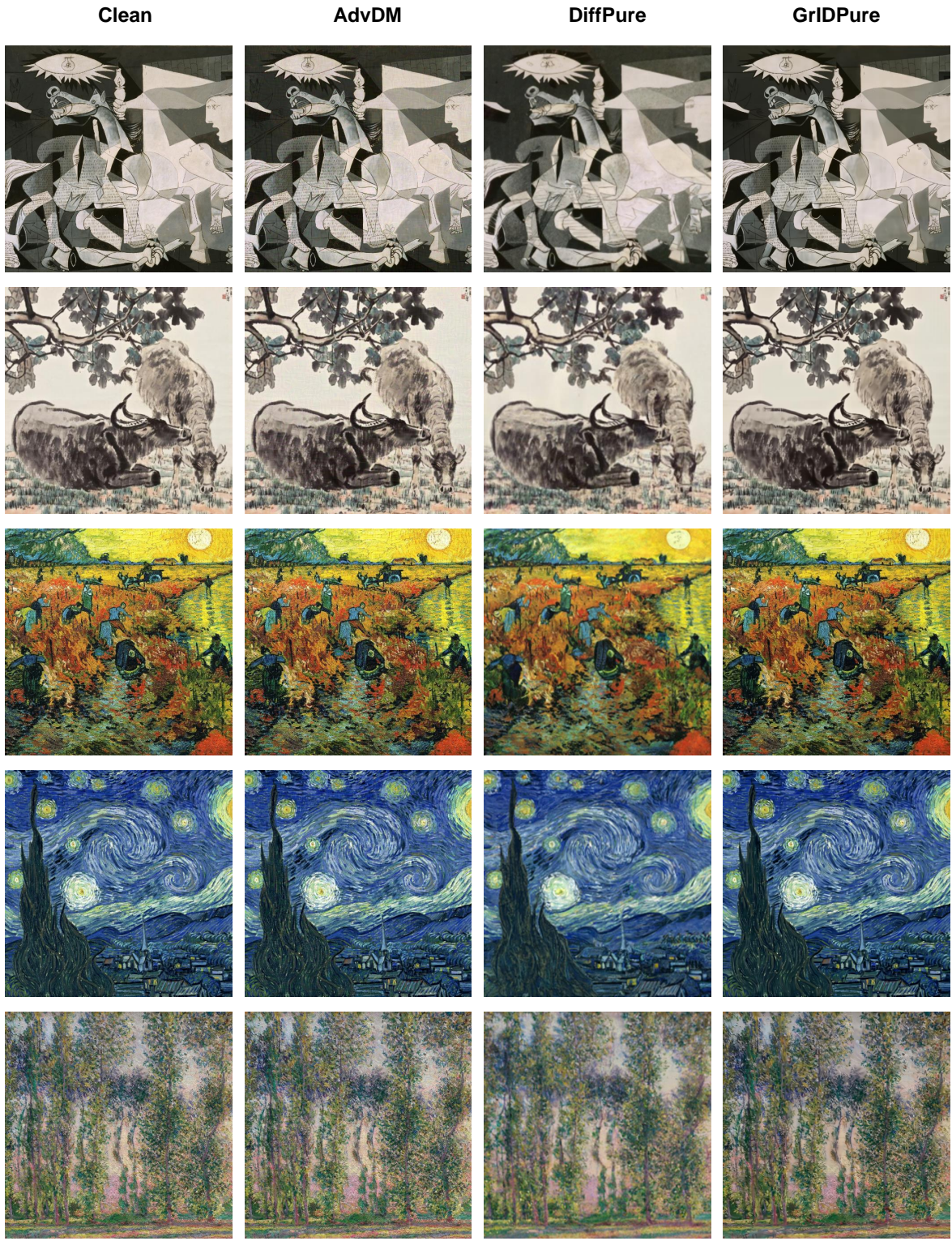


Figure 20. Visualization of clean training images, AdvDM-protected images and images purified by DiffPure and GrIDPure. Our GrIDPure can better preserve the quality (resolution and structure) of the original clean image.

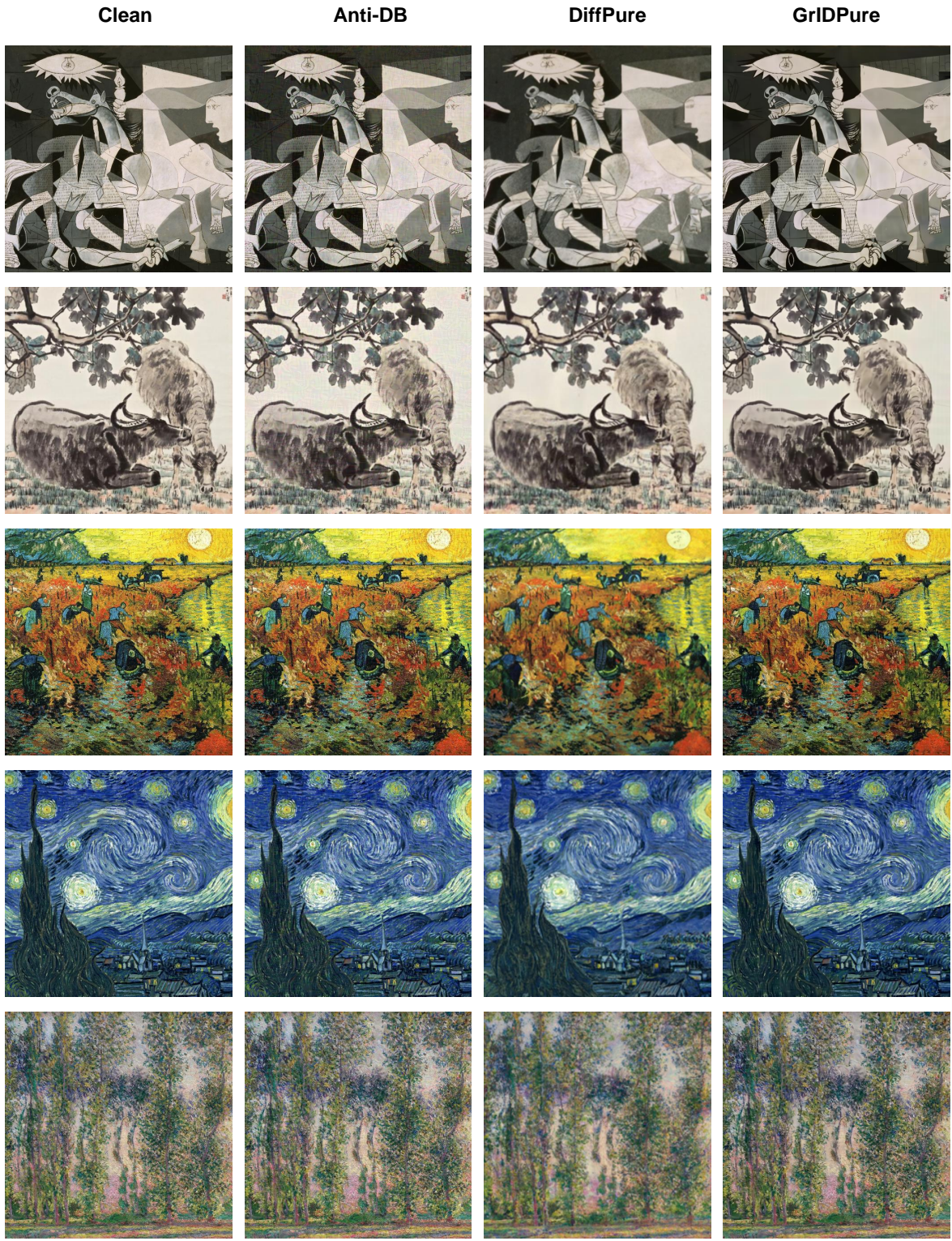


Figure 21. Visualization of clean training images, AntiDB-protected images and images purified by DiffPure and GrIDPure. Our GrIDPure can better preserve the quality (resolution and structure) of the original clean image.

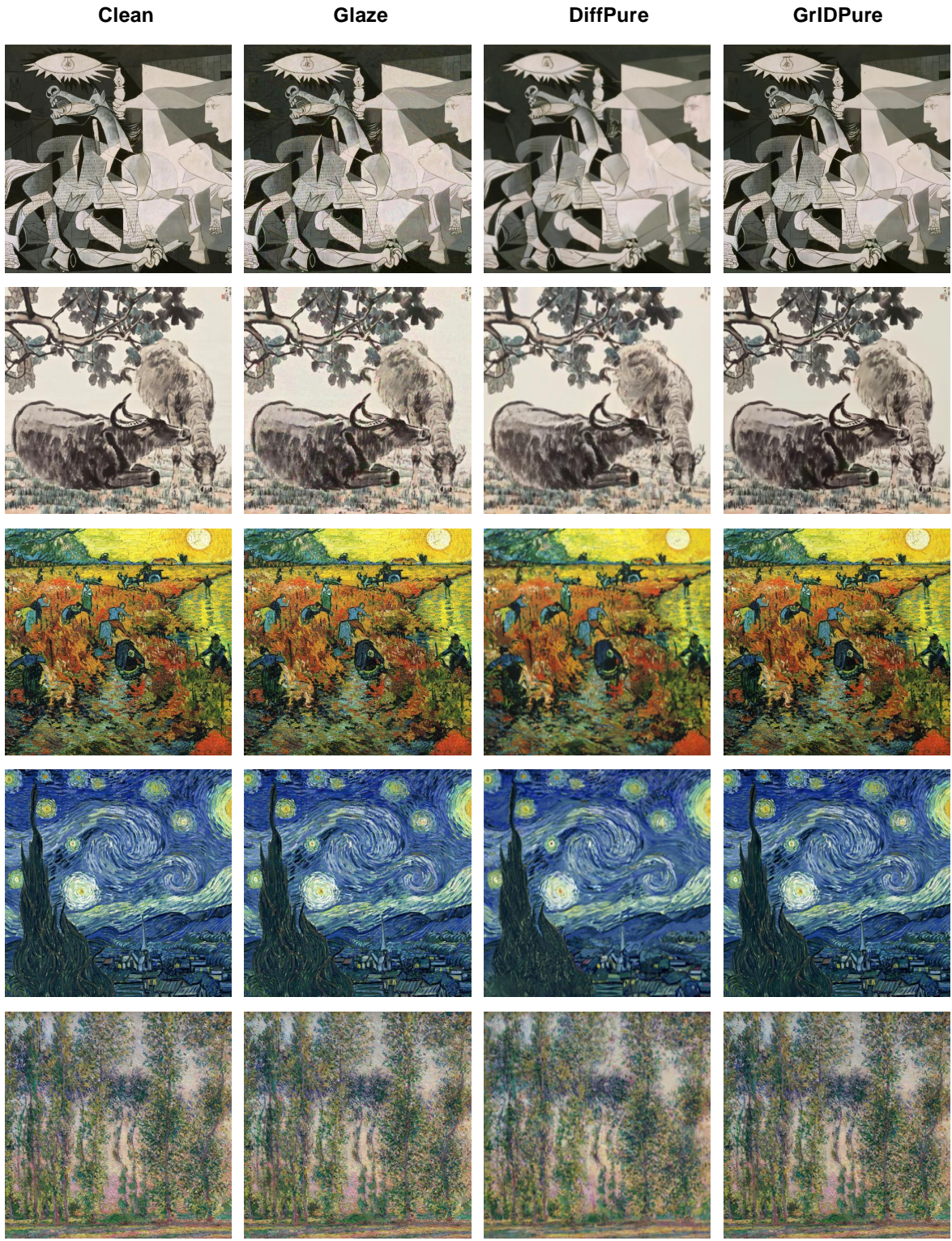


Figure 22. Visualization of clean training images, Glaze-protected images and images purified by DiffPure and GrIDPure. Our GrIDPure can better preserve the quality (resolution and structure) of the original clean image.

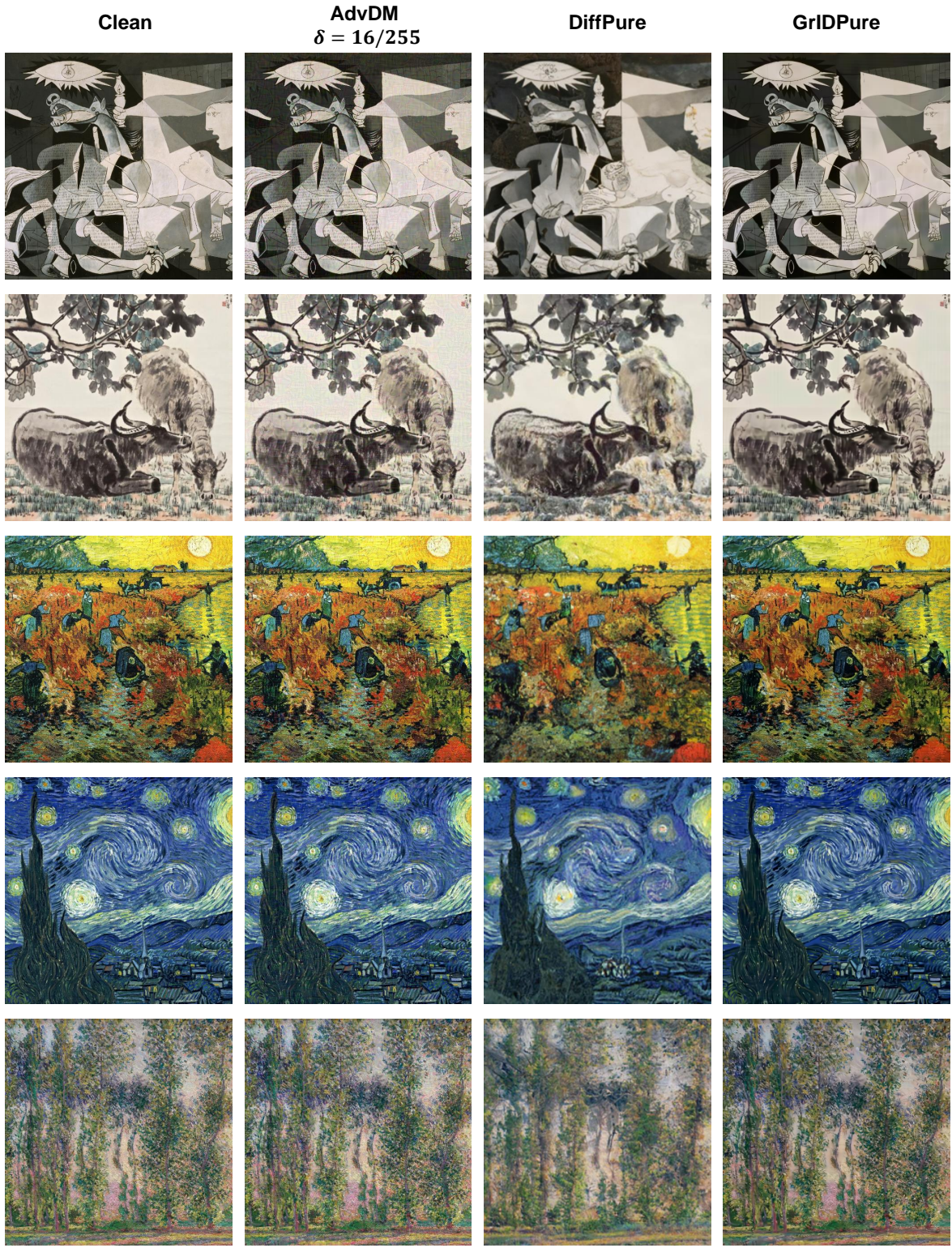


Figure 23. Visualization of clean training images, AdvDM-protected ($\delta = 16/255$) images and images purified by DiffPure and GrIDPure. Our GrIDPure can better preserve the quality (resolution and structure) of the original clean image.

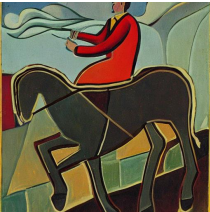

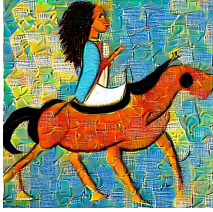
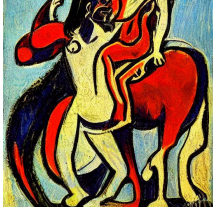


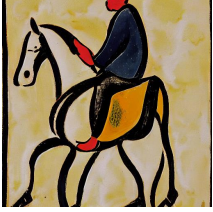

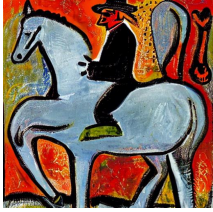

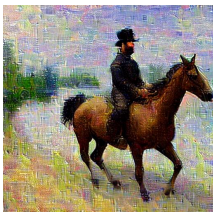
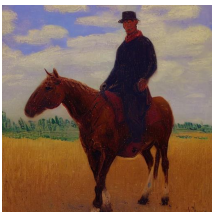
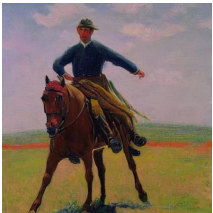

Clean Generation	Protected Generation	GrIDPure Generation
 <p data-bbox="256 781 412 890">"a painting of a man riding a horse in the style of Picasso"</p>	<p data-bbox="477 386 500 491">AdvDM16</p>  <p data-bbox="477 596 500 701">Anti-DB</p>  <p data-bbox="477 827 500 932">Glaze</p> 	     
 <p data-bbox="256 1482 412 1591">"a painting of a man riding a horse in the style of Monet"</p>	<p data-bbox="477 1083 500 1188">AdvDM16</p>  <p data-bbox="477 1293 500 1398">Anti-DB</p>  <p data-bbox="477 1524 500 1629">Glaze</p> 	     

Figure 24. Visualization of generated images by Stable Diffusion fine-tuning with images purified by GrIDPure.

Clean Generation	Protected Generation	GrIDPure Generation	
 <p data-bbox="267 783 418 888">"a painting of a man riding a horse in the style of Xu Beihong"</p>	<p data-bbox="479 384 506 489">AdvDM16</p>  <p data-bbox="479 594 506 699">Anti-DB</p>  <p data-bbox="479 825 506 930">Glaze</p> 	  	  
 <p data-bbox="267 1486 418 1591">"a painting of a panda in the style of Picasso"</p>	<p data-bbox="479 1087 506 1192">AdvDM16</p>  <p data-bbox="479 1297 506 1402">Anti-DB</p>  <p data-bbox="479 1507 506 1612">Glaze</p> 	  	  

Figure 25. Visualization of generated images by Stable Diffusion fine-tuning with images purified by GrIDPure.





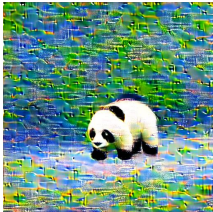

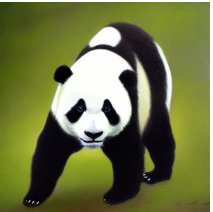







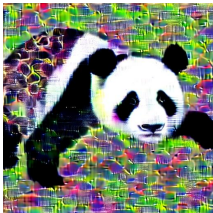




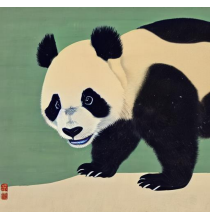
Clean Generation		Protected Generation	GrIDPure Generation	
 <p><i>"a painting of a panda in the style of Monet"</i></p>	AdvDM16			
	Anti-DB			
	Glaze			
 <p><i>"a painting of a panda in the style of Xu Beihong"</i></p>	AdvDM16			
	Anti-DB			
	Glaze			

Figure 26. Visualization of generated images by Stable Diffusion fine-tuning with images purified by GrIDPure.

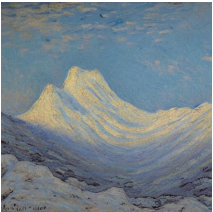
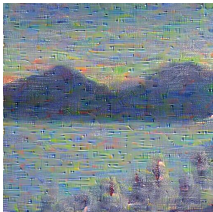
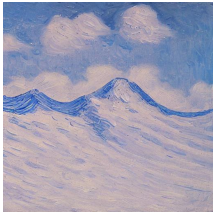




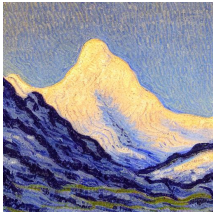
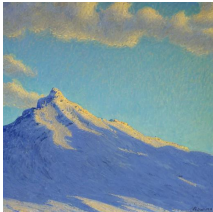


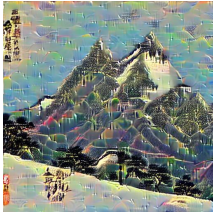
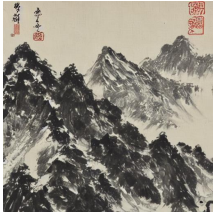

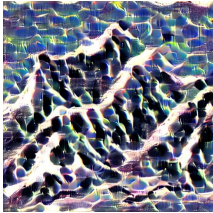
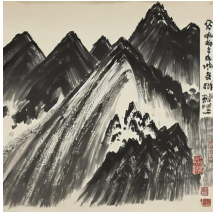

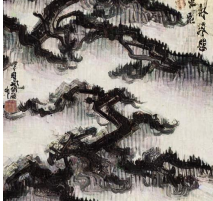


Clean Generation		Protected Generation	GrIDPure Generation	
 <p><i>“a painting of a snow mountain in the style of Monet”</i></p>	AdvDM16			
	Anti-DB			
	Glaze			
 <p><i>“a painting of a snow mountain in the style of Xu Beihong”</i></p>	AdvDM16			
	Anti-DB			
	Glaze			

Figure 27. Visualization of generated images by Stable Diffusion fine-tuning with images purified by GrIDPure.








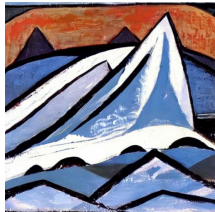












Clean Generation	Protected Generation	GrIDPure Generation
 <p data-bbox="256 783 412 888">"a painting of a snow mountain in the style of Picasso"</p>	<p data-bbox="477 384 505 489">AdvDM16</p>  <p data-bbox="477 604 505 709">Anti-DB</p>  <p data-bbox="477 842 505 905">Glaze</p> 	     
 <p data-bbox="256 1486 412 1539">"a photo of a sks person"</p>	<p data-bbox="477 1098 505 1182">AdvDM</p>  <p data-bbox="477 1308 505 1392">Anti-DB</p>  <p data-bbox="477 1560 505 1602">IAdvDM</p> 	     

Figure 28. Visualization of generated images by Stable Diffusion fine-tuning with images purified by GrIDPure.





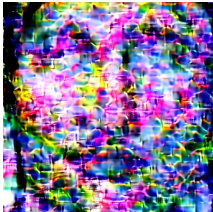









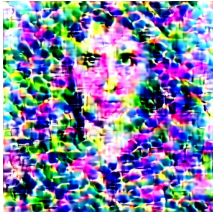





Clean Generation		Protected Generation	GrIDPure Generation	
 <p><i>"a photo of a sks person"</i></p>	AdvDM			
	Anti-DB			
	IAdvDM			
 <p><i>"a photo of a sks person"</i></p>	AdvDM			
	Anti-DB			
	IAdvDM			

Figure 29. Visualization of generated images by Stable Diffusion fine-tuning with images purified by GrIDPure.