

Supplementary for DCT-Shield: A Robust Frequency Domain Defense against Malicious Image Editing

A. DCT-Shield Algorithm

Algorithm 1: Frequency Domain Noise Optimization for Image Immunization

Input: Input image \mathbf{x} , VAE encoder \mathcal{E} , JPEG quality Q_{alg} , coefficient perturbation budget ϵ , list of channels to perturb $\mathcal{C} \in \{Y, Cb, Cr\}$, step size γ , number of optimization steps N , mask M

Output: Immunized image \mathbf{x}'

```

1 Initialize  $\delta \leftarrow 0$ 
  /*  $\alpha$  has 3 components Y, Cb, Cr */
2  $\alpha \leftarrow JPEG_E(\mathbf{x}; Q_{alg})$ 
3 if  $M \neq \emptyset$  then
4    $M \leftarrow component\_wise\_masks(M)$ 
5 for  $i = 0$  to  $N - 1$  do
6    $\eta \leftarrow (1 - i/N)\gamma$ 
7   for  $c \in \mathcal{C}$  do
8      $\alpha'_c \leftarrow \alpha_c + \delta_c$ 
9    $\mathbf{x}' \leftarrow JPEG_D(\alpha'; Q_{alg})$ 
10   $\mathcal{L}(\delta) \leftarrow \|\mathcal{E}(\mathbf{x}')\|_2$ 
11   $\delta \leftarrow \delta - \text{sign}(\nabla \mathcal{L}) \cdot \eta$ 
12   $\delta \leftarrow \text{clamp}(\delta, -\epsilon, +\epsilon)$ 
13  if  $M \neq \emptyset$  then
14     $\delta_c \leftarrow \delta_c \odot \mathbf{M}_c$ 

```

DCT-Shield defends against adversarial attacks by introducing subtle, imperceptible noise in the frequency domain, modifying the Discrete Cosine Transform (DCT) coefficients of an image. Our approach leverages the JPEG pipeline to add quantization-aware noise, ensuring strong protection even against JPEG-based purification methods. Algorithm 1 outlines the complete approach.

B. Dataset

For the instruction-based editing task, we use the OmniEdit test set, which consists of 700 images. From this set, we select 150 high-quality images where the editing model, In-

structPix2Pix, produces reliable outputs. Our algorithm is then applied to these selected images to generate protected versions. We subsequently perform edits on the protected images and compare the results with the edits of unprotected (clean) images to assess the impact of our approach.

Additionally, we evaluate our method on the inpainting task. To support this, we create a dataset of 56 images, combining samples from the PPR10k dataset with images collected from the web. PPR10k is a high-quality portrait dataset that provides human-region masks. The web-crawled images includes a diverse mix of popular celebrity photos and pet photos. For these, we generate segmentation masks using SAM [4]. Each image in the inpainting dataset is manually annotated with three suitable edit prompts. Examples of images and their corresponding masks are shown in Figure 1.

C. Evaluation Details

To assess the effectiveness of our image protection method, we employ a combination of quantitative metrics and human evaluations. The quantitative metrics evaluate image fidelity and robustness, while human evaluations provide qualitative insights into the effectiveness of the protection against malicious modifications. Below, we detail the evaluation methodologies used in our study.

C.1. Quantitative Metrics

C.1.1. Frechet Inception Distance (FID)

FID [2] measures the distributional difference between real and generated images based on activations from a pre-trained Inception network. We utilize FID to measure the distributional difference between the clean edits and protected edits. It is defined as

$$FID = \|\mu_r - \mu_g\|_2^2 + \text{Tr} \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right), \quad (1)$$

where μ_r , Σ_r and μ_g , Σ_g are the mean and covariance of the real and generated feature distributions.



Figure 1. Images and masks from the curated inpainting dataset.

C.1.2. Structural Similarity Index Measure (SSIM)

SSIM [7] evaluates the perceptual similarity between two images x and y based on luminance, contrast, and structure:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

where, μ_x, μ_y are the mean intensities of images x and y and σ_x^2, σ_y^2 are their variances. C_1 and C_2 are stabilizing constants.

C.1.3. PSNR

Peak Signal-to-Noise Ratio (PSNR) is a standard metric for evaluating the similarity between two images. It quantifies the ratio between the maximum possible signal power and the power of the noise that affects image quality. In our context, PSNR is used to compare the edited versions of the

unprotected image (denoted as \tilde{I}_u) and the protected image (denoted as \tilde{I}_p).

The PSNR is defined as:

$$\text{PSNR}(\tilde{I}_u, \tilde{I}_p) = 10 \cdot \log_{10} \left(\frac{L^2}{\text{MSE}(\tilde{I}_u, \tilde{I}_p)} \right) \quad (3)$$

where L is the maximum possible pixel value (e.g., 255 for 8-bit images), and the mean squared error (MSE) is computed as:

$$\text{MSE}(\tilde{I}_u, \tilde{I}_p) = \frac{1}{N} \sum_{i=1}^N (\tilde{I}_{u,i} - \tilde{I}_{p,i})^2 \quad (4)$$

Here, N is the total number of pixels in the image.

C.1.4. Visual Information Fidelity (VIF)

VIF [6] measures the amount of visual information retained in a distorted image relative to the original:

Human Eval on Image Protection Against Malicious Editing

The survey consists of two types of questions:


Image Fidelity Assessment: You will be shown images protected by different methods, including ours. Please rate them on a scale of 1 to 5 based on how well they maintain visual similarity to the original image. *Left to Right order - Input Image, Method 1, Method 2, ..., Method 5.*

Protection Effectiveness: You will compare images that have undergone malicious edits. Your task is to evaluate how well each protection method prevents or disrupts unauthorized modifications. **If the edited image doesn't resemble the original image or it doesn't fulfill the edit instruction, then the method of protection is good and vice-versa. Left to Right order - Edited Image from Input Image, Edited image from Method 1 protected image, Edited Image from Method 2 protected image, ..., Edited Image from Method 5 protected image.**

Note: Rating 5 is best and Rating 1 is worst. Give rating to all the methods.

Your feedback will help us refine our approach and improve image protection techniques. Thank you for your time and valuable input!

Rate on the basis of image fidelity retained wrt original image. *



	1	2	3	4	5
Method 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Method 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Human Eval Survey for Inpainting

Thank you for participating in this study! This survey is part of our research on protecting images from unauthorized edits by generative models. Our method aims to immunize images, making them resistant to malicious modifications.


In this survey, you will see an **immune image** along with an **editing instruction**. You will then be presented with inpainted images generated using our method and four other baseline methods. Your task is to **rate each method on a scale of 1 to 5 based on how well it prevented the requested edit**.

- If an edited image successfully follows the instruction, the corresponding method failed to immunize the source image and hence give that method a lower rating.
- If the edit is blocked or disrupted, the method was successful in preventing malicious modification and hence give that method a higher rating

Note: Left to Right Order - Input Image, Edited Image from Immunized Image of Method 1, Edited Image from Immunized Image of Method 2, ..., Edited Image from Immunized Image of Method 6.

Your feedback will help us evaluate and improve image immunization techniques. Thank you for your time and insights!

Edit Inst - a girl in a military camp



	1	2	3	4	5
Method 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Method 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2. Snapshots of Google forms used to collect human feedback for editing and inpainting tasks.

$$\text{VIF} = \frac{\sum_i I(C_i; F_i | S_i)}{\sum_i I(C_i; E_i | S_i)} \quad (5)$$

where, $I(C_i; F_i | S_i)$ is the mutual information between the original image and its features, $I(C_i; E_i | S_i)$ is the mutual information between the distorted image and its features.

C.1.5. Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS [8] evaluates perceptual similarity by comparing deep feature representations:

$$\text{LPIPS}(x, y) = \sum_l w_l \|f_l(x) - f_l(y)\|_2^2 \quad (6)$$

where, $f_l(x)$ and $f_l(y)$ are deep feature representations at layer l , w_l are learned weights for each layer.

We use the open-source library `pyiqa`¹ for computing the above mentioned image quality metrics.

C.1.6. CLIP Score for Text-to-Image Similarity

CLIP (Contrastive Language-Image Pre-training) [5] is a model that jointly trains vision and text encoders to create aligned representations of images and text. To evaluate how well an edit follows a given instruction, we measure the cosine similarity between the text representation of the target description, $\text{CLIP}_{\text{text}}(I_{\text{edit}})$, and the image representation

¹<https://github.com/chaofengc/IQA-PyTorch>

of the edited output, $\text{CLIP}_{\text{image}}(x_{\text{edit}})$, where x_{edit} is the inpainted image. A higher similarity score suggests that the edit better matches the intended changes. This metric helps assess how faithfully the edits reflect the given instructions.

C.2. Human Evaluation

To supplement our quantitative analysis, we conduct two human evaluation surveys to assess the effectiveness of our image protection method against malicious editing in two different tasks. We create Google Forms (Figure 2) and distribute them to 50 participants, receiving responses from 42. Each participant reviews and provides feedback on 30 samples per task. The specific evaluation criteria for comparing different methods are detailed below.

Comparison of Protected Image Fidelity and Editing Robustness. In this evaluation, participants compare the protected images generated by our method and various baseline methods. They rate each method on two aspects:

- **Fidelity to the Original Image:** Participants score the protected image’s similarity to the original image on a scale of 1 to 5, where 5 indicates high similarity and 1 indicates significant distortion.
- **Robustness to Malicious Editing:** Given an edit instruction, participants rate how well the protected image prevents the edit. A score of 5 means the edit instruction is not followed (indicating strong protection), whereas a score of 1 means the edit is well-executed (indicating weak protection).

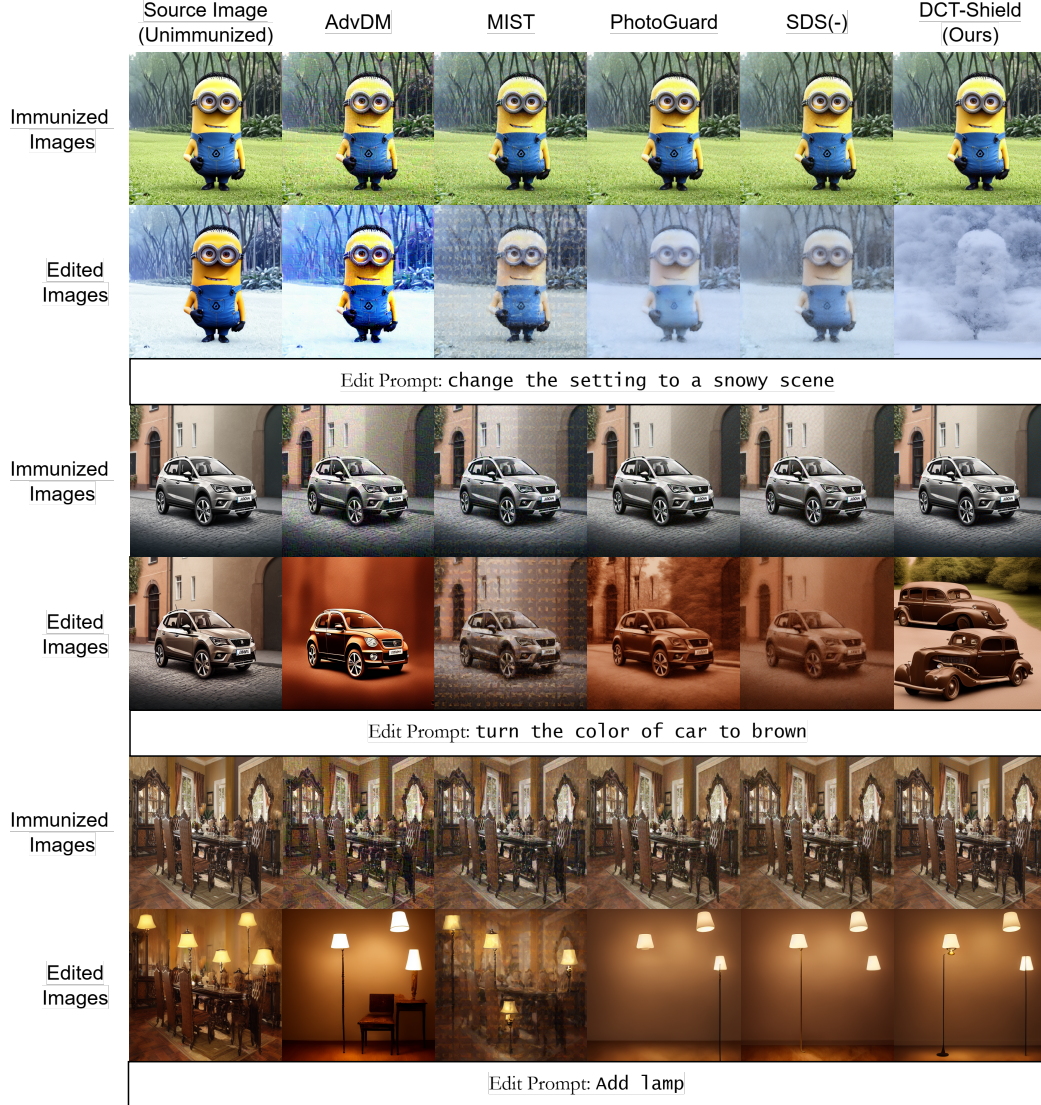


Figure 3. More Results on the Editing Task.

Comparison of Inpainting Results Under Malicious Editing. This evaluation focuses on the effectiveness of protection against malicious background modification. Participants assess images where an adversarial edit is applied, such as changing the background to a completely different scene (e.g., from a home setting to a jail).

Each method is rated on a 1 to 5 scale:

- 1 – If the edit is successfully applied, meaning the protection failed.
- 5 – If the edit is blocked, indicating effective protection.

By aggregating responses, we can evaluate how well our method prevents unauthorized modifications compared to existing approaches.

D. More Experimental Results

In this section, we present additional results and comparisons with baseline methods.

D.1. More Qualitative Results

Figure 3 visualizes the edits produced by different protection techniques. Notably, our method introduces the most significant alterations to the input image, effectively changing the subject’s identity and making it less recognizable. Compared to the baselines, our approach adds minimal perturbation artifacts, thus offering strong protection at lower imperceptibility of the added noise.

We provide additional qualitative results for the inpainting task in Figure 4. As shown in the immunized images,



Figure 4. More Results on the Inpainting Task.

our method confines noise to the masked region, whereas other approaches, such as AdvDM, MIST, and SDS(-), apply noise across the entire image. The qualitative results demonstrate that our approach is more effective at preventing malicious modifications, outperforming baseline methods in blocking adversarial intent.

D.2. Comparison across DCT-Shield Variants

In this section, we present a systematic ablation study to isolate and evaluate the individual contributions within the DCT-Shield framework. We investigate how introducing adversarial perturbations into different color-space channel coefficients affects overall robustness. Additionally, for the inpainting task, we examine the role of incorporating masking during the optimization process, analyzing its impact on edit protection. Through this careful comparison, we aim to elucidate which elements are most critical for enhancing adversarial resilience in our proposed method.

Y vs YCbCr. The Y-channel variant adds adversarial perturbations only in the Y-channel, offering improved JPEG robustness with lower perceptual noise. Figure 5 compares it with the base (YCbCr) variant in terms of the protection-perception trade-off. While the base variant per-

forms better before purification (green lines), the Y-channel variant consistently outperforms it after purification across JPEG levels.

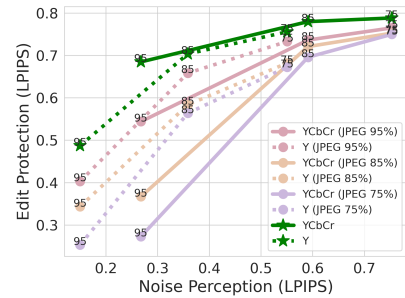


Figure 5. Comparison between the performance of Y-Channel and Base (YCbCr) Variants of DCT-Shield. The x-axis shows noise perceptibility (LPIPS; lower is better), and the y-axis shows edit protection (LPIPS; higher is better). Solid and dashed lines represent the base (YCbCr) and Y-channel variants, respectively. Green lines indicate edit protection before JPEG purification, while other colors indicate JPEG robustness after purification. Marker numbers denote the Q_{alg} values used by DCT-Shield for immunization.



Figure 6. **JPEG Robustness.** Figure shows the editing results of protected images generated at different compression levels Q_{alg} (INQ in figure) using DCT-Shield (left-right) under different JPEG purification levels (top-bottom))

Masked vs Unmasked. The masked variant was introduced specifically for inpainting tasks. In Table 1, we compare the masked and unmasked (base) variants of DCT-Shield. While the unmasked variant performs slightly worse for inpainting, it still provides reasonable protection, comparable to other baselines. When object masks are available, we recommend using the masked variant, as it offers stronger defense with lower perceptual noise confined to the masked region.

	LPIPS \uparrow	FID \uparrow	CLIP \downarrow
DCT-Shield (Masked)	0.547	199.082	0.674
DCT-Shield (Unmasked)	0.532	188.827	0.691

Table 1. Comparison of Masked and Unmasked (base) variants of DCT-Shield. Refer to Table 2 of main paper for baseline results.

D.3. Robustness against Other Defenses

In addition to the purification techniques evaluated in the main paper, we present further results in this supplementary section on Gaussian Noising and Noisy Upscaling from [3], as well as Impress, introduced in [1]. Fig. 7 shows that DCT-Shield achieves superior robustness against Gaussian Noise and Impress compared to other baselines. In contrast, all methods—including ours—exhibit only modest robustness against Noisy Upscaling, highlighting a promising direction for future work on improved immunization.

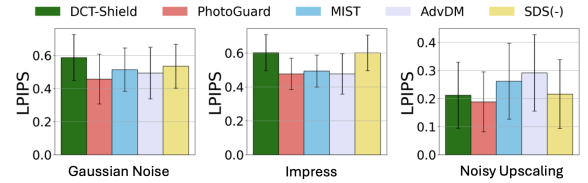


Figure 7. **Robustness against additional purification techniques.** Comparison of DCT-Shield with baseline methods in terms of LPIPS(\uparrow) between edits of original and purified images.

D.4. JPEG Robustness as a Function of Q_{alg}

In figure 6, we visualize the edit protection by varying the quality parameter Q_{alg} in DCT-Shield. We observe that running our algorithm at $Q_{alg} = q$ generates images that provide protection at attackers compression quality $q' \geq q$.

D.5. Cross-Model Transferability

Our method provides transferable defense, as demonstrated in Table 2. Images immunized with the masked variant of DCT-Shield (using the VAE from Stable Diffusion Inpainting 1.0) were subsequently edited using Stable Diffusion Inpainting 2.0. DCT-Shield outperforms baselines in terms of transferability, offering stronger edit protection.

	DCT-Shield	PhotoGuard	DiffGuard	AdvDM	MIST	SDS (-)
LPIPS \uparrow	0.527	0.506	0.495	0.477	0.502	0.495
CLIP \downarrow	0.714	0.716	0.717	0.74	0.722	0.726

Table 2. Cross-model transfer to Stable Diffusion Inpainting 2.0 from Stable Diffusion Inpainting 1.0.

E. Limitations

While our method provides strong protection against malicious edits, it is specifically designed for the LDM model, which limits its applicability to other editing frameworks. Advanced purification techniques, such as Noisy Upscaling, substantially diminish the effectiveness of both our method and baseline approaches, underscoring the need for further research and improvement.

References

- [1] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. In *Advances in Neural Information Processing Systems*, pages 10657–10677. Curran Associates, Inc., 2023. 6
- [2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 1
- [3] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative AI. In *The Thirteenth International Conference on Learning Representations*, 2025. 6
- [4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [6] Hamad R. Sheikh and Alan C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 2
- [7] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2
- [8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3