Anti-Tamper Protection for Unauthorized Individual Image Generation

Zelin Li Ruohan Zong Yifan Liu Ruichen Yao Yaokun Liu Yang Zhang Dong Wang
University of Illinois Urbana-Champaign
{zelin3, rzong2, yifan40, ryao8, yaokunl2, yzhangnd, dwang24}@illinois.edu

Abstract

With the advancement of personalized image generation technologies, concerns about forgery attacks that infringe on portrait rights and privacy are growing. To address these concerns, protection perturbation algorithms have been developed to disrupt forgery generation. However, the protection algorithms would become ineffective when forgery attackers apply purification techniques to bypass the protection. To address this issue, we present a novel approach, ***Anti-Tamper Perturbation (ATP)****. ATP introduces a tamper-proof mechanism within the perturbation. It consists of protection and authorization perturbations, where the protection perturbation defends against forgery attacks, while the authorization perturbation detects purification-based tampering. Both protection and authorization perturbations are applied in the frequency domain under the guidance of a mask, ensuring that the protection perturbation does not disrupt the authorization perturbation. This design also enables the authorization perturbation to be distributed across all image pixels, preserving its sensitivity to purification-based tampering. ATP demonstrates its effectiveness in defending forgery attacks across various attack settings through extensive experiments, providing a robust solution for protecting individuals' portrait rights and privacy. Our code is available at: https://github.com/Seeyn/Anti-Tamper-Perturbation .*

1. Introduction

In recent years, with the development of personalized generation technology, online services for creating customized individual images have become widely available [8, 14, 27]. Users can easily create customized individual images with text prompts by submitting requests to online service providers such as Civitai[1] and Midjourney[2]. However, personalized generation technology also raises serious ethical and legal concerns. As shown in Figure 1(a), forgery attackers can create fake individual images using online personalized generation services, infringing on the data owner's portrait rights and privacy. In this scenario, the online service providers may inadvertently become accomplices of attackers by providing easily accessible services [7].

To defend the forgery attack, a set of protection methods [18–21, 31] have been proposed. They disrupt the personalized generation process by injecting protection perturbations into data owners' images. As shown in Figure 1(b), the service provider injects protection perturbations to degrade the quality of generated images to prevent the forgery attacks. However, the protection perturbations can be easily purified by even naive purification methods (e.g., resizing or JPEG compression) [13, 34]. The forgery attacker can purify the protection perturbations to bypass the protection schemes and generate fake individual images again.

As the essence of purification is to tamper the protection perturbation, the service provider can defend it by adopting a tamper-proof mechanism. As shown in Figure 1(c), the tamper-proof mechanism alerts the service provider when the protection perturbation is altered, allowing the service provider to counteract attacks by refusing generation requests from tampered images. This approach is similar to the Not Safe For Work (NSFW) content filtering mechanism [2], where the service provider inspects the generation request and its output to ensure no harmful content is present. Likewise, the tamper-proof mechanism detects and prevents attacks targeting tampering protection perturbations, enhancing the service provider's defense capabilities. It is noteworthy that the tamper-proof mechanism is designed from the perspective of the service provider, where the purpose is to prevent the forgery attacker from misusing the service to launch a forgery attack. The attackers may still be able to launch the attack on their own devices. However, it is not the focus of this paper because it is not the service provider's responsibility and requires high-performance computing resources from the attacker rather than the effortless API calls used to exploit the service.

Implementing a tamper-proof mechanism is challenging, as tamper-proofing requires protection perturbations to contain verifiable information that can be checked for potential tampering. However, protection perturbations are funda-

[1] https://civitai.com
[2] https://www.midjourney.com

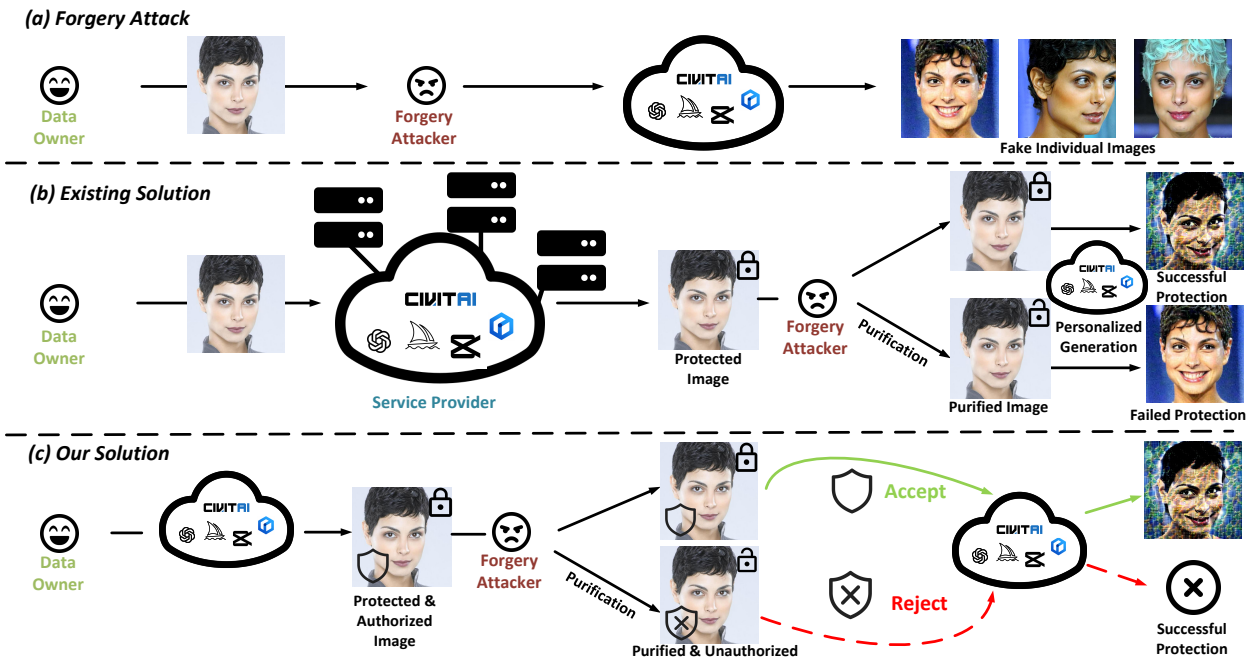


Figure 1. (a) The forgery attacker generates fake individual images of the data owner by taking pictures from social media and submitting them to the service provider. (b) The data owner can inject protection perturbation into their images with assistance from the service providers, causing low-quality results if an attacker tries to generate fake individual images. However, the protection can fail if the attacker purifies the protected images. (c) Our solution: We propose Anti-Tamper Perturbation with a tamper-proof mechanism. If no purification is applied, the perturbation protects the data owner's portrait rights and privacy by degrading the quality of generated images. Conversely, if a forgery attacker applies purification, the image becomes unauthorized, and the service provider rejects the generation request.

mentally adversarial noise, inherently designed to mislead deep learning models rather than encode structured information [9]. Its creation depends on specific model architectures, loss functions, and input images, making it unsuitable for information embedding.

Considering this challenge, we propose a new perturbation design, **Anti-Tamper Perturbation (ATP)**. The ATP consists of two components: *protection* and *authorization* perturbations. The protection perturbation is responsible for safeguarding the image against the forgery attacks. The authorization perturbation encodes an authorization message into the image. When purification occurs, the integrity of this message is disrupted, signaling that an unauthorized tamper attempt has occurred. It functions like a watermark [24, 33], yet the difference is that the existing watermark design is not sensitive to purification, making it incompatible with the authorization perturbation design. When both protection and authorization perturbations function simultaneously, we can achieve the objective of tamper-proofing. However, a fundamental challenge arises as the protection perturbation alters image information, making it conceptually a tampering manner. This creates an apparent dilemma: ***the authorization perturbation must remain intact despite changes induced by protection perturbation while still being vulnerable to removal by purification-based tampering attacks***. To address this challenge, we first adopt a Block Discrete Fourier Transformation (BDCT) to transform the image to the frequency domain. We design a gradient descent algorithm to generate protection perturbations in the frequency domain. An authorization perturbation network is proposed to generate the authorization perturbation, embedding an authorization message in the frequency domain. Both perturbations can be guided by a binary mask, which specifies the regions in the frequency domain where perturbations should be applied. The mask ensures that the authorization perturbation remains intact even after the protection perturbation is applied, as they are positioned in different regions of the frequency domain determined by the mask. Block Inverse Fourier Transformation (BIDCT) is then adopted to transform the image back to the pixel domain. Due to the transformation, the authorization perturbation is distributed across all pixels of the image, guaranteeing its sensitivity to purification-based tampering. Since ATP combines both protection and authorization perturbations, it is notable that ATP can work with various existing protection perturbation algorithms. The contributions of this work can be summarized as follows:

1. To the best of our knowledge, our work is the first to introduce a tamper-proof mechanism for individual im-

age generation protection, creating a novel approach to defend against forgery attacks with purification.
2. We design the **Anti-Tamper Perturbation (ATP)** to implement the tamper-proof mechanism. ATP comprises *protection* and *authorization* perturbations. The protection perturbation defends the image against forgery attacks, while the authorization perturbation remains unaffected by the protection perturbation yet retains sensitivity to purification-based tampering.
3. We evaluate the effectiveness of ATP in various attack scenarios through extensive experiments. The results show that ATP can be integrated with different protection perturbation designs. Existing solutions face an inevitable performance drop under attacks with purification. In contrast, ATP achieves a 100% protection success rate due to the sensitive tamper-proof mechanism triggered by purification tampering.

2. Related Works

Individual Image Generation. The diffusion model is a leading technique in image generation [6, 12, 26]. It can generate an image by using text as a condition to guide the generation process [28]. However, text conveys less detail than images, making it difficult to achieve specific results through text prompts alone, particularly for generating personalized content such as customized selfies [27]. To address this limitation, individual image generation methods (e.g., Text Inversion [8], DreamBooth [27]) were developed. These approaches aim to "learn" a unique token (e.g., *sks*) that can represent a specific person or object. Diffusion models can apply this token to generate images with specific subjects [8, 27]. Online service providers use individual image generation methods to provide the customized generation service, but the service might be misused for forgery attacks. ATP is designed to address this by preventing unauthorized use of personal images.

Protection Perturbation. Protection perturbation can be embedded within the image to safeguard against unauthorized generation. The perturbation is typically generated by maximizing the diffusion model's loss function, as first proposed by Liang et al. [19], who demonstrated that these perturbed images could act as adversarial examples for diffusion models. Le et al. [18] then introduced Anti-DB that enhanced AdvDM by incorporating Projected Gradient Descent (PGD) along with an Alternating Surrogate and Perturbation Learning strategy. Xu et al. [31] presented CAAT to demonstrate that the cross-attention layer is critical in training diffusion models. This indicates that targeting the perturbation to disrupt image-text mapping can effectively enhance protection performance. Liu et al. [21] proposed Metacloak that learns perturbations over a pool of surrogate models and applies the expectation-over-transformation technique to enhance the protection perturbation robustness against purification.

Perturbation Purification. A key limitation of the protection perturbation is its protection performance drop when purification occurs. The purification can disrupt the perturbation's integrity and weaken its protective capability [13, 34]. It is reported that naive purification techniques, such as image resizing and JPEG compression, would allow attackers to bypass the protection perturbation designs [13, 34]. Furthermore, Zhao et al. [34] introduced an advanced method, GridPure, to effectively purify protection perturbations. As reported by [13, 34], the protection performance drop caused by purification remains a significant challenge. To address this issue, we introduce ATP, which shifts the focus from resisting purification to verifying perturbation integrity. ATP implements a tamper-proof mechanism for protection perturbation, allowing the service provider to reject generation requests of purified images.

3. Anti-Tamper Perturbation

The pipeline of Anti-Tamper Perturbation is shown in Figure 2. The image is transformed by Block Discrete Cosine Transformation (BDCT) into the frequency domain. Guided by a binary mask, the authorization and protection perturbations are applied separately in the *frequency domain* before being transformed back to the *pixel domain*.

Mask-Guided Perturbation Blending. Both protection and authorization perturbation work by altering the image information. They can interfere with each other by altering the same image pixel. To address this problem, we propose distinguishing the perturbation by a mask as follows:

$$P_{AP}(I) = M \odot P_{Auth}(I) + (1 - M) \odot P_{Prot}(I), \quad (1)$$

where I denotes the image, and M represents the mask, composed of 0 and 1. The mask consists of the values $\{x|x \in \{0, 1\}\}$ sampled from the Bernoulli distribution $x \sim \text{Bernoulli}(p)$. As a result, we can adjust the hyperparameter p to control the ratio of 0s and 1s in the mask. The $P_{AP}(\cdot)$, $P_{Auth}(\cdot)$, $P_{Prot}(\cdot)$ refer to the Anti-Tamper Perturbation, Authorization Perturbation, and Protection Perturbation, respectively. The mask allows the two perturbations to function separately, ensuring they do not interfere with each other. However, this design also introduces a limitation: the perturbations become distinguishable in the pixel space, which may allow purification methods to target the protection perturbation selectively.

To address this issue, we can adopt a transformation function that maps the image to the frequency domain, ap-

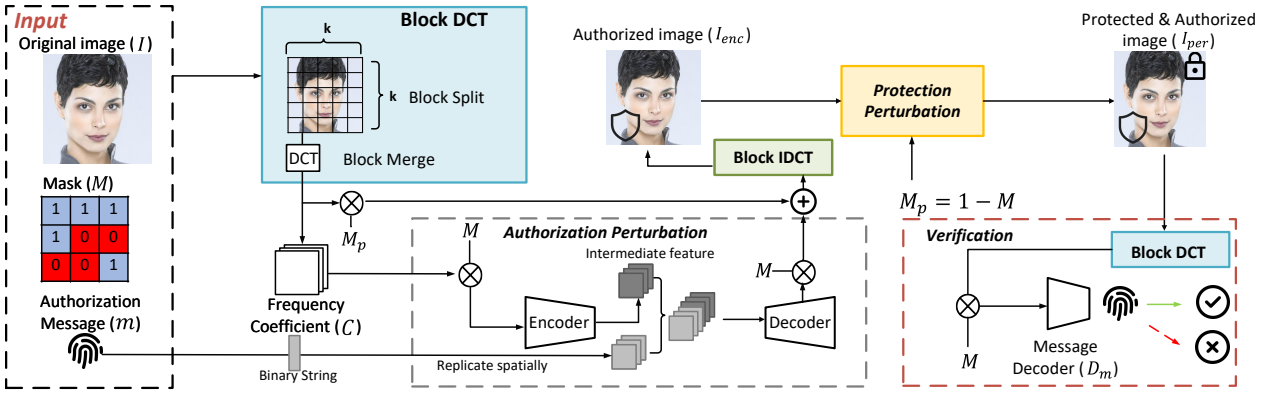


Figure 2. Pipeline of the Anti-Tamper Perturbation. The original image is first transformed to the frequency domain using Block Discrete Cosine Transformation (BDCT). Guided by a binary mask, the authorization and protection perturbations are independently applied in the frequency domain to obtain a protected and authorized image.

plying perturbations in the frequency domain:

$$P_{AP}(I) = F^{-1}[M \odot P_{Auth}(F(I)) + (1-M) \odot P_{Prot}(F(I))], \quad (2)$$

where $F(\cdot)$ and $F^{-1}(\cdot)$ denote the function projecting the image from the pixel domain to the frequency domain and its inverse function. As each pixel value is a linear combination of frequency domain coefficients, both perturbations are uniformly distributed within the pixel domain. This approach ensures that the authorization and protection perturbations are indistinguishable in the pixel domain. The uniform spread of the perturbation also improves the sensitivity of authorization perturbation to purification attempts.

Block Discrete Cosine Transformation. We employ a BDCT function as the transformation function F, which is more efficient than directly adopting Discrete Cosine Transform (DCT) [1] to the whole image [25]. Specifically, we first divide image I into non-overlapping blocks in the pixel domain. We record the position of the block in the pixel domain, for each block, we apply DCT as follows:

$$C_{u,v} = \alpha(u)\alpha(v) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} I_{i,j} \phi(u,i,N) \phi(v,j,N), \quad (3)$$

where $\phi(u,i,N) = \cos\left(\frac{\pi(2u+1)i}{2N}\right)$, $\alpha(i) = \sqrt{\frac{2}{N}}$ when $i = 0$, $\alpha(i) = \sqrt{\frac{1}{N}}$ otherwise. N represents the height and width of the block. C denotes the frequency coefficients. Then, we merge the blocks based on their positions in the pixel domain. As a result, we can obtain a tensor of frequency coefficients with the same scale as the original image. We do the mask-guided authorization and protection perturbation to the tensor. To transform the image back from the frequency domain to the image domain, we apply the Block Inverse Discrete Cosine Transformation (BIDCT). We split the tensor into blocks in the same way as BDCT and apply Inverse-DCT to obtain the perturbed image in the pixel domain:

$$I_{i,j} = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} \alpha(u)\alpha(v) C_{u,v} \phi(u,i,N) \phi(v,j,N). \quad (4)$$

Authorization Perturbation. Drawing inspiration from previous work on image steganography [15], we consider embedding the authorization message in the frequency domain as a viable approach for authorization perturbation. Referring to the work [35], which has been proven to be effective in hiding information in the pixel domain, we use a convolutional autoencoder $f_\theta(\cdot)$ to complete the authorization perturbation. As illustrated in Figure 2, the encoder extracts intermediate feature maps from masked input frequency coefficients. The authorization message m, represented as a binary string of length L composed of $\{0,1\}$, is spatially replicated and concatenated with the intermediate feature map. The decoder then reconstructs the coefficients from these modified feature maps. A message decoder D_m is also trained to retrieve the encoded information from the reconstructed coefficients. The entire pipeline of the authorization perturbation can be formulated as follows:

$$g_\theta(C) = (1-M) \odot C + M \odot f_\theta(M \odot C, m), I_{enc} = F^{-1}[g_\theta(F(I))], \quad (5)$$

where I_{enc} denotes the image with authorization perturbation. Referring to the model training of [35], we incorporate an image reconstruction loss $\mathcal{L}_{rec}$ and an adversarial loss $L_{adv,G}$ to minimize alterations to the image content after perturbation. For the accuracy of information hiding, we adopt a mask-guided message consistency loss and a regularization loss calculated in the frequency domain:

$$\mathcal{L}_{con} = ||D_m(M \odot f_\theta(C)) - m||_2^2, \mathcal{L}_{reg} = ||f_\theta(C) - C||_2^2. \quad (6)$$

The consistency loss facilitates message hiding, while the regularization suppresses significant changes in the frequency domain to keep the pixel domain values within the

Algorithm 1 Improved Frequency Domain PGD

Input: Loss for perturbation L, Image for perturbation I, Guiding Mask M_p, Block DCT F, Block IDCT F^{-1}, PGD radius ϵ, Step size α.
Output: Perturbed Image I_{per}.

$\nabla \leftarrow \frac{\partial \mathcal{L}}{\partial I}$
$\nabla_{freq} \leftarrow M_p \odot F(\nabla)$
$\nabla \leftarrow F^{-1}(\alpha \cdot \text{sgn}(\nabla_{freq}))$
$I_{per} \leftarrow I + \nabla$
$I_{per} \leftarrow F^{-1}(\Pi_{\epsilon,F(I)}(F(I_{per})))$ ▷ $\Pi_{\epsilon,I}(\cdot)$ constrain its output within an ϵ-ball around $F(I)$
return I_{per}

allowable range after inverse transformation. The total loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{con} + \lambda_{adv}\mathcal{L}_{adv,G} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{reg}\mathcal{L}_{reg}. \quad (7)$$

By combining these loss functions, the authorization perturbation can embed messages covertly within the frequency domain of the image.

Protection Perturbation. All the protection perturbation algorithms essentially generate the perturbation according to gradients derived from the diffusion model loss function. Based on the gradients, Projected Gradient Descent (PGD) [22] is the commonly used algorithm to update the perturbation. As suggested by [10], in the frequency domain, performing mask-guided perturbation is analogous to omitting gradients that would alter frequency coefficients outside the target region for modification:

$$I_{per} \leftarrow \Pi_{\epsilon,I}(I + \alpha \cdot \text{sgn}(F^{-1}(M_p \odot F(\nabla)))), \quad (8)$$

where ∇ denotes the gradient calculated from diffusion model loss and ϵ is the PGD radius. $\Pi_{\epsilon,I}(\cdot)$ constrains its output within an ϵ-ball around I. $\text{sgn}(\cdot)$ is the sign function and α is the step size. M_p is a binary mask, where entries with value 1 indicate the frequency coefficients to be updated by the gradient, and entries with value 0 indicate those to remain unchanged. As shown in Figure 2, we can flip the mask M for authorization perturbation to obtain this mask. However, we find that updating the perturbation by Equation 8 can invalidate the mask guidance. The reason is that each pixel value is a linear combination of all frequency coefficients. The modification of a single pixel value influences all the frequency coefficients. The transformation of $\Pi(\cdot)$, $\text{sgn}(\cdot)$ in the pixel domain inevitably alters the frequency coefficients the mask intended to preserve. To address this problem, we propose our improved frequency domain PGD algorithm in Algorithm 1. $\Pi(\cdot)$, $\text{sgn}(\cdot)$ are moved to apply in the frequency domain to ensure accurate modification of specific frequency coefficients based on the mask. A comparison experiment is placed in Appendix B.1 to show the accuracy improvement.

Since existing protection perturbation methods rely on PGD for optimization, their algorithms can be directly adapted to use our Improved Frequency Domain PGD (Algorithm 1), facilitating integration into ATP with minimal changes.

4. Experiments

In this section, we evaluate our ATP design from different perspectives. First, we assess the effectiveness of ATP in defending against forgery attacks. Three attack scenarios are evaluated: 1) attacks with purification, 2) attacks without purification, and 3) adaptive attacks for ATP. Second, we evaluate the aesthetic impact of ATP on the image. Third, we assess the robustness of the authorization perturbation to modifications induced by the protection perturbation and its sensitivity to purification tampering.

Datasets. To train the authorization perturbation network, we utilize the FFHQ dataset [17], which comprises 70,000 high-quality images of human faces. For generating the protection perturbation, we follow the dataset selection of [18], using subsets of high-quality face image datasets CelebA-HQ [16] and VGGFace2 [3]. We select 50 subjects from each dataset, each represented by eight images, and split them equally into two subsets. One subset is used for training the protection perturbation, and the other is reserved for the algorithm requirement of [18] and metric calculations.

Evaluation Metrics. For authorization performance evaluation, we calculate the bit error of the extracted and original authorization messages, denoted as **Bit-error**. This metric represents how accurately the authorization perturbation hides the information in the image. The Bit-error increases when the purification attempts alter the authorization perturbation. Following the choice of [18, 21], for protection performance evaluation, we take Stable Diffusion v2-1 [26] as the base generation model and apply DreamBooth [27] for personalized image generation. We generate 16 images for each subject by prompt "a photo of *sks* person." (Additional results for a different generation model, personalized generation algorithm, and prompt can be found in Appendix B.2 & B.3). We assess the protection performance using four metrics: 1) **CLIP-IQAC**: CLIP-IQAC is proposed by [21] and adapted from CLIP-IQA [29], this metric evaluates human images in quality. 2) **LIQE**: LIQE is an image quality assessment metric that aligns well with human perception and applies to human and non-human images [32]. 3) **Face Detection Failure Rate (FDFR)**: FDFR measures the failure rate of face detection using Retinaface [4]. 4) **Identity Score Matching (ISM)**: ISM evaluates the identity consistency between the generated image and the original image. The identity embedding of generated images and the original images are extracted by Arcface [5]. ISM is computed by measuring the cosine similarity between the generated

image's identity embedding and the average embedding of the original images.

The four metrics assess the impact of perturbations on generation quality from different perspectives. However, there is no clear standard for defining a "successful protection." As a result, we propose the Protection Success Rate (PSR), which quantifies the effectiveness of protection by establishing a threshold for CLIP-IQAC, where any generated image with a quality score below this threshold is deemed as successfully protected. We also define a Bit-error threshold. Any generation request with an unauthorized image whose Bit-error exceeds the threshold will be rejected. If a user submits four images for generation, and at least one of them is unauthorized, the service provider can reject the generation request. This is considered a successful protection, resulting in a PSR of 1.0. The experiments about why we select the CLIP-IQAC to calculate PSR and how we set the threshold for Bit-error and CLIP-IQAC are placed in Appendix A.2 & A.3.

Baselines. As ATP can be integrated with different protection algorithms, we select four state-of-the-art protection algorithms for our experiments: Anti-DB (2023)[18], AdvDM (2023)[19], CAAT (2024)[31], MetaCloak (2024)[21]. By modifying the gradient descent approach of these protection algorithms as described in Algorithm 1, we can adapt the baseline protection perturbation to ATP. The details of ATP's implementation are in Appendix A.1. The mask ratio $p = 0.5$ and the BDCT block width $N = 16$ are determined based on the ablation studies in Appendix B.4 and B.5.

Protection Performance Under Attacks with Purification. In this experiment, we demonstrate the ATP design is effective for protecting individual image generation under attacks with purification. We select two purification methods for naive purification: JPEG compression and image resizing, which have been reported to be effective [21, 34] to purify the protection perturbation. For the advanced purification method, we choose GridPure (2024) [34]. The PSR metric is used to evaluate different methods across two datasets, and the results are shown in Figure 3. The results show, in the "clean" condition where no purification is applied, all methods achieve relatively high protection success rates. In this setting, ATP's tamper-proof mechanism is not triggered, and no unauthorized image is found. When purification is applied, the performance of different methods faces an unavoidable reduction in protection performance. Such performance decline is especially noticeable in settings like Resize 4x, JPEG 50, and GridPure (Visual results of the generated images after purification are provided in Appendix B.9). Even MetaCloak, a robust protection perturbation algorithm designed against purification, its PSR still inevitably drops. However, when combined with ATP, MetaCloak and other baselines all achieve a 100% Protection Success Rate, as ATP's tamper-proof mechanism reliably detects various purifications and triggers generation rejection. It indicates the effectiveness of ATP in defending against attacks with purification.

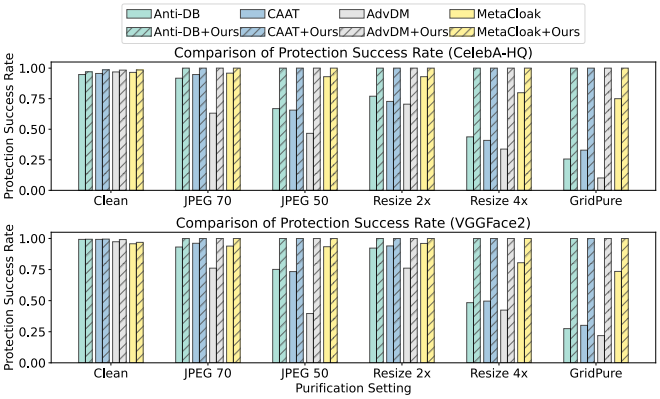


Figure 3. Comparison of Protection Success Rate for different methods across various purification settings.

	CelebA-HQ				VGGFace2			
	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑
Anti-DB	-0.2870	1.1168	0.4619	0.4575	-0.4340	1.0282	0.3201	0.6788
Anti-DB+Ours	-0.3139	1.0741	0.4647	0.5213	-0.3864	1.0984	0.2549	0.8188
AdvDM	-0.3361	1.0287	0.4166	0.6638	-0.3797	1.0163	0.4061	0.6100
AdvDM+Ours	-0.3621	1.0312	0.4119	0.6675	-0.4370	1.0355	0.2540	0.8638
CAAT	-0.3261	1.0872	0.4577	0.4700	-0.5008	1.0139	0.2977	0.7825
CAAT+Ours	-0.3568	1.0768	0.4315	0.6338	-0.4673	1.0291	0.2497	0.7863
MetaCloak	-0.3418	1.4243	0.4911	0.4850	-0.3358	1.2247	0.4857	0.5025
MetaCloak+Ours	-0.3639	1.3140	0.4048	0.7038	-0.3628	1.1573	0.4179	0.6338

Table 1. Quantitative results for CelebA-HQ and VGGFace2 datasets across various metrics.

Protection Performance Under Attacks without Purification. In this experiment, we show the protection performance of adopting an ATP design under attacks without purification, where the tamper-proof mechanism is not triggered. The results are shown in Table 1. By adding baseline protection perturbation with ATP design, we observe a consistent increase in FDFR, a decrease in ISM and CLIP-IQAC in most cases, and similar results for LIQE. As illustrated in Figure 4, each baseline retains its original capability to degrade image quality after adapting to ATP. The results show that although tamper-proof mechanism will not be triggered under attacks without purification, ATP can still achieve performance comparable to the original protection perturbation.

Figure 4. Qualitative comparison of original perturbation algorithms and their ATP modified versions in CelebA-HQ. More visual results from the VGGFace2 are provided in Appendix B.10.

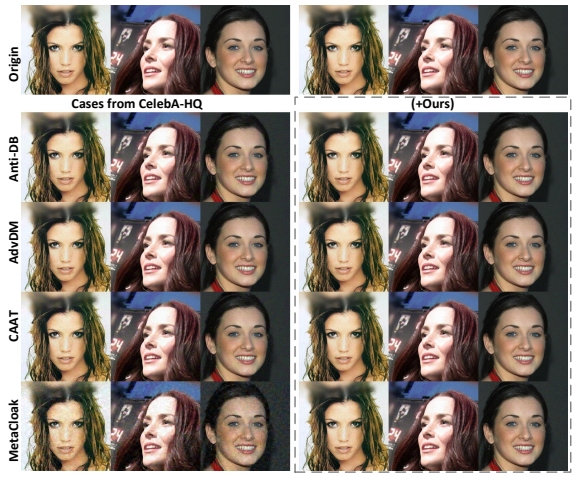


Figure 5. Perturbed images of different methods from CelebA.

	CelebA		VGGFace	
	CLIP-IQAC↑	ISM↑	CLIP-IQAC↑	ISM↑
Anti-DB	0.3925	0.7569	0.4220	0.7224
Anti-DB+Ours	0.5744	0.7716	0.4195	0.6934
AdvDM	0.5115	0.7686	0.4986	0.7289
AdvDM+Ours	0.5785	0.7703	0.4280	0.6910
CAAT	0.3886	0.7608	0.4211	0.7230
CAAT+Ours	0.5733	0.7712	0.4519	0.7088
MetaCloak	-0.0139	0.7126	0.2046	0.6766
MetaCloak+Ours	0.4923	0.7581	0.4390	0.6982

Table 2. Quantitative evaluation of the aesthetic impact of different perturbation algorithms across two datasets.

	BDCT	Improved-PGD	Mask	Bit-error (e^{-3})↓
(a)				349.84
(b)			✓	42.031
(c)	✓			360.31
(d)	✓		✓	81.719
Ours	✓	✓	✓	0.4688

Table 3. Comparison of different fusion designs with Bit-error values. **BDCT** determines whether the BDCT is used to transform the image to the frequency domain; If not, the perturbation is applied directly in the pixel domain. **Improved-PGD** indicates whether we adopt Algorithm 1; If not, we adopt Equation 8. **Mask** specifies whether we use the guiding mask of ratio 0.5; If not, the perturbations are applied without the guidance of the mask.

Protection Performance Under Adaptive Attacks. For the adaptive attack, we consider three settings: (1) the attacker knows the mask value but not the BDCT hyper-parameters, (2) the attacker knows the hyper-parameters but not the mask value, and (3) the attacker knows both. We conduct experiments on MetaCloak with ATP using the VGGFace2 dataset and apply a rounding function to purify perturbations in the frequency domain. In setting (1), we change the BDCT block size from 16 to 8, leading to a significant increase in Bit-error from $6.25 \times e^{-4}$ to $3.18 \times e^{-1}$, causing the authorization verification to fail. For (2), we apply the rounding action to all frequency coefficients, which raises the Bit-error from $6.25 \times e^{-4}$ to $5.04 \times e^{-1}$, also resulting in verification failure. In (3), the attack successfully bypasses verification, reducing the Protection Success Rate to 0.33. While ATP is ineffective in setting (3), its strong resistance in (1) and (2) highlights its robustness against attackers with partial knowledge. The vulnerability in (3) is due to the complete leakage of both BDCT hyper-parameters and the mask value, which is extremely unlikely to happen. Even without considering variations in BDCT hyperparameters, the search space of a binary random mask (with a size of $512 \times 512 \times 3$) remains $C_{786432}^{393216} \approx 2^{786414}$ given a known mask ratio of 0.5. Such a large search space makes it practically impossible for an attacker to retrieve the mask without human-induced leakage. Therefore, ATP remains a viable defense under adaptive attack conditions, particularly when attackers have only partial knowledge.

Aesthetic Impact of Perturbation. The perturbation may degrade image quality and compromise identity consistency, potentially leading to poor aesthetics and discouraging image owners from adopting such techniques. In this experiment, we evaluate the aesthetic impact of ATP and other protection perturbations using ISM and CLIP-IQAC. Since both metrics are computed on the perturbed images rather than the generated ones, higher values indicate less aesthetic impact. As shown in Table 2, the combination with ATP generally leads to increases in both ISM and CLIP-IQAC, suggesting that ATP introduces minimal aesthetic degradation. This observation is further supported by the qualitative comparisons in Figure 16. Additional visual results from VGGFace2 are provided in Appendix B.8.

Robustness to Protection Perturbation. In this experiment, we evaluate different perturbation fusion strategies to analyze the factors that influence the robustness of the au-

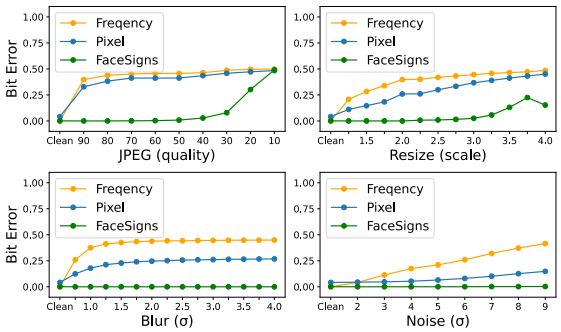


Figure 6. Sensitivity of ATP to different types of purification. The x-axis indicates the hyperparameter of different purifications, while the y-axis indicates the Bit-error.

thorization perturbation against interference from the protection perturbation. We incorporate Anti-DB into ATP for this experiment on CelebA-HQ. VGGFace2 results are provided in Appendix B.7. The robustness is evaluated by the Bit-error, the lower the error, the higher the robustness. From the results in Table 3, we observe that settings (a) and (c) exhibit a large Bit-error, underscoring the importance of mask guidance. Comparing setting (b) with "Ours", we find that encoding messages in the frequency domain leads to less Bit-error than in the pixel domain. The comparison between setting (d) and "Ours" indicates that Equation 8 invalidates the mask guidance, which results in an increased Bit-error. The experiment highlights the importance of using a mask to guide the fusion of authorization and protection perturbations. It also demonstrates that frequency-domain authorization perturbation is more effective at concealing information than its pixel-domain counterpart. Moreover, the results underscore the necessity of Algorithm 1 for accurate mask guidance.

Sensitivity to Purification. In this experiment, we compare the sensitivity of frequency-domain ATP versus pixel-domain ATP to purification. We also compare the performance with the existing watermark method FaceSigns [24]. For the pixel-domain design, we adopt setting (b) from the robustness experiment. We incorporate Anti-DB into ATP for this experiment on CelebA-HQ. VGGFace2 results are provided in Appendix B.7. We select four types of purification to purify the images with ATP: Resize, JPEG, Gaussian Blur, and Gaussian Noise. The hyperparameters include the JPEG compression quality, downsampling scale for resizing, the sigma value for Gaussian blur (with a kernel size of 3), and Gaussian noise. The results are shown in Figure 6. The Bit-error rises sharply as purification intensity increases, indicating the high sensitivity of ATP to various types of purification. Compared to its pixel-domain counterpart, the frequency-domain implementation of ATP achieves a lower Bit-error when no purification is applied, but exhibits a steeper increase under purification. This suggests that frequency-domain authorization perturbations are both more effective at concealing messages and more sensitive to purification. This increased sensitivity arises from the fact that frequency-domain perturbations are uniformly distributed across the pixel domain, making them more vulnerable to purifications. Additional theoretical analysis of this phenomenon is provided in Appendix A.4. In contrast, FaceSigns exhibits low sensitivity to purification due to its robustness against such operations, indicating that existing solutions are not directly compatible with ATP design.

5. Discussion

While ATP demonstrates strong effectiveness in our evaluated scenarios, it introduces additional computational cost and degrades to a regular protection perturbation when attackers operate on their own devices to bypass the verification process. To address concerns about deployment cost, we provide a scalability analysis in Appendix B.11, showing that ATP remains computationally affordable for the service provider. Despite these limitations, ATP still offers a practical and robust defense mechanism by preventing service providers from inadvertently contributing to forgery attacks. Meanwhile, since image generation requires significantly more computational resources than purification attacks, the need for a generation-capable device raises the barrier for potential attackers. These limitations also point to a promising direction for future work: eliminating the need for an explicit verification process from the service provider. One possible direction is to design authorization perturbations whose disruption degrades the generation result, thereby removing the reliance on explicit verification.

6. Conclusion

This paper introduces a novel perturbation design called Anti-Tamper Perturbation (ATP), motivated by the challenge that forgery attackers can bypass protection perturbation defenses through purification techniques. To address this issue, the ATP incorporates a tamper-proof mechanism. When purification occurs, the integrity of the ATP is compromised, signaling to the service provider that the image has been altered. This allows the provider to reject the unauthorized image and mitigate the threat of forgery attacks with purification. Extensive experiments conducted on two datasets demonstrate that ATP consistently outperforms state-of-the-art baselines in preventing forgery attacks under a wide range of purification methods. This paper highlights the potential of ATP as a solution for resisting forgery attacks, offering greater feasibility for safeguarding portrait rights and personal privacy in real-world scenarios.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974. 4

[2] Artsmart. What is nsfw in ai image generation?, 2024. 1

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018. 5

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5202–5211. Computer Vision Foundation / IEEE, 2020. 5

[5] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 5

[6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 3

[7] Benj Edwards. Thanks to ai, it's probably time to take your photos off the internet. *Ars Technica*, 2022. Accessed: 2024-11-01. 1

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1, 3

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[10] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, pages 1127–1137. AUAI Press, 2019. 5

[11] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 5

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[13] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative AI. *CoRR*, abs/2406.12027, 2024. 1, 3

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1

[15] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2021. 4

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 5

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 5

[18] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2116–2127. IEEE, 2023. 1, 3, 5, 6, 2

[19] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 20763–20786. PMLR, 2023. 3, 6

[20] Hanwen Liu, Zhicheng Sun, and Yadong Mu. Countering personalized text-to-image generation with influence watermarks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 12257–12267. IEEE, 2024.

[21] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24219–24228. IEEE, 2024. 1, 3, 5, 6, 2

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 5

[23] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial

domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 2

[24] Paarth Neekhara, Shehzeen Hussain, Xinqiao Zhang, Ke Huang, Julian J. McAuley, and Farinaz Koushanfar. Facesigns: Semi-fragile watermarks for media authentication. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(11): 337:1–337:21, 2024. 2, 8

[25] K. R. Rao and P. Yip. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press, San Diego, CA, revised edition edition, 2014. 4

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 3, 5

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510. IEEE, 2023. 1, 3, 5

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 3

[29] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 2555–2563. AAAI Press, 2023. 5, 2

[30] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[31] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24534–24543. IEEE, 2024. 1, 3, 6, 2

[32] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14071–14081. IEEE, 2023. 5, 2

[33] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11964–11974, 2024. 2

[34] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24398–24407. IEEE, 2024. 1, 3, 6

[35] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 682–697. Springer, 2018. 4

Anti-Tamper Protection for Unauthorized Individual Image Generation

Supplementary Material

In the appendix, we will include more experimental results and the detailed settings for anti-tamper perturbation (ATP).

A. Details of the Anti-tamper Perturbation

A.1. Hyper-parameter configuration

Experiment Environment. All experiments were conducted on a server equipped with 4 L40S GPUs (each with 48G) and an Intel(R) Xeon(R) Gold 6426Y CPU. The system had 251 GB of RAM. The software environment included Pytorch 2.4.1 running on Ubuntu 22.04.4 LTS, with CUDA 12.3 and cuDNN 9.1.0.70 for GPU acceleration. We didn't do distributed training, so the experiment can be conducted using one GPU.

Authorization Perturbation Hyper-Parameters. The authorization perturbation network is trained on FFHQ for 65,000 steps with a batch size of 8. For the weights of the loss function: $\lambda_{adv} = 1e-3$, $\lambda_{rec} = 0.7$, $\lambda_{reg} = 10$. The length of the authorization message m is 32, and the default mask ratio p is 0.5.

Protection Perturbation Hyper-Parameters. APT can adopt the existing protection design, and different baselines have varying choices for PGD radius and step size. Unlike the baselines, our method performs calculations in the frequency domain, so we did not select the same hyperparameters as the baseline.

Method	CelebA-HQ	VGGFace2
	Radius / Step Size	Radius / Step Size
Anti-DB+ours	5e-2 / 5e-3	250e-3 / 25e-3
AdvDM+ours	1e-1 / 2e-3	250e-3 / 25e-3
CAAT+ours	5e-2 / 5e-3	250e-3 / 25e-3
MetaCloak+ours	150e-3 / 5e-3	200e-3 / 5e-3

Table 4. PGD Radius and Step Size for different methods on CelebA-HQ and VGGFace2.

We observed the loss performance after adapting the baseline to our algorithm and selecting the appropriate PGD radius and step sizes. However, we did not perform detailed hyperparameter tuning experiments, as our main objective was to demonstrate that our method does not degrade the baseline's protection performance.

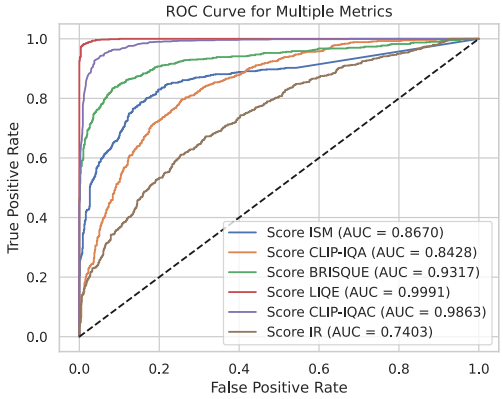


Figure 7. The ROC curve of different metrics.

A.2. Metric Selection

To select suitable metrics for evaluating the protection perturbation, we choose six metrics from the metrics adopted by existing works [18, 21, 31]: ISM [18], CLIP-IQA [29], BRISQUE [23], LIQE [32], CLIP-IQAC [21], IR [30]. When the model can't detect the face, the ISM value is set to -1 to guarantee that all generated images can get a corresponding ISM value. We first generate individual images using unprotected images from CelebA-HQ and those protected with Anti-DB. For each subject, we generate 16 images (50 subjects in total). We then calculate the value of the generated image's six metrics accordingly. We assume that the Anti-DB can often successfully protect the image when attackers do not make purification attempts. As a result, if the metric can classify images generated from protected images and those generated from unprotected images, it should be a reliable metric for evaluating protection performance. Images generated from protected images are categorized as negative samples, while those generated from unprotected images are categorized as positive samples. We then draw the ROC curves of the protection performance metrics, as shown in Figure 7. Among the metrics, CLIP-IQAC and LIQE show the highest AUC values, demonstrating the strongest discriminatory ability. As a result, we adopt them for the **Standard Protection Performance Comparison** in Section 4 (FDFR and ISM are also adopted, as ISM is the only metric among them that is directly related to facial identity. Furthermore, FDFR and ISM are typically computed together [18].). For the experiment of the **Protection Performance Under Attack Scenario**, we need to select one metric for calculating the Protection Success Rate (PSR). We use the property of the ROC curve to decide the

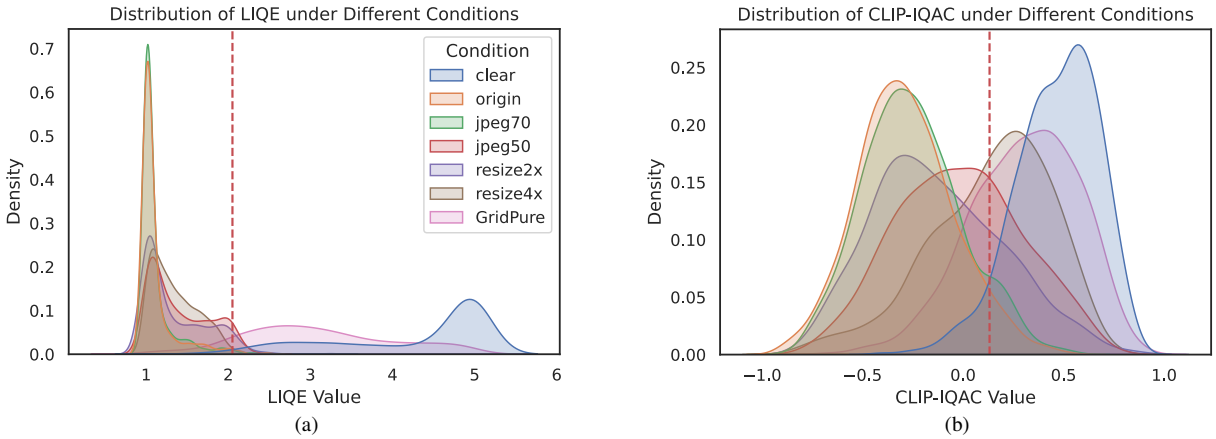


Figure 8. (a) Distributions of generated images evaluated by LIQE metric. (b) Distributions of generated images evaluated by CLIP-IQAC metric. The red dashed line illustrates the PSR threshold.

threshold of PSR. We select the threshold that can minimize $\sqrt{(1-TPR)^2+FPR^2}$, where TPR denotes true positive rate and FPR denotes false positive rate. **The threshold for CLIP-IQAC and LIQE are 0.1318359375, 2.05078125, respectively**.

Subsequently, we evaluate the performance of these metrics in capturing the impact of purification attempts on the protection mechanism. The distribution of the metrics for generated images is visualized through kernel density estimation. Specifically, "clear" and "origin" represent the generation results using unprotected and protected images, respectively. At the same time, the remaining categories correspond to the outcomes of applying the respective purification methods to protected images before generation.

As shown in Figure 8, the results demonstrate that CLIP-IQAC and LIQE effectively reflect the influence of purification attempts. Notably, following "resize 4x", "jpeg 50", and "GridPUre", the resulting distributions exhibit a convergence trend toward those of "clear." However, it can be observed that the PSR threshold of LIQE fails to capture the trend, as the majority of the samples fall to the left of the threshold. In contrast, CLIP-IQAC does not exhibit this issue, making it the preferred choice for calculating PSR.

A.3. Threshold Setting

We adopt Anti-DB for ATP to perturb the images in the CelebA-HQ test set. Then, we adopt purification techniques to purify the image. Figure 9 shows distinct differences in bit-error rate with and without purification. Since we aim to detect the occurrence of purification through the bit-error threshold, when the occurrence of purification significantly impacts the distribution of bit-errors, setting the threshold becomes a straightforward task. As a result, **we set the bit-error threshold of PSR to 3/32**. We adopt this value across different datasets and various protection perturbations, consistently finding that it can be effectively used to reject purification attempts.

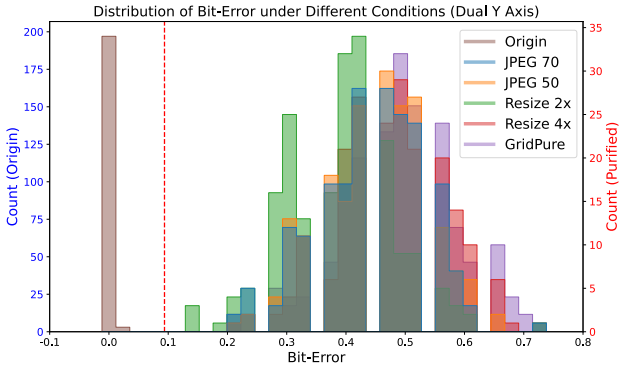


Figure 9. The distribution of bit-error under different purification settings. "Origin" denotes no purification applied. The red dashed line illustrates the PSR bit-error threshold.

A.4. Frequency-domain Sensitivity Analysis

In this section, we analyze why frequency-domain perturbation is inherently more sensitive (i.e., vulnerable) than the pixel-domain perturbation to the purifications. For pixel-domain purification (e.g., resizing), the vulnerability arises because the Block DCT computes each frequency coefficient as a weighted sum of all pixel values in a block. Thus, even a minor modification to a single pixel can affect all frequency coefficients. For frequency-domain purification (e.g., JPEG), the vulnerability stems from the fact that JPEG compression directly quantizes the frequency coefficients. These changes may be smoothed out in the pixel domain due to the inverse DCT and pixel rounding. To support this explanation, we define the change rate as the proportion

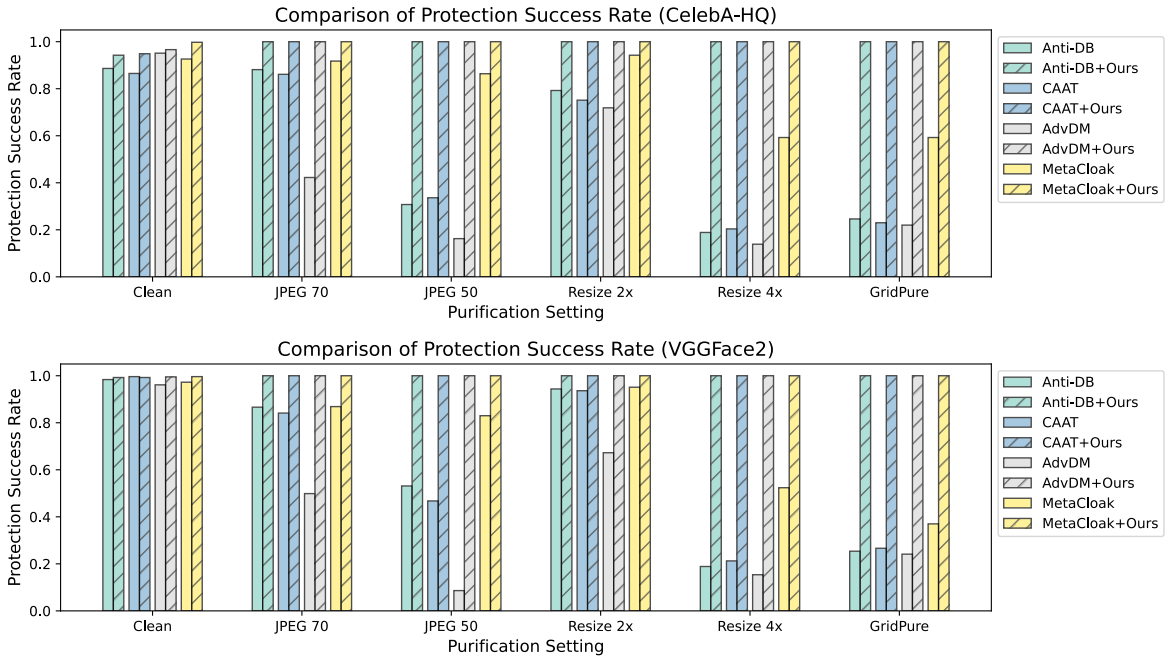


Figure 10. Comparison of Protection Success Rate for different methods across various purification settings. (Generated by prompt "a dslr portrait of sksperson")

Average Change Rate	Frequency Domain	Pixel Domain
4× Resizing	0.9714	0.7788
JPEG Compression (Q=50)	0.9594	0.8445

Table 5. The average change rate of coefficients and pixel values after performing different purifications.

of frequency coefficients or pixel values that vary before and after purification. We evaluate it on CelebA-HQ. As shown in Table 5, the frequency-domain perturbations have a higher probability of being changed after the purification, resulting in the inherent vulnerability.

Comparison of high- vs. low-frequency resilience to purification. While it is commonly assumed that high-frequency components are more vulnerable to traditional purification methods (e.g., resizing), our findings show that advanced purification techniques such as GridPure challenge this assumption. We want to share that different purification techniques have different preferences for altering frequency bands. We computed the average normalized variance of the DCT coefficient differences (within 16×16 blocks) before and after purification. As shown in the Figure 11, resizing primarily affects higher frequency bands (green-box region), whereas GridPure significantly alters low-frequency bands (red-box region).

Consequently, we adopt a random and uniform perturbation design in this project to ensure sensitivity to different purifications.

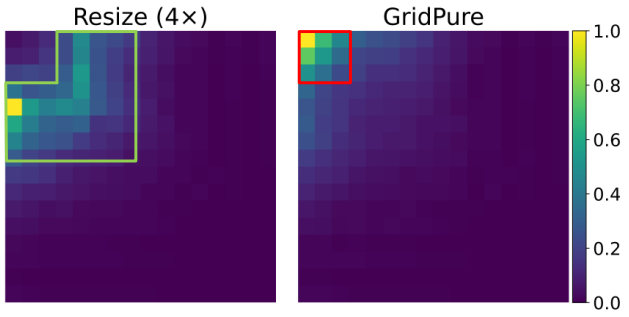


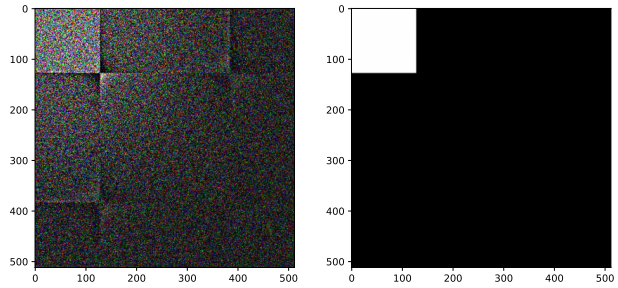
Figure 11. Visualization of the average normalized variance of the DCT coefficient differences (within 16×16 blocks) before and after purification.

B. More Experiment Results

B.1. Influence of Algorithm Design on Mask-guidance

In this experiment, we aim to demonstrate that Algorithm 1 validates the mask guidance. As outlined in the methodology section, the design of the projected gradient descent algorithm using Equation 8 is intended to invalidate the mask guidance. We verify this through a simulation experiment.

Specifically, we generate random gradients in the frequency domain, ensuring they are concentrated in the top-left 128 × 128 region. A 512 × 512 image is transformed into the frequency domain via DCT, and a one-step gradi-



(a) By Equation 8 (b) By Algorithm 1

Figure 12. Visualization of change in the frequency domain after the gradient descent. The visualizations depict the changes in frequency domain coefficients after the updates, where black represents no change, and brighter values indicate greater changes.

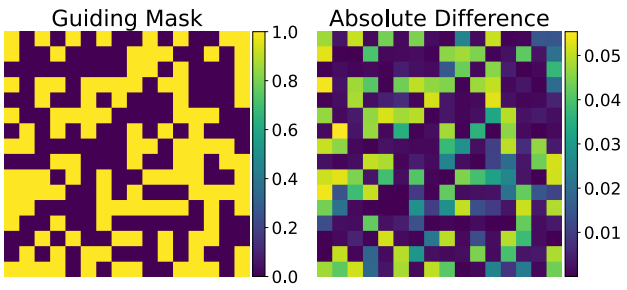


Figure 13. Visualization of the absolute difference in one 16×16 DCT coefficient map before and after applying the protection perturbation, along with the corresponding guiding mask.

ent descent is conducted using Equation 8 and Algorithm 1 (the step size is 1, and the PGD radius is 1). Subsequently, we visualize the changes in frequency domain coefficients after the single gradient descent step. As illustrated in Figure 12, our algorithm successfully confines the coefficient updates to the designated region in the frequency domain, whereas the original algorithm fails to achieve such precise localization.

In addition to the simulation experiment, we also provide a visualization of the absolute difference in one 16×16 DCT coefficient map before and after applying the protection perturbation using Algorithm 1, along with the corresponding guiding mask used during optimization. As illustrated in Figure 13, the perturbation primarily affects regions where the guiding mask is activated (i.e., mask value = 1), confirming that the mask guidance effectively constrains the perturbation by Algorithm 1 (Improved Frequency Domain PGD).

B.2. Repeat Main Experiments with different prompt

Following the experimental setup described in [18, 21], we evaluate the protection performance of our method using an alternative prompt: "a dslr portrait of sks person", to generate individual images. We adopt the same experimental setup described in Section 4, with the sole distinction being the prompt utilized.

Figure 10 shows that, with the new prompt, ATP continues to safeguard individual image generation effectively. This is because the integrity-check mechanism prevents generation before the prompt is utilized, ensuring that the performance of this mechanism remains unaffected by variations in the prompt. Table 6 reveals that, under the new prompt, ATP still performs comparably to the original protection perturbation approaches when the purification techniques are not applied.

B.3. Generalizability Analysis

We report the protection performance of ATP (CAAT) trained on SD2.1 when applied to a different diffusion model (SD1.5) and personalization method (SVDiff [11]) using CelebA in Table 7. We compare the protection performance of ATP against that of the unprotected baseline (i.e., without any perturbation applied).

The results demonstrate that ATP is generalizable across diffusion models and personalization techniques.

B.4. Performance Trade-off on Mask Ratio

The authorization and protection perturbations in the frequency domain can be distinguished based on the random mask M. The mask ratio p controls the region in the frequency domain used for authorization versus protection. This experiment shows that adjusting the mask ratio achieves a performance balance between protection and authorization for the ATP.

For example, as the mask ratio increases, a larger portion of the frequency domain will be allocated to authorization. As shown in Table 9 and Table 10, the increase in mask ratio leads to a decrease in bit-error, reflecting an improvement in message embedding accuracy. It also decreases protection performance, as LIQE, CLIP-IQAC, ISM, and FDFR scores indicate. Thus, we adopt a mask ratio of 0.5 as the default setting to achieve a balanced trade-off between authorization and protection performance.

B.5. Performance Trade-off on Block Size

The frequency domain transformation is achieved by BDCT. One of the hyperparameters for it is the size of the Block. In this section, we report the influence of this hyperparameter on the information hiding of authorization perturbation. We train the authorization model using different block sizes and evaluate it on CelebA-HQ. Figure 14 visualizes the variation in Bit-error under different block sizes. Since the block is square-shaped, we use its side length to represent the block size. It can be observed that a size of

	CelebA-HQ				VGGFace2			
	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑
AntiDB	-0.2047	1.3403	0.3944	0.3775	-0.4274	1.0228	0.3233	0.7950
AntiDB+Ours	-0.3085	1.1027	0.3509	0.4513	-0.4635	1.0250	0.3073	0.6850
AdvDM	-0.2979	1.0450	0.3193	0.6325	-0.3763	1.0305	0.3650	0.6213
AdvDM+Ours	-0.3367	1.0459	0.3634	0.4638	-0.4703	1.0126	0.3103	0.6538
CAAT	-0.1927	1.3018	0.4139	0.3025	-0.4890	1.0080	0.2819	0.7888
CAAT+Ours	-0.3257	1.0999	0.3725	0.4075	-0.4902	1.0192	0.2914	0.6963
Metacloak	-0.2573	1.4254	0.3892	0.5000	-0.4485	1.1075	0.3513	0.8613
Metacloak+Ours	-0.4049	1.0891	0.3488	0.7975	-0.4694	1.0447	0.3500	0.8875

Table 6. Quantitative results for CelebA-HQ and VGGFace2 datasets across various metrics. (Generated by prompt "a dslr portrait of sks person")

SD1.5 + DreamBooth				
	CLIP-IQAC↓	LIQE↓	ISM↓	FDFR↑
Origin	0.5007	4.5427	0.6824	0.0125
ATP	-0.2893	1.1929	0.4329	0.2988
SD2.1 + SVDiff				
	CLIP-IQAC↓	LIQE↓	ISM↓	FDFR↑
Origin	0.3837	4.3704	0.6679	0.1338
ATP	-0.3307	1.0484	0.3861	0.5575

Table 7. Protection Performance of ATP when generation model and algorithm are changed.

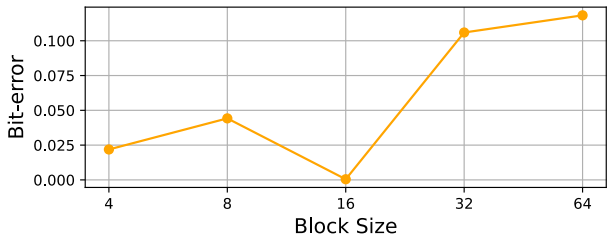


Figure 14. The Bit-error variation under different block size.

16×16 yields the lowest Bit-error, supporting our design choice adopted in the project.

B.6. Protection Performance Achieved Using Only Authorization Perturbation

In this section, we discuss the protection performance when we don't include protection perturbation in the ATP design. We compare the protection performance of images with no perturbation, images with authoirzation perturbation and images with ATP (taking CAAT as protection perturbaiton) in CelebA-HQ. As shown in the Table 8, authorization perturbation alone fails to provide strong protection when purification is not applied.

As a result, the combination of protection perturbation and authorization perturbation (ATP) is crucial for achieving reliable protection.

	CLIP-IQAC↓	LIQE↓	ISM↓	FDFR↑
Origin (No Perturb)	0.4659	4.2340	0.7053	0.0975
Authorization Alone	0.3258	3.6935	0.6414	0.1075
ATP (CAAT)	-0.3568	1.0768	0.4315	0.6338

Table 8. The protection performance using only the authorization perturbation is significantly worse than that of ATP.

Ratio	Bit-error (e^{-3}) ↓	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑
0.25	0.7813	-0.3561	1.0471	0.3805	0.6400
0.50	0.4688	-0.3139	1.0741	0.4647	0.5213
0.75	0.3125	-0.1480	1.3582	0.5765	0.2225

Table 9. Performance comparison for different mask ratios on CelebA-HQ.

Ratio	Bit-error (e^{-3}) ↓	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑
0.25	3.1250	-0.422682	1.135657	0.205465	0.91375
0.50	0.7813	-0.386397	1.098367	0.254911	0.81875
0.75	1.2500	-0.371436	1.03989	0.340458	0.70625

Table 10. Performance comparison for different mask ratios on VGGFace2.

	BDCT	Improved-PGD	Mask	Bit-error (e^{-3})↓
(a)				366.09
(b)			✓	43.594
(c)	✓			503.75
(d)	✓		✓	79.844
Ours	✓	✓	✓	0.7813

Table 11. Comparison of different fusion designs with Bit-error values on VGGFace2.

B.7. Repeat Experiments on VGGFace2

We repeat the experiment on VGGFace2 to further validate the credibility of our conclusions in Section 4. We adopt the same experimental setup described in Section 4, with the sole distinction being the dataset utilized. The experiment results are shown in Table 11 and Figure 15.

B.8. Visualization of Perturbed Images

In Figure 16, we present perturbed images generated using different methods from the CelebA-HQ and VGGFace2

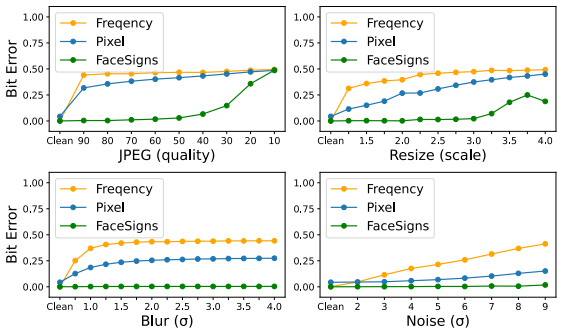


Figure 15. Sensitivity of ATP to different types of purification. The x-axis indicates the hyperparameter of different purifications, while the y-axis indicates the Bit-error. The images are from VGGFace2.

datasets. We observe that while perturbations are difficult to detect at normal scales, they become noticeable when viewed at an enlarged scale. This remains an unresolved challenge in the field and a focus of our future research efforts.

B.9. Visualization of Generation Results Applied Purification Techniques

We prepared visual cases to illustrate how purification techniques bypass existing protection mechanisms. Specifically, we present the results of individual image generation for images from CelebA-HQ and VGGFace2 after applying different protection perturbation algorithms. As demonstrated in Figure 17 and Figure 18, purification can bypass the protection provided by protection perturbation, compromising the safeguarding of individual image generation.

B.10. More Qualitative Results of Main Experiments

We present additional qualitative comparison results across various methods under two datasets (i.e., CelebA-HQ, VGGFace2) and two different prompts (i.e., a photo of sks person, a dslr portrait of sks person) in Figure 19, Figure 20, Figure 21, and Figure 22.

B.11. Scalability and Computational Efficiency Analysis

Scalability. A safety checker is deployed by the widely used diffusion model library "diffusers", which takes up 1159.60 MB. The authorization model only takes up 1.58 MB, which should be affordable by the service providers.

Computational Efficiency. The ATP requires extra time in authorization message hiding and verification. With batch size = 4, the averaged inference time costs are: Autoencoder encoding/decoding: 0.0201s/0.0274s; BDCT + IBDCT: 0.0016s. In the protection phase, ATP using CAAT as protection perturbation performs autoencoder encoding once and applies mask-guided PGD, which requires two additional BDCT+IBDCT operations per PGD step. This results in **0.38% increase** of the total protection time compared to the original CAAT protection (77.33s). In the generation phase, autoencoder decoding is performed once. When considering a generation method like DreamBooth (341.9s), the added decoding introduces **0.008% increase**.

Figure 16. Perturbed images of different methods from two datasets.

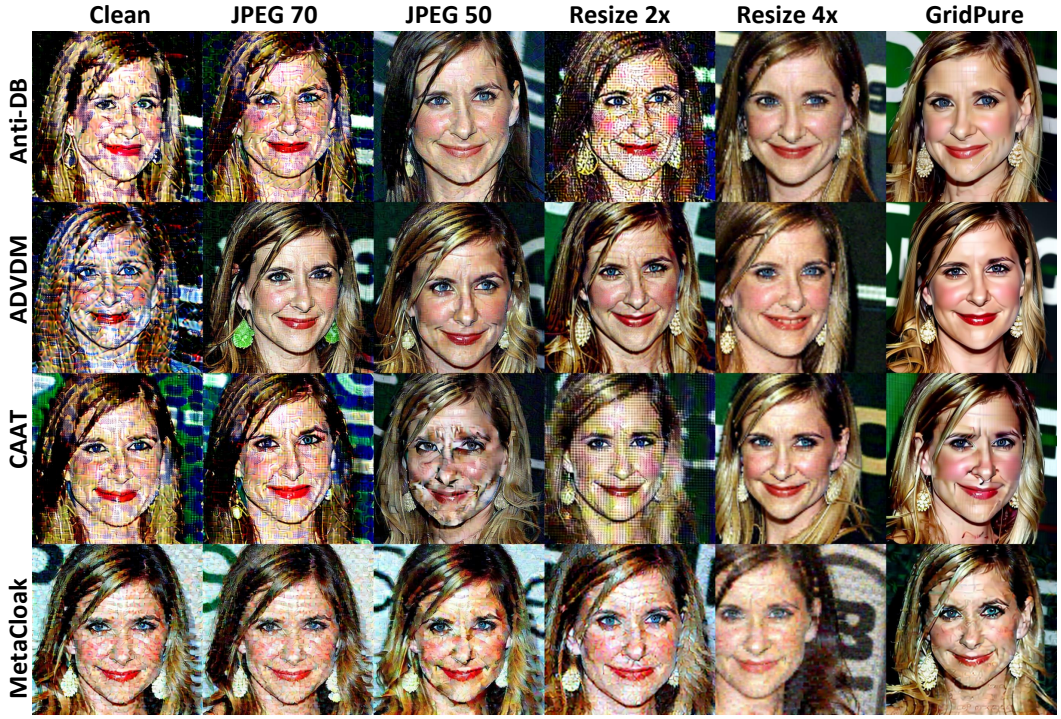


Figure 17. Visual cases showing the purification results bypassing the protection mechanisms on images from the CelebA-HQ dataset. "Clean" indicates no purification applied.

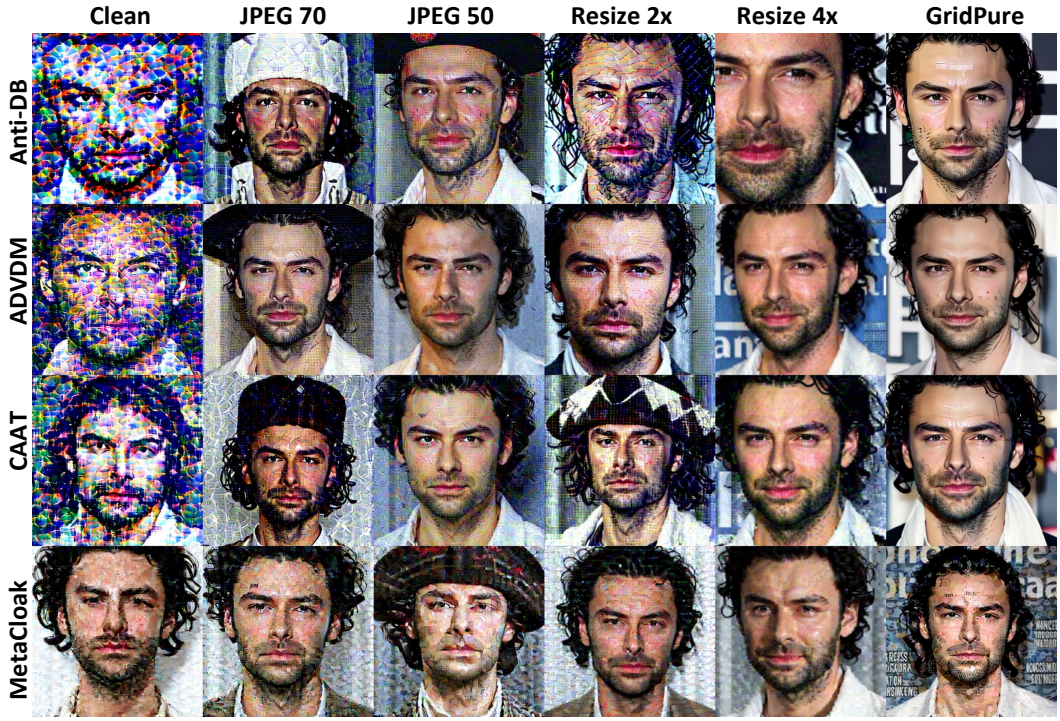


Figure 18. Visual cases showing the purification results bypassing the protection mechanisms on images from the VGGFace2 dataset. "Clean" indicates no purification applied.

Figure 19. Qualitative comparison of original perturbation algorithms and their ATP modified versions in CelebA-HQ.

Figure 20. Qualitative comparison of original perturbation algorithms and their ATP modified versions in CelebA-HQ.

Figure 21. Qualitative comparison of original perturbation algorithms and their ATP modified versions in VGGFace2

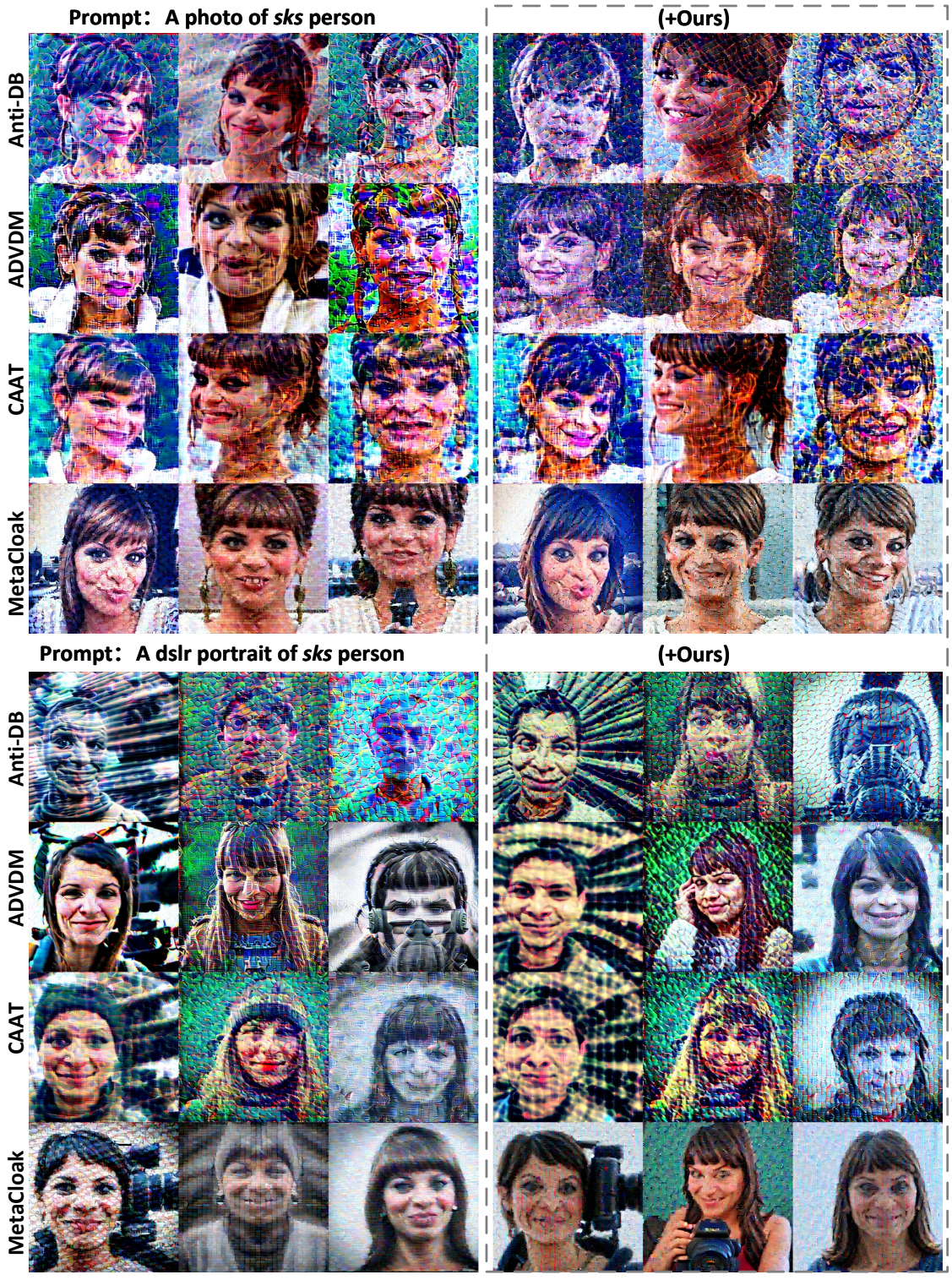


Figure 22. Qualitative comparison of original perturbation algorithms and their ATP modified versions in VGGFace2