Visual-Friendly Concept Protection via Selective Adversarial Perturbations

Xiaoyue Mi[1,2], Fan Tang[1,2*], You Wu[1,2], Juan Cao[1,2], Peng Li[3], Yang Liu[3,4]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Institute for AI Industry Research (AIR), Tsinghua University
[4]Department of Computer Science & Technology, Tsinghua University
mixiaoyue19s@ict.ac.cn, tfan.108@gmail.com, wwwuyou99@gmail.com, caojuan@ict.ac.cn, pengli09@gmail.com, liuyang2011@tsinghua.edu.cn

Abstract

Personalized concept generation by tuning diffusion models with a few images raises potential legal and ethical concerns regarding privacy and intellectual property rights. Researchers attempt to prevent malicious personalization using adversarial perturbations. However, previous efforts have mainly focused on the effectiveness of protection while neglecting the visibility of perturbations. They utilize global adversarial perturbations, which introduce noticeable alterations to original images and significantly degrade visual quality. In this work, we propose the Visual-Friendly Concept Protection (VCPro) framework, which prioritizes the protection of key concepts chosen by the image owner through adversarial perturbations with lower perceptibility. To ensure these perturbations are as inconspicuous as possible, we introduce a relaxed optimization objective to identify the least perceptible yet effective adversarial perturbations, solved using the Lagrangian multiplier method. Qualitative and quantitative experiments validate that VCPro achieves a better trade-off between the visibility of perturbations and protection effectiveness, effectively prioritizing the protection of target concepts in images with less perceptible perturbations.

Introduction

With the advent of popular image generative models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Song, Meng, and Ermon 2020) such as Stable Diffusion (Rombach et al. 2022) and GPT-4o (Achiam et al. 2023), people lacking expertise in drawing or photography can effortlessly create realistic or artistic works using simple textual descriptions. However, the success of these models has raised significant concerns about privacy, intellectual property rights, and various legal and ethical issues (Luo et al. 2024). For example, an adversary could easily generate sensitive specific concepts, such as personal fake images, or imitate renowned artworks for commercial purposes, using only a few reference images and concept personalization techniques such as Textual Inversion (Gal et al. 2022) or DreamBooth (Ruiz et al. 2023).

One prevailing direction to mitigate these potential risks is to leverage adversarial attack techniques, transforming original images into adversarial examples, termed "protected images". These protected images can misguide the personalized diffusion model and deceive its generation process, which resists malicious editing or personalization (Salman et al. 2023; Liang et al. 2023; Liang and Wu 2023; Le et al. 2023; Shan et al. 2023; Liu et al. 2024). For instance, AdvDM (Liang et al. 2023) employs adversarial attacks in the inversion stage of Stable Diffusion to safeguard against malicious imitation of the style of a specific artist. Later, its updated version, Mist (Liang and Wu 2023), enhances the protection efficacy of protected images by adding a textual loss, extending its application from Textual Inversion to DreamBooth. Unlike AdvDM and Mist, which focus on art style protection, Anti-DreamBooth (Le et al. 2023) undermines DreamBooth model generation quality to enhance privacy by adding adversarial perturbations to human face images before posting online. Furthermore, MetaCloak (Liu et al. 2024) leverages a meta-learning strategy and data transformations to generate more effective and robust protected images against DreamBooth.

However, these methods primarily focus on preventing the personalization methods from generating high-quality corresponding images, often resulting in noticeable and unacceptable perturbations. They usually use $11/255$ (Le et al. 2023; Shan et al. 2023; Liu et al. 2024) or $17/255$ (Liang and Wu 2023) as perturbation size in their paper, which is generally unacceptable for owners of face photos, significantly limiting the usability of these methods in real-world applications. Therefore, we pose the following question in this work: *How can we find a better trade-off between the visibility of perturbations and protection effectiveness?*

To answer this question, we highlight the sparsity of images under concept protection tasks: The critical information that deserves protection constitutes only a part of the image. Previous approaches aim to protect the entire image, including background regions and other non-essential information, enhancing the visibility of perturbations. As shown in Fig. 1, the protected images generated by Mist, Anti-DreamBooth, MetaCloak, SDS(-), and PhotoGuard, exhibit noticeable odd textures on the entire image: face, neck, and background, and their final protective effects akin to a special style picture of the target person. In contrast, we prioritize protecting essential information within an image, utilizing a more stealthy adversarial perturbation.

To this end, we introduce a Visual-Friendly Concept

*Corresponding author

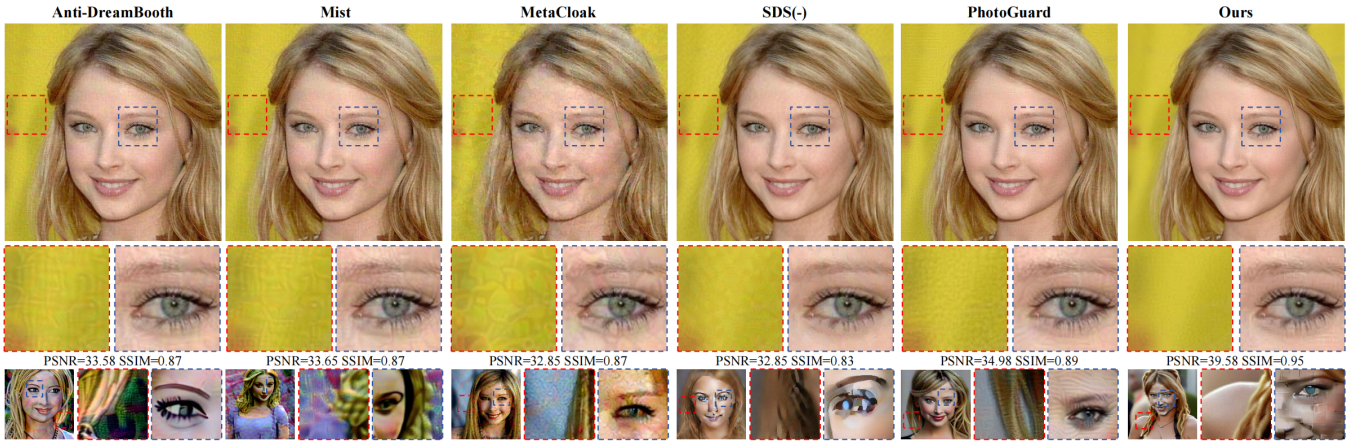

Figure 1: The first row shows protected images from Anti-DreamBooth (Le et al. 2023), Mist (Zhang et al. 2022), MetaCloak (Liu et al. 2024), SDS(-) (Xue et al. 2024), PhotoGuard (Salman et al. 2023), and our VCPro. Second and third rows show magnified regions and Textual Inversion results. At ϵ=8/255, our method better balances protection effectiveness and visual quality (higher PSNR/SSIM).

Protection (VCPro) framework to counteract unauthorized concept-driven text-to-image synthesis. This framework learns selective adversarial perturbations targeting important regions. Unlike discriminative tasks where important information is class-related and provided by target model gradients, identifying crucial information in generative tasks is challenging. VCPro utilizes user-provided masks for target concept protection, which can be supplemented by other privacy detection tools for online platforms. In that way, we propose a regional adversarial loss using spatial information to focus on selected areas. To further enhance visual quality, we apply a Lagrangian multiplier-based solution, shifting from maximizing protection effectiveness to minimizing perceptibility while ensuring effective protection. Considering human sensitivity to low-frequency changes, we measure perturbation perceptibility in the frequency domain. Experiments on models like Textual Inversion and DreamBooth validate VCPro's effectiveness. Our approach yields subtler adversarial perturbations compared to baselines like Mist and Anti-DreamBooth, especially FID, which is reduced from 96.24 to 27.04.

Our contributions are summarized as follows:

- We point out that the existing image protection methods over-emphasize the final protection effectiveness while neglecting the visual appearance of the protected images.
- We propose a visual-friendly concept protection framework that uses regional adversarial loss to protect essential image information. Considering human sensitivity, we measure perturbation perceptibility in the frequency domain and optimize for the smallest feasible perturbations rather than the strongest ones within size constraints.
- Experiments demonstrate that our approach can focus on crucial concepts specified by users with lower perceptibility than baselines, achieving a better trade-off between protection effectiveness and perturbation visibility.

Related Work

Personalization of Diffusions Models. Personalization for specific concepts (attributions, styles, or objects) has been a long-standing goal in the image generation field. In text-to-image diffusion models, previous researchers have primarily concentrated on prompt learning and test-time tuning of pre-trained models to generate images based on specific target concepts using special language tokens. Textual Inversion adjusts text embeddings of a new pseudoword to describe the concept (Gal et al. 2022). DreamBooth fine-tunes denoising networks to connect the novel concept and a less commonly used word token (Ruiz et al. 2023). Based on that, more works (Voynov et al. 2023; Zhang et al. 2023b; Kumari et al. 2023) are proposed to improve controllability and flexibility in processing image visual concepts. In this paper, we have selected Textual Inversion and DreamBooth as the techniques used by the adversary due to their popularity and representativeness.

Imperceptibility Adversarial Attack. Adversarial examples (Szegedy et al. 2013; Carlini and Wagner 2017; Duan et al. 2021; Mi et al. 2023) are initially introduced by adding imperceptible noise to original data, fooling classifiers into misclassifying with high confidence. Recently, more and more researchers aim to improve the imperceptibility of adversarial examples, and they make use of a variety of tools such as perceptual color distance (Zhao et al. 2021), low-frequency spatial constraints (Luo et al. 2022), hybrid attacks in frequency and spatial domain (Jia et al. 2022), and invertible neural networks (Chen et al. 2023), etc. But they are mainly aimed at discriminative tasks such as image classification, where important information in the image can be fed back relatively accurately by the gradient of the target model, whereas we target a diffusion-based generative model whose gradient is also for the whole image.

Adversarial Examples Against Unauthorized Diffusion Generation. Unauthorized AI generation poses significant safety risks, driving research into mitigation approaches.

While passive defenses focus on detecting synthetic images (Wu et al. 2022; Li, Luo, and Huang 2017), adversarial attacks offer promising protection against unauthorized generation (Ruiz, Bargal, and Sclaroff 2020; Wang et al. 2022a,b; Zhu et al. 2023). Recent works target diffusion models specifically. Photoguard (Salman et al. 2023) attacks VAE encoders to prevent malicious editing, while AdvDM (Liang et al. 2023) protects artistic styles by maximizing denoising loss. Mist (Liang and Wu 2023) enhances AdvDM with texture loss in latent space. GLAZE (Shan et al. 2023) measures art style similarity using pre-trained style-transfer models. Anti-DreamBooth (Le et al. 2023) protects privacy from DreamBooth learning, with MetaCloak (Liu et al. 2024) improving robustness via meta-learning. Xue *et al.* (Xue et al. 2024) introduce Score Distillation Sampling (SDS) loss to reduce computational costs. However, except for GLAZE, existing methods protect entire images with significant noise, degrading user experience and potentially overemphasizing backgrounds while missing critical information. We propose prioritizing limited noise to protect important semantic regions like faces and specific IPs, unlike GLAZE which targets art style.

Preliminaries

Personalization based on Diffusion. Concept-driven personalization customizes generative model outputs to align with specific concepts. Most techniques apply to latent diffusion models (LDMs, parameterized by θ) consisting of image encoder $\mathcal{E}_\theta$, decoder $\mathcal{D}_\theta$, condition encoder τ_θ, and denoising UNet ε_θ. For image x with latent code $z_0 = \mathcal{E}_\theta(x)$, the training objective is:

$$\mathcal{L}_\theta := \mathbb{E}_{z\sim\mathcal{E}(x),y,g\sim\mathcal{N}(0,1),t}\left[\|g - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right]. \tag{1}$$

Textual Inversion learns embeddings v for pseudo tokens sks in prompts like "a photo of sks [class noun]" by optimizing:

$$\arg\min_v \mathbb{E}_{z,y,g,t}\left[\|g - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right]. \tag{2}$$

DreamBooth fine-tunes LDM parameters on training images and class examples x^p to prevent catastrophic forgetting:

$$\begin{aligned}\theta := \arg\min_\theta \mathbb{E}_{z,y,g,t}\Bigg[&\|g - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2 \\ &+ \|g - \varepsilon_\theta\left(z_t^p, t^p, \tau_\theta(y^p)\right)\|_2^2\Bigg].\end{aligned} \tag{3}$$

Protected Image for Diffusions. Protected images add imperceptible adversarial perturbations δ to original images x to disrupt personalization. The protected image x' is formalized as:

$$\begin{aligned}x' := \arg\max_{x'} \mathcal{L}_\theta(x', y),\\ \text{s.t.} \quad \|x - x'\| \le \epsilon,\end{aligned} \tag{4}$$

where ϵ limits the perturbation budget. These adversarial attacks create samples difficult to denoise, enhancing LDM optimization challenges for image protection.

Visual-Friendly Concept Protection

Overview

Fig. 2 shows the pipeline of the proposed framework for visual-friendly concept protection. Accurately describing spatial positions through language can be challenging for users, but it can be precisely achieved using masks. By leveraging SAM (Kirillov et al. 2023) or other segmentation tools, users can generate a mask m for important concepts within a given image x. The user-provided masks and original images are then collectively fed into the protected image generation module as described in Eq. (8). In this module, we propose a regional adversarial learning loss to reduce the visibility of protected images through precise protection and a Lagrangian multiplier-based solution to minimize perturbations while maintaining successful protection.

In this section, we start by formulating the regional adversarial learning framework for diffusion models in Sec. and then move on to the solution of the proposed optimization objectives in Sec. . The total learning process is shown in Alg. 1.

Formulation

Regional Adversarial Loss. Unlike previous studies, we aim to achieve precise concept protection to reduce the perceptibility of protected images. We use mask m to indicate the spatial positions of important information in the feature map enabling precise concept protection. This precise optimization allows prioritized protection of the most critical information in the image with a smaller adversarial perturbation. The optimization objective of protected images in our method can be defined as:

$$\begin{aligned}x' := \arg\max_{x'} \mathcal{L}'_\theta\left(x', y, m\right),\\ \text{s.t.} \quad \|x - x'\| \le \epsilon,\end{aligned} \tag{5}$$

and the regional adversarial loss $\mathcal{L}'_\theta$ combines a "push" term for protected regions and a "pull" term for non-protected regions.

$$\begin{aligned}\mathcal{L}'_\theta :=& \mathbb{E}_{z\sim\mathcal{E}(x),y,g\sim\mathcal{N}(0,1),t}\left[l_{mask}(z_t, y, g, t, m)\right], \\ l_{mask} :=& \|\Delta \odot m\|_2^2 - \|\Delta \odot (1-m)\|_2^2,\end{aligned} \tag{6}$$

where $\Delta = g - \varepsilon_\theta(z_t, t, \tau_\theta(y))$. This loss operates through a balanced mechanism of opposing forces. The push term maximizes the distance between the predicted noise ε_θ and ground truth noise g in masked regions ($m = 1$), effectively disrupting the denoising process for protected concepts. Simultaneously, the pull term minimizes this distance in unmasked regions ($m = 0$), preserving visual quality in non-protected areas.

During optimization, these components generate distinct gradient signals: push gradients divert predictions away from ground truth in protected regions, while pull gradients maintain accuracy elsewhere. This dual approach is crucial for effective protection—without the pull component, regions outside the mask cannot provide sufficient gradient feedback, significantly impairing the optimization process. Our ablation studies in Table 2 demonstrate that removing the pull component substantially degrades protection effectiveness.

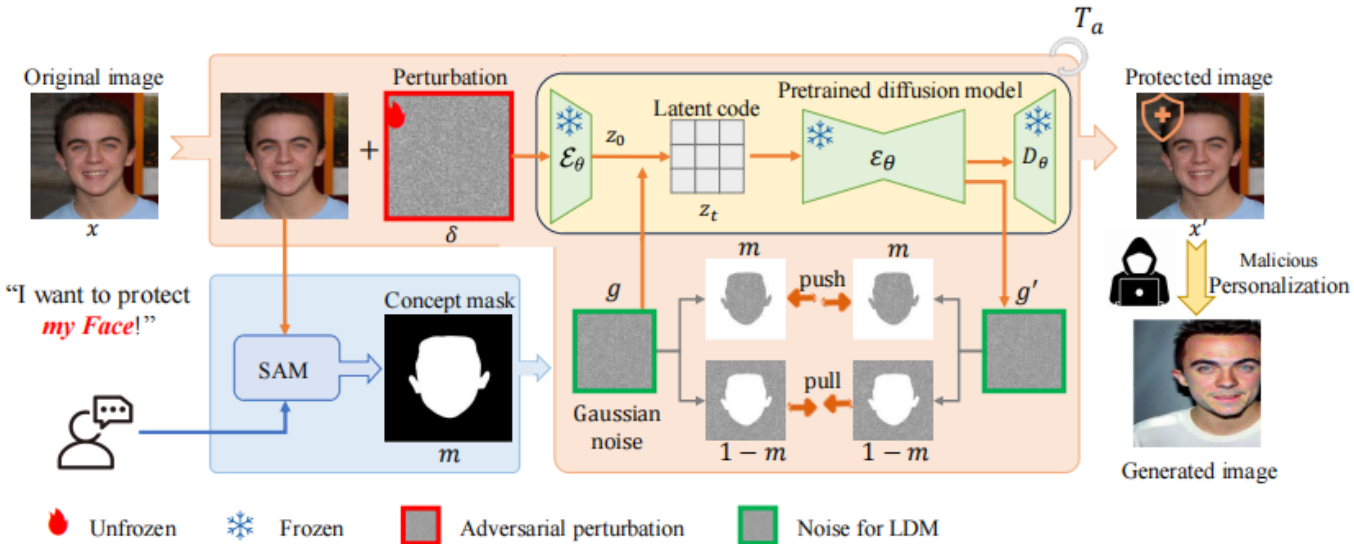

Figure 2: VCPro pipeline. Users create protective masks via SAM or other tools. The masks and images are fed into our protection module (Eq. 8), which uses regional adversarial learning to minimize perturbation visibility while maintaining protection.

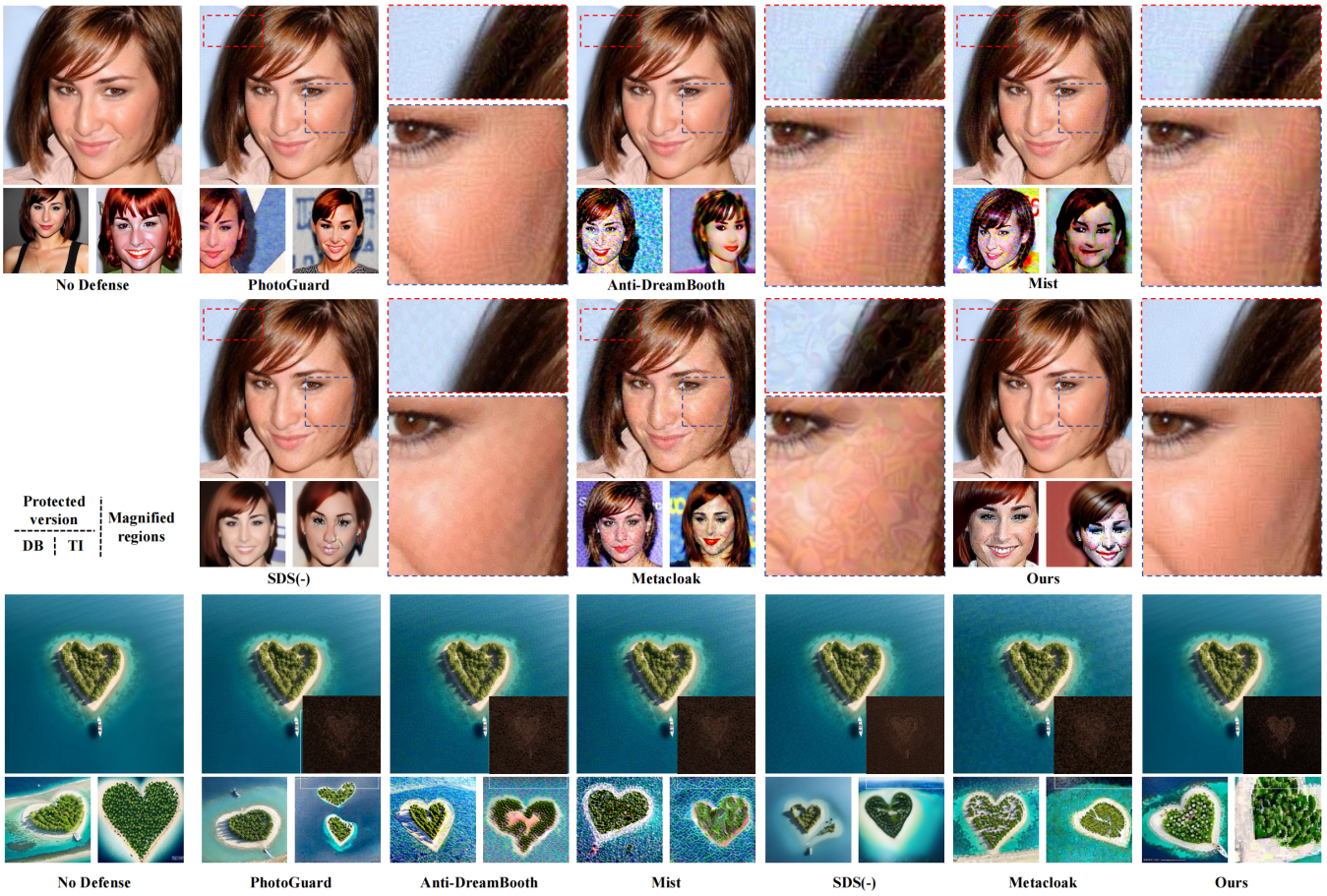

Figure 3: Qualitative comparison ($\epsilon = 8/255$, SD v1-4). Rows show original and protected images from baseline methods and ours, with TI/DB results. Perturbations visualized in black-yellow. Please zoom in.

Algorithm 1: Visual-Friendly concept protection (VCPro) framework

Require: Image x, diffusion model with parameter θ, number of time steps T, text condition y, attack steps T_a, step size α, and adversarial perturbation size ϵ

Initialize $x' = x$, $i = 0$
Get mask m by SAM or user-provide
while $i < T_a$ **do**
  Sample $t \sim [0, T]^n$
  $\delta = \text{Uniform}(-\epsilon, \epsilon)$
  Calculate $\mathcal{L}_{final}(x', y, m)$ by Eq. (8)
  $\delta = \delta - \alpha \cdot \text{sign}\left(\nabla_\delta \mathcal{L}_{final}(x', y, m)\right)$
  $\delta = \max(\min(\delta, \epsilon), -\epsilon)$
  $x' = x' + \delta$
  $x' = \max(\min(x', 255), 0)$
  $i = i + 1$
end while
return Protected image x'

Lagrangian Multiplier-based Loose Solution

Based on Eq. (6), protected images will destroy the target area as much as possible. However, for the generation task, the output can still be recognized as synthetic as long as there are some clear signs of protection. Typical image classification attacking methods (Carlini and Wagner 2017; Luo et al. 2022; Chen et al. 2023) also present similar viewpoints that minimize the impact on normal visual perception while maintaining the effectiveness of the attack. Hence, we propose a loosed optimization objective by attempting to find the minimal adversarial perturbation δ that can attack diffusion models successfully. For convenience, we set a heuristic method: if $\mathcal{L}'_\theta > \alpha$, the attack is successful. The problem can be translated as:

$$\begin{aligned} \text{minimize} \quad & D(x, x+\delta), \\ \text{such that} \quad & -\mathcal{L}'_\theta + \alpha \leq 0, \\ & \delta \in [-\epsilon, \epsilon]^n, \\ & x + \delta \in [0, 255]^n, \end{aligned} \tag{7}$$

where $D(\cdot)$ is a distance metric.

For ease of solution, we use Lagrangian multiplier method and get the alternative formulation:

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}_{final} = c \cdot D(x, x+\delta) - \mathcal{L}'_\theta + \alpha, \\ \text{such that} \quad & \delta \in [-\epsilon, \epsilon]^n, \\ & x + \delta \in [0, 255]^n, \end{aligned} \tag{8}$$

where a constant $c > 0$ is appropriately selected. The equivalence of Eq. (7) and Eq. (8) can be understood by the existence of a positive constant c ensuring the best solution for the second formulation aligns with that of the first.

Considering that the human visual system is more sensitive to low-frequency regions (Luo et al. 2022), we use $D(\cdot)$ to limit perturbations into the high-frequency regions. Specifically, we use discrete wavelet transform (DWT) to transform images from the spatial domain to the frequency domain. DWT will decompose the image x into one low-frequency and three high-frequency components, i.e., $x_{ll}, x_{lh}, x_{hl}, x_{hh}$, and inverse DWT (IDWT) uses all four components to reconstruct the image.

$$\begin{aligned} x_{ll} = LxL^T, \quad x_{lh} = HxL^T, \\ x_{hl} = LxH^T, \quad x_{hh} = HxH^T, \end{aligned}$$

where L and H are an orthogonal wavelet's low-pass and high-pass filters, respectively. x_{ll} preserves the low-frequency information of the original image, whereas x_{lh}, x_{hl} and x_{hh} are associated with edges and drastic variations.

In this work, we drop the high-frequency components and reconstruct an image with only the low-frequency component as $\widetilde{x} = \phi(x)$, where $\phi(x) = L^T x_{ll} L = L^T(LxL^T)L$, and $D(\cdot) = ||\widetilde{x} - \widetilde{x'}||_2^2$.

Experiments

This section introduces our experimental settings and qualitative and quantitative experiments to demonstrate our effectiveness in generating protected images with less visual perceptibility while safeguarding key concepts in user-provided images.

Experimental Settings

We validate our method on CelebA-HQ (Karras et al. 2017) and VGGFace2 (Cao et al. 2018) datasets, randomly selecting 50 identities from each with at least 15 images exceeding 500×500 resolution, following Anti-DreamBooth (Le et al. 2023) and MetaCloak (Liu et al. 2024). Using Stable Diffusion v1-4 with 512×512 resolution, we generate protected images with 720 iterations, step size $\alpha = 1/255$, and perturbation size $\epsilon = 8/255$. We test protection performance against Textual Inversion (Gal et al. 2022) (learning rate 5×10^{-4}, 3000 steps) and DreamBooth (Ruiz et al. 2023) (learning rate 5×10^{-7}, 1000 steps), comparing with five SOTA baselines: Mist (Liang et al. 2023), Anti-DreamBooth (Le et al. 2023), PhotoGuard (Salman et al. 2023), SDS(-) (Xue et al. 2024), and MetaCloak (Liu et al. 2024). We evaluate visual perception using FID (Heusel et al. 2017), SSIM (Wang et al. 2004), and PSNR between protected and original images, and assess protection effectiveness using LIQE (Zhang et al. 2023a) for full image quality and CLIP-FACE (Liu et al. 2024) for face regional quality evaluation. LIQE measures image quality on a five-point scale: $c \in C = \{1, 2, 3, 4, 5\} = \{$"bad", "poor", "fair", "good", "perfect"$\}$. CLIP-FACE is based on CLIP-IQA for visual quality by considering additional class information. All experiments are conducted on NVIDIA A100 GPU 40GB with parameters $c = 0.1$ and $\alpha = 0.5$. We provided detailed experimental settings in **Supplementary Materials**.

Main Results

Quantitative Results. Table 1 shows automatic and human evaluation results for protected image visibility and protection effectiveness. For visibility metrics, our method generates protected images most faithful to originals with stable performance (lowest FID variance). Compared to Anti-DreamBooth baseline, we achieve up to 69.20 FID reduction, 0.09 SSIM improvement, and 5.58 PSNR increase.

Dataset	Method	Visibility			TI		DB		Human Eval		
		FID↓	SSIM↑	PSNR↑	LIQE↓	CLIP-F↓	LIQE↓	CLIP-F↓	Vis.	TI↑	DB↑
VGGFace2	No Defense	–	–	–	3.61	0.24	3.95	0.38	**64**/25/11	–	–
	Mist	105.9	0.83	32.5	1.24	0.11	**1.02**	0.25	4/6/**90**	98.3	97.2
	Anti-DB	96.2	0.82	32.4	**1.18**	0.05	1.03	0.26	3/9/**88**	97.4	**100.0**
	PhotoGuard	62.1	0.84	33.2	1.50	0.13	1.26	0.27	13/19/**68**	**98.7**	84.0
	SDS(-)	47.5	0.81	32.9	2.09	0.07	2.74	0.33	13/16/**71**	94.6	77.1
	MetaCloak	204.3	0.82	32.1	1.86	0.19	1.33	0.25	2/10/**88**	91.0	90.4
	VCPro	**27.0**	**0.90**	**35.3**	2.31	**0.03**	2.06	**0.21**	–	98.6	97.4
CelebA-HQ	No Defense	–	–	–	4.40	0.41	4.84	0.63	**77**/19/4	–	–
	Mist	78.3	0.86	33.8	**1.45**	0.26	**1.04**	0.44	5/4/**91**	96.3	**100.0**
	Anti-DB	78.5	0.86	33.7	1.81	0.28	1.04	0.44	0/4/**96**	97.0	96.5
	PhotoGuard	45.2	0.88	35.1	1.81	0.30	1.23	0.48	2/22/**76**	94.3	94.3
	SDS(-)	34.5	0.82	33.0	2.29	0.09	2.92	0.55	9/11/**80**	98.6	84.3
	MetaCloak	161.9	0.86	33.1	1.47	0.25	1.59	0.44	0/4/**96**	96.0	95.5
	VCPro	**16.2**	**0.95**	**39.3**	2.61	**0.09**	2.62	**0.44**	–	**100.0**	95.7

Table 1: Comparison of protection methods on VGGFace2 and CelebA-HQ with $\epsilon = 8/255$. Metrics include visibility (FID, SSIM, PSNR), quality degradation (LIQE, CLIP-FACE), and protection efficacy (TI/DB: % synthetic images detected). Human evaluation visibility shows Lose/Tie/Win rates for VCPro vs. baselines. ↑/↓ indicates higher/lower is better. **Bold**: best performance.

Our method significantly outperforms current SOTA methods PhotoGuard and SDS (-) in image quality, though SDS (-) achieves low FID by introducing noticeable brightness increases that hurt SSIM performance.

For protection effectiveness, we achieve significant quality degradation in full-image metric LIQE (from "good/perfect" to "poor") and excel in face region quality metric CLIP-FACE. Both metrics demonstrate reduced visibility against perturbations while maintaining strong protection.

Qualitative Results. Fig. 3 shows protected images and effects on identity information and landscapes under Textual Inversion and DreamBooth. Adversarial perturbations are visualized using [0,1] normalization with colormaps. PhotoGuard, Mist, Anti-DreamBooth, and MetaCloak add obvious strange textures throughout images. SDS(-) causes excessive blurriness with limited DreamBooth protection, adding circular blob artifacts. Our approach achieves subtler perturbations while ensuring protective effectiveness. Our method prevents personalization methods from generating high-fidelity results. For faces, landmarks, and buildings, our perturbations are less noticeable (especially in backgrounds) while effectively distorting crucial textures and features. This prevents unauthorized use of personal or copyrighted material while maintaining better quality balance.

Textual Inversion protection is easier than DreamBooth since DreamBooth fine-tunes most Stable Diffusion parameters while Textual Inversion only adds word embeddings, as supported by Table 1.

User Study. We evaluate VCPro against five methods: PhotoGuard (Salman et al. 2023), Mist (Liang and Wu 2023), Anti-DreamBooth (Le et al. 2023), SDS(-) (Xue et al. 2024), and MetaCloak (Liu et al. 2024). 50 participants (58% male, 42% female, ages 18-55, mean: 24.5) with social media proficiency participated. We randomly sampled protected and generated images from 100 identities in VGGFace2 and CelebA-HQ plus six non-face groups. The study includes three parts: perturbation visibility (2064 valid votes), Textual Inversion protection (1996 valid votes), and DreamBooth protection (1990 valid votes).

User Study I. Participants compared VCPro-protected images with other methods' results, voting on visual quality ("A wins", "tie", or "B wins"). VCPro received 68%-96% of votes against competing methods with Kendall coefficient of 0.71 ($p < 0.05$), indicating substantial inter-rater agreement.

User Study II. Participants determined whether images generated via Textual Inversion from protected images were synthetic. All methods achieved $\geq 91\%$ synthetic recognition, with VCPro showing most effective protection. Cohen's Kappa = 0.88 ($p < 0.05$) indicates strong participant agreement.

User Study III. For DreamBooth protection evaluation, participants recognized 97.01% protective success rate for VCPro. Cohen's Kappa = 0.88 ($p < 0.05$) confirms study consistency.

Ablation Study

As shown in Table 2, we conduct ablation experiments on two VCPro modules: Regional Adversarial Loss (RAL) and Lagrangian Multiplier-based Loose Solution (LMLS), using Anti-DreamBooth as baseline. Anti-DreamBooth+RAL reduces perturbation visibility with protection concentrated in mask areas, resulting in slight LIQE decrease but stable CLIP-FACE performance. Smaller masks further reduce perturbation visibility but limit protection range. Anti-DreamBooth+LMLS reduces noise visibility and preserves high-frequency components, but final protection creates streaky textures across the entire image. This confirms that adversarial perturbations in specific spatial/frequency

Method	Visibility			TI		DB	
	FID↓	SSIM↑	PSNR↑	LIQE↓	CLIP-F↓	LIQE↓	CLIP-F↓
No Defense	–	–	–	3.61	0.24	3.95	0.38
Anti-DB	96.2	0.82	32.4	**1.18**	0.05	**1.03**	0.26
Anti-DB+RAL	68.2	0.83	32.9	1.62	0.03	1.53	0.18
Anti-DB+RAL (Small Mask)	58.4	0.84	33.1	2.11	0.05	2.21	0.27
Anti-DB+LMLS	30.1	0.90	35.0	2.52	0.17	2.16	0.30
VCPro (w/o Pull)	**17.2**	**0.94**	**36.8**	2.74	0.14	2.55	0.35
VCPro	27.0	0.90	35.3	2.31	**0.03**	2.06	**0.21**
Anti-DB+Input-Mask	41.3	0.90	35.2	2.05	0.10	2.03	0.23
VCPro+Input-Mask	18.6	0.93	36.6	2.41	0.07	2.68	0.31

Table 2: Ablation study on VGGFace2 with $\epsilon = 8/255$. Input-Mask constrains perturbations within the mask region on input images (vs. our method which constrains the optimization objective, allowing perturbations across the entire image). ↑/↓ indicates higher/lower is better. **Bold**: best performance.

ϵ	Visibility			Protection	
	FID↓	SSIM↑	PSNR↑	LIQE↓	CLIP-F↓
0	–	–	–	3.78	0.31
4	**23.0**	**0.93**	**36.5**	2.55	0.12
8	27.0	0.90	35.3	2.19	0.12
16	48.7	0.82	32.6	**1.84**	**0.09**

Table 3: Impact of ϵ on VGGFace2. Protection shows averaged LIQE and CLIP-F across TI/DB. ↑/↓: higher/lower is better.

domains directly affect corresponding protection domains. VCPro without pull-loss (VCPro (w/o Pull)) reduces training gradient feedback for protection while maintaining perturbation visibility constraints, decreasing both visibility and protection effectiveness. Comparing direct mask-constrained updates (Input-Mask) with our loss-guided approach shows both achieve regional protection. Direct masking improves pixel-level metrics (SSIM/PSNR) but makes perturbations more obvious with high FID. Our combined approach achieves better visual perception while maintaining target protection.

Perturbation Size Influence. Perturbation size ϵ controls the maximum allowable change in pixel values of adversarial perturbation. As shown in Table 3, different levels of adversarial perturbation size noticeably influence the protection outcomes: Compared to low ϵ, large ϵ presents worse invisibility of adversarial perturbations while the more obvious protection effects. For Textual Inversion, when the perturbation size is low, the image preserves the facial area with alterations in facial texture and feature distribution. Upon reaching a perturbation size of 16/255, the facial areas experience complete degradation. The results under DreamBooth and quantitative experiment show a similar trend. Compared with DreamBooth, Textual Inversion is easier to achieve concept protection.

Qualitative ablation experiments, **hyper-parameter (training iterations, c, α)**, analysis of adversary settings, and **frequency domain analysis** please see supplementary materials.

Conclusion

In this paper, we show that existing approaches utilizing adversarial perturbations to safeguard images from malicious personalization often overemphasize the final protection effectiveness, resulting in more noticeable perturbations. To mitigate this problem, on the one hand, we protect the important concept regions rather than the full images in previous works, leveraging the sparse nature of images and designing a user-specified image concepts protection framework. On the other hand, we change the optimization objective from generating the most protective adversarial perturbation to generating the least perceptible adversarial perturbation that exactly achieves the required protective effect. Quantitative and qualitative experiments demonstrate that we can protect important user-specified concepts and greatly reduce the degree of naked-eye visibility of adversarial perturbations.

Future Works and Limitations. Our future efforts will focus on finding efficient methods to produce protected images swiftly while maintaining high visual quality. By doing so, we aim to significantly enhance the overall user experience and ensure that our solutions meet the highest standards of both functionality and aesthetics.

Ethical Considerations. VCPro empowers individuals against AI content power imbalances. Technical measures complement limited legal protections. Dual-use potential—adversaries could theoretically exploit VCPro to evade accountability. However, the primary threat today is unauthorized personalization with limited user consent, and empowering individuals is essential given current power asymmetries.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under No. 62572458, and the Innovation Funding of ICT, CAS under Grant No. E561160.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition*, 67–74. IEEE.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy*, 39–57. IEEE.

Chen, Z.; Wang, Z.; Huang, J.-J.; Zhao, W.; Liu, X.; and Guan, D. 2023. Imperceptible adversarial attack via invertible neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 414–424. AAAI Press.

Duan, R.; Chen, Y.; Niu, D.; Yang, Y.; Qin, A. K.; and He, Y. 2021. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7506–7515. IEEE.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Jia, S.; Ma, C.; Yao, T.; Yin, B.; Ding, S.; and Yang, X. 2022. Exploring frequency adversarial attacks for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4103–4112. IEEE.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026. IEEE.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941. IEEE.

Le, T. V.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N.; and Tran, A. 2023. Anti-DreamBooth: Protecting Users from Personalized Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2116–2127. IEEE.

Li, H.; Luo, W.; and Huang, J. 2017. Localization of diffusion-based inpainting in digital images. *IEEE transactions on information forensics and security*, 12(12): 3050–3064.

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

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Zhengui, X.; Ma, R.; and Guan, H. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples.

Liu, Y.; Fan, C.; Dai, Y.; Chen, X.; Zhou, P.; and Sun, L. 2024. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24219–24228.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15315–15324. IEEE.

Luo, X.; Jiang, Y.; Wei, F.; Wu, Y.; Xiao, X.; and Ooi, B. C. 2024. Exploring privacy and fairness risks in sharing diffusion models: An adversarial perspective. *IEEE Transactions on Information Forensics and Security*.

Mi, X.; Tang, F.; Yang, Z.; Wang, D.; Cao, J.; Li, P.; and Liu, Y. 2023. Adversarial Robust Memory-Based Continual Learner. *arXiv preprint arXiv:2311.17608*.

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

Ruiz, N.; Bargal, S. A.; and Sclaroff, S. 2020. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 236–251. Springer.

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

Salman, H.; Khaddaj, A.; Leclerc, G.; Ilyas, A.; and Madry, A. 2023. Raising the Cost of Malicious AI-Powered Image Editing. *arXiv preprint arXiv:2302.06588*.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting artists from style mimicry by text-to-image models.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. $P+$: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522*.

Wang, R.; Huang, Z.; Chen, Z.; Liu, L.; Chen, J.; and Wang, L. 2022a. Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 761–767. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wang, X.; Huang, J.; Ma, S.; Nepal, S.; and Xu, C. 2022b. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14920–14929. IEEE.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, H.; Zhou, J.; Tian, J.; Liu, J.; and Qiao, Y. 2022. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 17: 443–456.

Xue, H.; Liang, C.; Wu, X.; and Chen, Y. 2024. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*.

Zhang, J.; Li, B.; Xu, J.; Wu, S.; Ding, S.; Zhang, L.; and Wu, C. 2022. Towards Efficient Data Free Black-Box Adversarial Attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15115–15125. IEEE.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023a. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14071–14081. Vancouver, BC, Canada: IEEE.

Zhang, Y.; Dong, W.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Lee, T.-Y.; Deussen, O.; and Xu, C. 2023b. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Trans. Graph.*, 42(6).

Zhao, R.; Liu, T.; Xiao, J.; Lun, D. P.; and Lam, K.-M. 2021. Invertible image decolorization. *IEEE Transactions on Image Processing*, 30: 6081–6095.

Zhu, Y.; Chen, Y.; Li, X.; Zhang, R.; Tian, X.; Zheng, B.; and Chen, Y. 2023. Information-Containing Adversarial Perturbation for Combating Facial Manipulation Systems. *IEEE Transactions on Information Forensics and Security*, 18: 2046–2059.