MYOPIA: Protecting Face Privacy from Malicious Personalized Text-to-Image Synthesis via Unlearnable Examples

Zhihao Wu[1], **Yushi Cheng**[1,2*], **Tianyang Sun**[2], **Xiaoyu Ji**[1], **Wenyuan Xu**[1]

[1]USSLAB, Zhejiang University
[2]ZJU-UIUC Institute, Zhejiang University
{zhihaowu, yushicheng, xji, wyxu}@zju.edu.cn, tianyang.21@intl.zju.edu.cn

Abstract

Personalized text-to-image synthesis models, such as DreamBooth, have demonstrated significant potential in creating lifelike images tailored to a specific individual by fine-tuning from a limited set of face images and simple prompts. However, if misused, these models could pose a serious risk of privacy infringement by generating harmful images containing violent or pornographic content. To tackle this issue, this paper introduces MYOPIA, a method that renders facial images unlearnable by incorporating error-minimizing perturbations. These meticulously designed perturbations enable the model to quickly overfit to them, resulting in a swift reduction in loss and the cessation of model fine-tuning, effectively preventing the model from capturing genuine facial features. Moreover, to ensure the imperceptibility and robustness of the perturbations, we utilize the Just-Noticeable-Difference and Expectation-of-Transformation techniques to regulate both their location and intensity. Evaluation on two face datasets, i.e., VGGFace2 and CelebA-HQ, with various model versions illustrates the effectiveness of our approach in preserving personal privacy. Furthermore, our method showcases robust transferability across diverse model versions and demonstrates resilience against various image pre-processing techniques.

Code — https://github.com/ZhihaoWu95/myopia

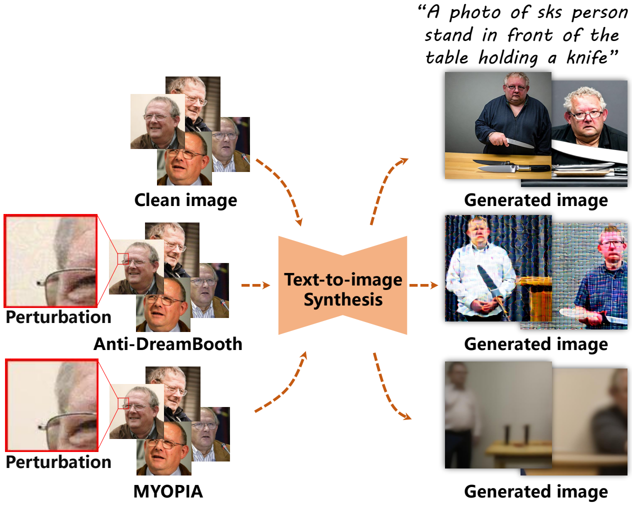

Figure 1: Adversaries can misuse personal images to create harmful content via T2I synthesis. To prevent it, we propose MYOPIA, using unlearnable examples to prevent models from learning genuine facial features, thus protecting facial privacy. Unlike adversarial-based methods, MYOPIA achieves similar protection with less perceptible distortions.

Introduction

Text-to-image (T2I) generation tasks have made remarkable strides in recent years, enabling the creation of realistic images from textual descriptions. This advancement has revolutionized various applications such as content creation, design prototyping, and visual storytelling (Roose 2022; Dhariwal and Nichol 2021). These models empower users to generate high-quality images rich in intricate details that closely mimic real-world visuals (Rombach et al. 2022), often making them indistinguishable from genuine photographs (Ingram, Goode, and Nair 2022). In particular, personalized text-to-image synthesis tools like DreamBooth (Ruiz et al. 2022) provide users with a convenient means to produce images featuring specific individuals. By fine-tuning the T2I model with a limited number of examples, users can effortlessly create diverse images of the target person, ranging from portraits to candid snapshots.

However, this capability poses a double-edged sword. Through these tools, adversaries can exploit a limited set of face images of an individual to generate content containing harmful elements, such as pornography or violent imagery (Qu et al. 2023), as illustrated in Figure 1. To counter this, a straightforward approach, akin to Anti-DreamBooth (Van Le et al. 2023), involves leveraging adversarial examples to disrupt facial representations in the feature space, compelling the model to produce poor-quality or distorted images.

In this paper, inspired by unlearnable examples in image classification tasks (Huang et al. 2021), we ask *"Can we prevent T2I models from capturing genuine facial features from protected face images, thus addressing privacy concerns at their core?"* Unlearnable examples introduce easily learn-

[*]Corresponding Author

able perturbations into original examples and strategically manipulate the model gradient during training to create an "optimization trap". By training the model on these unlearnable examples, it quickly overfits to the perturbation, driving its prediction loss close to zero and converging to a local optimum. Consequently, the target model becomes unable to optimize further and loses the ability to extract genuine features from the original examples. Moreover, these shortcut features make unlearnable examples inherently transferable and effective against black-box systems (Yu et al. 2021), such as uncontrolled T2I models used by adversaries. With these benefits in mind, we delve into expanding the use of unlearnable examples for privacy protection in personalized T2I models.

However, several challenges arise when applying unlearnable examples in diffusion-based models in real-world scenarios: (1) How to efficiently generate the required perturbations from a diffusion model with gradients that vary based on the current timestamp? and (2) How to ensure that the introduced perturbations remain imperceptible to adversaries while being resilient against specific image distortions in uncontrolled conditions?

To address these challenges, we introduce the error-minimizing perturbation generation algorithm to create shortcut features for rapid model overfitting. To enhance the effective exploration of gradient information relevant to facial features and promote gradient stability, we utilize a Gaussian-based timestamp sampling technique. Additionally, we incorporate Just-Noticeable-Difference (JND) to improve the imperceptibility of the added perturbation. Finally, we apply Expectation-of-Transformation (EOT) to enhance resilience against various image processing algorithms. Extensive experiments demonstrate that our methods effectively prevent personalized T2I models from capturing genuine facial features of the target individual, exhibiting strong portability and robustness. In summary, our contributions include follows:

- We explore the potential of leveraging unlearnable examples to safeguard facial privacy in personalized Text-to-Image synthesis, pioneering the adaptation of this concept to diffusion-based generative models.
- We introduce MYOPIA, a solution that effectively preserves facial privacy by introducing imperceptible perturbations to prevent the target model from capturing genuine facial features, rather than disrupting the feature representation.
- Our evaluation on VGGFace2 and CelebA-HQ datasets using different model versions showcases the effectiveness of MYOPIA in protecting personal privacy. Furthermore, MYOPIA demonstrates robust transferability across a range of models and resilience against various image pre-processing techniques.

Related Works

Personalized Text-to-image Synthesis

T2I generation merges computer vision and natural language processing to create images from textual descriptions, which can be categorized into four types: (1) auto-regressive models (Yu et al. 2022), (2) masked-generative-transformer models (Chang et al. 2023), (3) generative adversarial networks (GANs)(Sauer et al. 2023), and (4) diffusion models. Notably, diffusion-model-based approaches like Imagen(Saharia et al. 2022), DALL-E (Ramesh et al. 2022), and Stable Diffusion (Rombach et al. 2021) have garnered significant public interest for their ability to generate high-quality, realistic images by leveraging the CLIP (Radford et al. 2021) technique and extensive datasets like LAION-5B (Schuhmann et al. 2022).

In recent years, personalized T2I synthesis has emerged as a research focus to tailor model outputs for specific individuals or objects, using model fine-tuning (Chen et al. 2023; Gal et al. 2022; Ruiz et al. 2024) or incorporating new concepts directly from image embeddings (Chen et al. 2024; Jia et al. 2023; Wei et al. 2023). Among these methods, DreamBooth, known for its robust capabilities on various public platforms (getimg.ai 2024; toolify.ai 2024), raises misuse concerns due to its accessibility, making it our target.

Adversarial Examples

Adversarial examples, initially designed to deceive machine learning models (Szegedy et al. 2013; Carlini and Wagner 2017), now play a role in data privacy through image cloaking. This approach involves introducing adversarial perturbations to images before their public release to prevent potential misuse in model training or malicious activities. For instance, Fawkes (Shan et al. 2020) employs targeted adversarial attacks to alter a user's identity within the embedding feature space, thwarting privacy violations by unauthorized face recognition systems. Furthermore, LowKey (Cherepanova et al. 2021) enhances transferability to black-box models by utilizing an ensemble of surrogate models.

For generative models, methods such as Anti-forgery (Wang et al. 2022) and UnGANable (Li et al. 2023) have been developed to defeat GAN-based Deepfakes. For diffusion models, initiatives like GLAZE (Shan et al. 2023a), Nightshade (Shan et al. 2023b), and AdvDM (Liang et al. 2023) aim to protect data privacy in personalized T2I diffusion models through image cloaking, focusing on safeguarding against artistic mimicry rather than face privacy concerns. Additionally, Anti-DreamBooth targets the malicious use of personalized T2I models and proposes effective preventive measures, making it a pertinent comparison method for this paper.

Unlearnable Examples

In contrast to adversarial examples, the aim of unlearnable examples is not to deceive the model's predictions but rather to impede the model from capturing specific patterns or features during training. To achieve it, unlearnable examples establish a strong correlation between perturbations and true labels, causing the model to quickly overfit these perturbations instead of learning the authentic data features. Consequently, the model's loss rapidly converges to 0, hindering effective learning. Expanding on this concept, Unlearnable clusters (Zhang et al. 2023) utilize a CLIP surrogate model to improve transferability across various models. Furthermore, these studies (Yu et al. 2021; Ren et al. 2022) emphasize the importance of linear separability in unlearnable examples, enhancing their transferability across diverse datasets. However, these investigations predominantly focus on CNN-based image classification models and do not directly apply to generative models. In this paper, we delve into optimizing unlearnable examples within generative models and propose a series of methods to tailor generative models for this purpose.

Methodology

Background

Unlearnable Examples. The primary strategy to render an example unlearnable involves expediting the target model's loss convergence to 0 during training, effectively stalling further learning. This is achieved by crafting a perturbation that the model can effortlessly learnt, leading it to swiftly converge to a local optimum. The key to generating such perturbations lies in solving the following bi-level optimization problem:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\min_{\delta} \mathcal{L}\left(f_\theta(x+\delta), y\right) \right] \quad \text{s.t. } \|\delta\|_p \leq e \tag{1}$$

where (x, y) is a pair of a example and its true label selected from the training dataset $\mathcal{D}$, $f_\theta(\cdot)$ is the target model with the weights θ, $\mathcal{L}(\cdot)$ is the loss function for training the target model, and the δ is the L_p-norm bounded perturbations.

In this bi-level optimization function, there are two components to be optimized with the same objective, including the perturbation δ and the model weights θ. To effectively optimize them, an alternating optimization approach has been utilized, i.e., optimizing the δ after every M steps of training θ.

For each perturbation optimization step, the Projected Gradient Descent (PGD) (Madry et al. 2017) is commonly used to solve the inner optimization problem and update the perturbation as follows:

$$x'_{k+1} = \Pi_e\left(x'_k - \alpha \cdot \text{sign}\left(\nabla_x \mathcal{L}\left(f\left(x'_k\right), y\right)\right)\right) \tag{2}$$

where the k is the current perturbation optimization step, $\nabla_x \mathcal{L}\left(f\left(x'_k\right), y\right)$ us the gradient of the loss with the current input sample x'_k, Π is used to clip the pixel value to restrain the perturbation in e, and the perturbation of each step can be presented as $\delta = x'_{k+1} - x'_k$.

Diffusion model. In contrast to one-step generative models, such as GAN, the diffusion model performs the following two processes: a forward process and a backward process (Song, Meng, and Ermon 2020). In the forward process, the model adds noise to the input image until it becomes an entire Gaussian noise. In the backward process, the model learns to reverse the forward process and obtain the input-like image from the Gaussian noise. During the forward process, the model adds the increasing level of random noise by the scheduler $\{\beta_t : \beta_t \in (0,1)\}_{t=1}^{T}$, and obtain the noised image sequence $\{x_1, x_2, ..., x_T\}$. For each x_t, the noise level is based on its timestamp t as follows:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha_t}} \epsilon \tag{3}$$

where the $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \Pi_{s=1}^{t} \alpha_s$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

During the backward process, the model learns to denoise the sample x_{t+1} into the previous-step noise-added sample x_t by training a noise prediction model ϵ_θ, where the denoised sample can be described as $x_t = x_{t+1} - \epsilon_\theta(x_{t+1}, t)$. To achieve this denoising process, the noise prediction model is trained by minimizing the ℓ_2 distance between the predicted noise and the truly added noise ϵ for each t step:

$$\mathcal{L}\left(\theta, x_0\right) = \mathbb{E}_{x_0, t, \epsilon \in \mathcal{N}(0,1)} \left\| \epsilon - \epsilon_\theta\left(x_{t+1}, t, c\right) \right\|_2^2 \tag{4}$$

where the t is uniformly samples from $\{1, \ldots, T\}$, c is the prompt that guide the image generation.

DreamBooth. To enhance the personalized capabilities of the T2I model, the DreamBooth method has been introduced as a fine-tuning technique. This approach involves using a specific prompt, such as "a photo of sks [class]" to facilitate the model in learning the distinct features of the designated object sks, with "[class]" indicating the object type. For example, users can fine-tune the model with a prompt like "a photo of sks person", enabling it to generate images featuring the specified sks person. However, training the model directly with such a prompt may cause it to overfit that specific prompt, potentially limiting its ability to generate images of other individuals.

To address it and make the training effective, DreamBooth utilizes a prior preservation loss to make the model learn the generic features of "[class]" while allowing it to learn unique features of the "sks" object only using a small set of its images. Specifically, DreamBooth firstly uses a generic prior prompt "a photo of [class]" to generate a set of images for this class, then fine-tuning the model with both the class examples x' and the sks examples x. Therefore, the loss of DreamBooth can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{db}\left(\theta, x_0\right) = \mathbb{E}_{x_0, t, t'} &\left\| \epsilon - \epsilon_\theta\left(x_{t+1}, t, c\right) \right\|_2^2 \\ &+ \zeta \left\| \epsilon' - \epsilon_\theta\left(x'_{t'+1}, t', c_{pr}\right) \right\|_2^2 \end{aligned} \tag{5}$$

where the c_{pr} is a generic prior prompt, ϵ, ϵ' are both sampled from $\mathcal{N}(0, \mathbf{I})$, t, t' are both sampled from $\{1, \ldots, T\}$, and ζ is the weight to control the importance of prior term.

Problem Definition

While DreamBooth empowers users to create personalized images, there exists a risk of misuse for malicious purposes that could compromise personal privacy, such as generating harmful or sexually explicit content using someone's facial images without authorization. To address it, this paper aims to protect individuals' privacy by preventing T2I models from capturing the unique features of an individual through the introduction of imperceptible unlearnable perturbations. By fine-tuning the model using these unlearnable examples, it becomes incapable of generating images that depict specific individuals. Subsequently, we outline the definition of this problem in the following sections.

Given a set of clean images $\mathcal{X}_c = \{x_c^i\}_{i=1}^N$ for the person to be protected, our goal is to transform it into the unlearnable image set $\mathcal{X}_u = \{x_u^i\}_{i=1}^N$. We donate each unlearnable

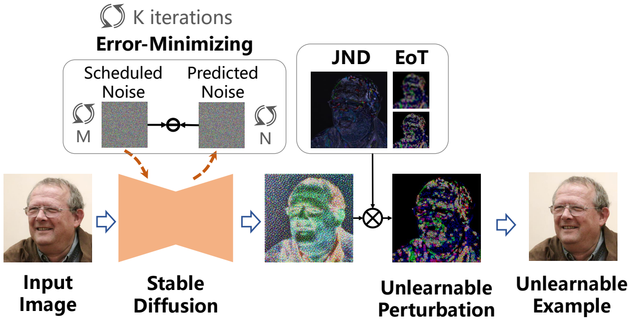

Figure 2: Overview of MYOPIA: The error-minimizing perturbation is generated via alternating optimization, constrained by Just-Noticeable-Difference and Expectation of Transformation, and added to the input image to create the unlearnable example.

image as $x_u^i = x_c^i + \delta^i$, where δ^i is the imperceptible perturbation that bounded by L_p norm, i.e., $\|\delta_i\|_p \leq e$. Therefore, the goal of this paper becomes the optimization of the perturbation set $\Delta = \{\delta_i\}_{i=1}^N$ while minimizing the face similarity between the origin images and the images generated from the unlearnable-example-trained DreamBooth. This problem can be formulated as follows:

$$\begin{aligned} \Delta = \arg\min_{\delta} \mathcal{J}_{x\sim\mathcal{X}_c}\left(f_{\theta^*}(c), x\right) \\ \text{s.t. } \theta^* = \arg\min_{\theta} \mathbb{E}_{x\sim\mathcal{X}_u}\left(\mathcal{L}_{db}(\theta, x)\right) \end{aligned} \tag{6}$$

where the $\mathcal{J}(\cdot)$ is a judgment metric that measures the face similarity between the generated image and the origin image, and the $f_{\theta^*}(c)$ is the image generated by the trained model with any prompt c.

The Proposed Method

Error-Minimizing Perturbation Generation. Our objective is to facilitate the target model in learning a easily learnable perturbation distinct from actual data features. To achieve this, we try to solve the similar bi-level optimization problem as mentioned in Eq. 1. Given the absence of labels in the diffusion model, we aim to minimize the ℓ_2 distance between the predicted noise and the scheduled noise during training for the target entity, i.e., the first part of the DreamBooth loss function. Therefore, the optimization problem can be formulated as follows:

$$\arg\min_{\delta} \mathbb{E}_{x\sim\mathcal{X}_c}\left[\arg\min_{\theta} \mathbb{E}_{x_0,t}\left\|\epsilon - \epsilon_\theta\left(x_{t+1} + \delta, t, c\right)\right\|_2^2\right] \tag{7}$$

where $\|\delta\|_p \leq e$.

Similar to Eq. 1, we solve the above problem by alternating optimization approach. As shown in Figure 2, we first train the diffusion model with clean images, then fix the model weight and optimize the perturbations. By alternating several iterations of this process, we obtain the unlearnable perturbation. A detailed implementation of the algorithm can be found in the Appendix.

Typically, PGD utilizes gradients to update perturbations. However, the gradient of the diffusion model varies across different timestamps. To mitigate this variability, a common approach involves estimating the aggregate gradient direction by uniformly sampling timestamps from the set $\{1, \ldots, T\}$. Nevertheless, our experiments detailed in Appendix reveal that training with larger timestamps significantly enhances the reconstruction of facial details. As a result, to effectively generate unlearnable examples, we opt to optimize perturbations using larger timestamp values. Consequently, we sample timestamps from a Gaussian distribution, as opposed to a uniform distribution, during the perturbation optimization process. The approach is outlined below:

$$t \in |(1 - \mathcal{N}(0, 0.25))|T \quad \text{s.t. } 1 \leq t \leq T \tag{8}$$

where $|\cdot|$ is the absolute value function, and the T is the max value of timestamps.

Just Noticeable Difference. While the perturbations are constrained to be small by the L_p-norm, they can still be perceptible to the human visual system, which may compromise protection performance. To diminish the visual impact of perturbations without compromising their efficacy, we incorporate the Just Noticeable Difference (JND) model (Wu et al. 2017) for constraint. The JND model defines the minimum image distortion perceptible to the human eye and provides a saliency map that indicates the permissible range of variation in pixel values. For each pixel p in the image x, the JND value can be computed as follows:

$$\mathcal{F}_{JND}(p) = \mathcal{L}_A(p) + \mathcal{M}_S(p) - C \cdot \min\{\mathcal{L}_A(p), \mathcal{M}_S(p)\} \tag{9}$$

where C is the gain reduction parameter (set to 0.3 in this paper), and $\mathcal{L}_A(p)$ is a luminance adaptation function that constrains the value of p by the background luminance. $\mathcal{M}_S(p)$ is a spatial masking function of pixel p, which is calculated by pattern masking function $\mathcal{M}_P$ and contrast masking function $\mathcal{M}_C$ as follows:

$$\mathcal{M}_S(p) = \max\{\mathcal{M}_P(p), \mathcal{M}_C(p)\} \tag{10}$$

where $\mathcal{M}_P(p)$ and $\mathcal{M}_C(p)$ donates the pattern complexity and the luminance contrast around the pixel p, respectively. All the definition and calculation of $\mathcal{L}_A$, $\mathcal{M}_P$, and $\mathcal{M}_C$ can be found in Appendix.

While directly applying the JND model to constrain the perturbation can render it nearly imperceptible, it may also significantly diminish the perturbation's effectiveness. To retain perturbations that are both effective and visually acceptable, we employ a hard-clipping technique. This method restricts the perturbation to regions where the JND value is high, ensuring that the perturbation remains influential while aligning with perceptual thresholds. The process is outlined as follows:

$$\hat{\mathcal{F}}_{JND}(p) = \begin{cases} 1 & \text{If } \mathcal{F}_{JND}(p) \geq \gamma \\ \mathcal{F}_{JND}(p) & \text{If } F_{JND}(p) < \gamma \end{cases} \tag{11}$$

where γ is the threshold to control the perturbation added region, which we set 0.25 in this paper.

After calculating the JND value for each pixel, we project the unlearnable perturbations to the space where the human is less susceptible to image changes as follows:

$$\delta_{JND} = \lambda \hat{\mathcal{F}}_{JND}(x)\delta \tag{12}$$

where the λ is a weight parameter that controls the strength of the JND constraint.

Expectation of Transformation. Beyond addressing human perceptibility constraints, image preprocessing methods like Gaussian blur or compression can inadvertently distort the shape of unlearnable perturbations, potentially compromising their efficacy in safeguarding facial privacy. To combat this issue, we leverage the Expectation of Transformation (EoT) (Athalye et al. 2018) technique to bolster the resilience of perturbations during the optimization phase. The EoT approach involves constructing a distribution $\mathcal{T}$ of transformation functions encompassing image preprocessing techniques such as Gaussian blurring and JPEG compression. Subsequently, in each iteration of perturbation optimization, a transformation function $t_f \in \mathcal{T}$ is randomly selected to convert the perturbation into a transformed state as follows:

$$\delta_{EoT} = \arg\min_{\delta} \mathbb{E}_{t_f \sim \mathcal{T}}[\mathcal{L}(x, \delta)] \tag{13}$$

where $\mathcal{L}(x, \delta)$ donates to the error-minimizing optimization function of Eq. 7.

Experiments

Experimental Setting

Datasets. We evaluate MYOPIA on two well-known face datasets VGGFace2 (Cao et al. 2018) and CelebA-HQ (Liu et al. 2015). Similar to Anti-DreamBooth, we select a subset from each dataset comprising images with resolutions exceeding 500×500 and select 50 identities for each dataset. Each identity comprises two subsets: a target-protected image set and a clean image set for reference. Both subsets consist of four images resized to 512×512.

Model Training. For both the clean image set and the protected image set, we train a Stable Diffusion model using the DreamBooth fine-tuning technique on an NVIDIA H800 GPU (80GB). The training parameters includes a learning rate of $5e-7$, a batch size of 2, and 1000 training steps. We choose the Sable Diffusion model version 2.1 as the default model. During training, we follow the same setting as DreamBooth, utilizing the instance prompt "a photo of sks person" and the prior prompt "a photo of person".

Perturbation Generation. We generate the unlearnable perturbation by alternating optimization with default parameters: the iteration K set to 10, the PGD steps N of 100, the step size α of 0.005, the JND weight λ of 0.985, and the perturbation bounded at L_∞-norm 0.5.

Metrics. As our goal is to safeguard facial privacy from malicious T2I synthesis while also ensuring that the introduced protective noise remains undetectable to potential attackers, the unlearnable examples we generate shall meet two key criteria: (1) the generated images should show minimal resemblance to the target individual, and (2) the protected image should closely mirror the original. To achieve these, we evaluate both the quality of the perturbed images and the effectiveness of the generated images.

In evaluating the efficacy of the generated images, we consider both face similarity and image quality. To measure face similarity, we utilize the **Identity Mismatch Score (ISM)** (Van Le et al. 2023), which computes the cosine similarity between the faces in the generated image and the reference clean image using ArcFace(Deng et al. 2019). For assessing image quality, we employ the classical image quality assessment method **BRISQUE** (Mittal, Moorthy, and Bovik 2012), where a higher score indicates lower image quality. To prevent the model from learning the real features of the target person instead of directly disturbing the generated images, we use the **FID** metric (Heusel et al. 2017) to evaluate the realism of the generated image, with a lower score indicating a closer resemblance to real-world images.

For assessing the image quality of perturbed images, we utilize **SSIM** (Wang et al. 2004) and **LPIPS** (Zhang et al. 2018) as metrics. SSIM measures the similarity between two images, with a higher score indicating a closer resemblance. LPIPS considers human perception by using a deep neural network to evaluate perceptual differences between images, where a lower LPIPS score signifies higher perceptual similarity.

Qualitative Results

We first evaluate the effectiveness of MYOPIA on two datasets with two distinct prompts: (1) a face-related prompt "A DSLR portrait of sks person", and (2) a malicious prompt "A photo of sks person stand in front of the table holding a knife". To compare its performance against both an undefended model and the adversarial-attack-based defense method, Anti-DreamBooth, we train the target model for each identity using: (1) clean image sets, (2) adversarial examples generated from Anti-DreamBooth, and (3) unlearnable examples. Subsequently, we generate 12 images for evaluation for each prompt.

Compare to Undefended Model. The results presented in Table 1 indicate that when compared to the undefended model, MYOPIA significantly reduces the ISM metric across different datasets and prompts. This highlights the effectiveness of our approach in thwarting the model from capturing authentic facial features. Furthermore, as depicted in Figure 3 (left), we observe that the images generated by MYOPIA visually show a loss of facial features, particularly noticeable with the malicious prompt.

Compare to Anti-DreamBooth. We further compare MYOPIA with the adversarial-attack-based protection method Anti-DreamBooth. The results in Table 1 demonstrate that MYOPIA achieves a comparable performance to Anti-DreamBooth in terms of ISM and BRISQUE metrics while introducing more imperceptible perturbations, as illustrated in Figure 3 (right). In terms of the quality of the generated images, our method produces images with minimal visual distortions, as evidenced by their lower FID scores. Figure 3 (left) showcases the generated images, highlighting that Anti-DreamBooth leads to highly distorted images, whereas MYOPIA generates more realistic ones. This distinction arises from our approach preventing the model from capturing genuine facial features instead of directly disrupting them, thereby safeguarding the model from interference when generating other objects within the image.

More detailed comparisons with additional prompts and individuals are provided in the Appendix.

Datasets	Method	Quality		Prompt 1			Prompt 2		
		SSIM↑	LPIPS↓	ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
VGGFace2	No Defense	-	-	0.47	31.2	251	0.32	36.4	387
	Anti-DB	0.746	0.139	0.28	49.6	448	0.13	47.8	401
	MYOPIA	0.878	0.059	0.30	59.0	262	0.10	58.9	368
CelebA-HQ	No Defense	-	-	0.41	34.2	221	0.20	37.2	408
	Anti-DB	0.817	0.115	0.33	46.3	338	0.12	41.6	442
	MYOPIA	0.890	0.077	0.32	30.5	214	0.14	35.9	426

- Prompt 1: *A DSLR portrait of sks person*
- Prompt 2: *A photo of sks person stand in front of the table holding a knife*

Table 1: Performance comparison of undefended images, Anti-DreamBooth, and MYOPIA on various datasets and prompts.

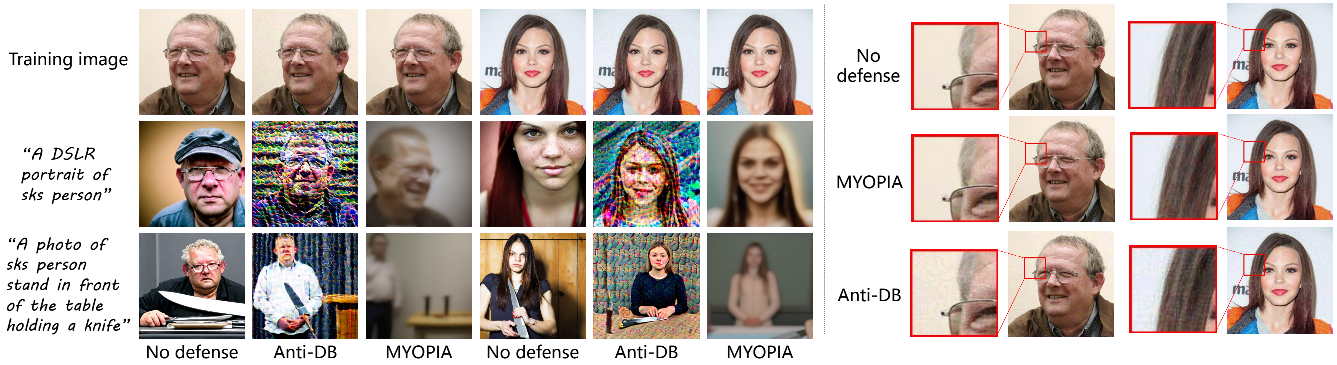
Training image
“A DSLR portrait of sks person”
“A photo of sks person stand in front of the table holding a knife”
No defense
Anti-DB
MYOPIA
No defense
Anti-DB
MYOPIA
No defense
MYOPIA
Anti-DB

Figure 3: Illustration of the images generated by the different methods (left) and the perturbation added (right).

Ablation Studies

We evaluate MYOPIA under the default settings, but the protective performance may vary across different model versions and noise budgets. Hence, we conduct ablation studies focusing on these two aspects.

Model Versions. Since Stable Diffusion is the only open-source T2I model series currently available, we conduct a comprehensive evaluation of MYOPIA using different versions of Stable Diffusion, specifically versions 1.4 and 1.5. These versions differ in both model structure and the training datasets used, providing a robust basis for assessing the generalizability of our approach. The results presented in Table 2 demonstrate that MYOPIA consistently performs well across different model versions.

Noise Budgets. In this paper, we control the perturbation using two key parameters: the L_∞-norm boundary e and the JND weight λ. We assess the effectiveness of MYOPIA across different noise budgets by adjusting these parameters individually. To isolate the impact of each parameter, we keep the other parameter constant to prevent mutual interference. From the findings presented in Table 3, we observe that variations in e beyond 0.1 have minimal influence on the efficacy of MYOPIA, primarily due to the stricter constraint imposed by the JND. Regarding the JND weight, higher values of λ enhance performance but may compromise the invisibility of the perturbation. Nevertheless, even with e set as 0.05 or λ set as 0.975, MYOPIA effectively protects facial privacy, achieving an ISM reduction of 0.06 compared to the unprotected model.

Uncontrolled Conditions

In real-world protection scenarios, adversaries may employ unknown models, partially acquire protected images, or employ image pre-processing methods to fine-tune a T2I model for generating malicious images. To assess the effectiveness of MYOPIA under these uncontrolled conditions, we evaluate its transferability, resilience to various poisoning rates, and image pre-processing techniques.

Transferability. To assess the transferability of MYOPIA across different models, we employ lower versions of Stable Diffusion as surrogate models to produce unlearnable examples, and subsequently test the protection performance on higher versions. The outcomes presented in Table 4 demonstrate MYOPIA maintains consistent efficacy even when surrogate and target models vary. Notably, MYOPIA achieves high transferability without the need for ensemble methods. This success can be attributed to MYOPIA creating an easily learnable perturbation that serves as a training shortcut, enabling it to function effectively across diverse models.

Poisoning Rates. Adversaries may obtain images from diverse sources, including both protected and unprotected images. Therefore, we evaluate MYOPIA across various poisoning rates (i.e., the ratio of protected images to all fine-tuning images). Given that DreamBooth utilizes 4 images for model fine-tuning, we consider three settings where the number of clean images ranges from 1 to 3. From the results shown in Table 5, we find that MYOPIA retains its efficacy when utilizing more than 2 protected images. However, as the rate of clean images rises, the protective performance di-

Version	Method	Prompt 1			Prompt 2		
		ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
V2.1	No Defense	0.47	31.2	251	0.32	36.4	387
	Myopia	0.30	59.0	262	0.10	58.9	368
V1.5	No Defense	0.43	32.9	256	0.31	43.7	376
	Myopia	0.35	42.3	227	0.16	46.2	362
V1.4	No Defense	0.44	32.0	256	0.33	41.0	375
	Myopia	0.32	41.4	234	0.19	43.0	326

Table 2: Protect performance comparison of Myopia on VGGFace2 with different model version.

e	λ	Quality		Prompt 1			Prompt 2		
		SSIM↑	LPIPS↓	ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
0.5	1	0.578	0.431	0.21	66.0	278	0.08	56.1	374
	0.995	0.768	0.148	0.29	62.1	260	0.07	56.3	370
	0.985	0.878	0.059	0.30	59.0	262	0.10	58.9	368
	0.975	0.915	0.037	0.39	57.9	259	0.17	55.1	350
0.25	0.985	0.877	0.059	0.32	59.8	236	0.10	56.1	372
0.10		0.885	0.053	0.34	59.5	257	0.11	56.6	372
0.05		0.928	0.027	0.42	54.9	206	0.21	52.2	371

Table 3: Protect performance comparison of Myopia on VGGFace2 with different perturbation budgets.

Train	Test	Prompt 1			Prompt 2		
		ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
V 1.4	V 1.5	0.34	41.0	230	0.17	45.4	363
V 1.4	V 2.1	0.32	59.3	272	0.09	57.2	370
V 1.5	V 2.1	0.32	60.0	259	0.10	58.2	361

Table 4: Transferability of Myopia across model versions.

Poisoning Rates	Prompt 1			Prompt 2		
	ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
4/4	0.30	59.0	262	0.10	58.9	368
3/4	0.39	51.6	259	0.21	45.7	376
2/4	0.40	41.2	261	0.22	40.9	378
1/4	0.44	34.7	239	0.24	38.1	382

Table 5: Protect performance comparison of Myopia on VGGFace2 with different image poisoning rates.

minishes due to the model's ability to learn genuine facial features from the clean images.

Image Pre-processing. The effectiveness of publicly shared protected images may be compromised by various image processing methods, potentially reducing their protective capabilities. To assess this impact, we evaluate Myopia with two commonly used image pre-processing techniques: Gaussian blur and JPEG compression. For Gaussian blur, we vary the blur kernel size from 3 to 9 to evaluate different levels of blurring, with the outcomes detailed in Table 6. Our findings demonstrate that Myopia maintains a high level of protection performance across these varying blur levels. For JPEG compression, we adjust the quality parameter from 10 to 70. The results in Table 7 reveal that Myopia consistently sustains its performance. This resilience can be attributed to the utilization of our EoT method, which equips Myopia with robustness when confronted with these image pre-processing techniques.

Blur Kernel	Prompt 1			Prompt 2		
	ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
K=3	0.36	48.1	240	0.16	46.2	337
K=5	0.37	40.0	249	0.18	42.5	378
K=7	0.38	39.6	249	0.17	42.4	353
K=9	0.37	39.8	254	0.17	41.8	375

Table 6: Protect performance comparison of Myopia on VGGFace2 with different Gaussian blur kernel size.

JPEG Quality	Prompt 1			Prompt 2		
	ISM↓	BRISQUE↑	FID↓	ISM↓	BRISQUE↑	FID↓
Q=10	0.27	33.1	263	0.15	39.1	388
Q=30	0.28	33.9	261	0.15	38.6	387
Q=50	0.28	33.6	262	0.14	38.9	390
Q=70	0.29	33.5	262	0.14	39.5	393

Table 7: Protect performance comparison of Myopia on VGGFace2 with different JPEG compression quality.

Conclusion

In this paper, we investigate the potential of unlearnable examples for safeguarding facial privacy within personalized T2I models. Building on this concept, we propose Myopia, a method that can prevent the model from capturing the true facial features of the target individual, thus securing the associated images from unauthorized generation. Comprehensive experiments demonstrate that our approach can efficiently preserve facial privacy while minimizing the perceptibility of perturbation, and exhibiting strong transferability and robustness against image pre-processing methods.

Acknowledgments

We thank the anonymous shepherd and reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China (NSFC) Grant 62271280, 62071428, 62222114.

References

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*, 284–293. PMLR.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

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

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*.

Chen, W.; Hu, H.; Li, Y.; Ruiz, N.; Jia, X.; Chang, M.-W.; and Cohen, W. W. 2024. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36.

Cherepanova, V.; Goldblum, M.; Foley, H.; Duan, S.; Dickerson, J.; Taylor, G.; and Goldstein, T. 2021. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

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. *arXiv preprint arXiv:2208.01618*.

getimg.ai. 2024. Transform your images into custom DreamBooth AI models. [EB/OL]. https://getimg.ai/tools/dreambooth.

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. *Advances in neural information processing systems*, 30.

Huang, H.; Ma, X.; Erfani, S. M.; Bailey, J.; and Wang, Y. 2021. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*.

Ingram, D.; Goode, J.; and Nair, A. 2022. You against the machine: Can you spot which image was created by AI. *NBC News, Dec*, 1.

Jia, X.; Zhao, Y.; Chan, K. C.; Li, Y.; Zhang, H.; Gong, B.; Hou, T.; Wang, H.; and Su, Y.-C. 2023. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*.

Li, Z.; Yu, N.; Salem, A.; Backes, M.; Fritz, M.; and Zhang, Y. 2023. {UnGANable}: Defending against {GAN-based} face manipulation. In *32nd USENIX Security Symposium (USENIX Security 23)*, 7213–7230.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ren, J.; Xu, H.; Wan, Y.; Ma, X.; Sun, L.; and Tang, J. 2022. Transferable unlearnable examples. *arXiv preprint arXiv:2210.10114*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

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.

Roose, K. 2022. An AI-Generated Picture Won an Art Prize. Artists Arent Happy.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation.

Ruiz, N.; Li, Y.; Jampani, V.; Wei, W.; Hou, T.; Pritch, Y.; Wadhwa, N.; Rubinstein, M.; and Aberman, K. 2024. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6527–6536.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sauer, A.; Karras, T.; Laine, S.; Geiger, A.; and Aila, T. 2023. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, 30105–30118. PMLR.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023a. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2187–2204.

Shan, S.; Ding, W.; Passananti, J.; Zheng, H.; and Zhao, B. Y. 2023b. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*.

Shan, S.; Wenger, E.; Zhang, J.; Li, H.; Zheng, H.; and Zhao, B. Y. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, 1589–1604.

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*.

toolify.ai. 2024. Create Your DreamBooth with Ease. [EB/OL]. https://www.toolify.ai/stable-video-diffusion/create-your-dreambooth-with-ease-297729.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. 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*, 2116–2127.

Wang, R.; Huang, Z.; Chen, Z.; Liu, L.; Chen, J.; and Wang, L. 2022. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. *arXiv preprint arXiv:2206.00477*.

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.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Wu, J.; Li, L.; Dong, W.; Shi, G.; Lin, W.; and Kuo, C.-C. J. 2017. Enhanced just noticeable difference model for images with pattern complexity. *IEEE Transactions on Image Processing*, 26(6): 2682–2693.

Yu, D.; Zhang, H.; Chen, W.; Yin, J.; and Liu, T.-Y. 2021. Indiscriminate poisoning attacks are shortcuts.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Zhang, J.; Ma, X.; Yi, Q.; Sang, J.; Jiang, Y.-G.; Wang, Y.; and Xu, C. 2023. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3984–3993.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.