Imperceptible Protection Against Style Imitation from Diffusion Models

Namhyuk Ahn[1], Wonhyuk Ahn[1], KiYoon Yoo[2],
Daesik Kim[1], and Seung-Hun Nam[1]

[1] NAVER WEBTOON AI
[2] Seoul National University

Abstract. Recent progress in diffusion models has profoundly enhanced the fidelity of image generation. However, this has raised concerns about copyright infringements. While prior methods have introduced adversarial perturbations to prevent style imitation, most are accompanied by the degradation of artworks' visual quality. Recognizing the importance of maintaining this, we develop a visually improved protection method that preserves its protection capability. To this end, we create a perceptual map to identify areas most sensitive to human eyes. We then adjust the protection intensity guided by an instance-aware refinement. We also integrate a perceptual constraints bank to further improve the imperceptibility. Results show that our method substantially elevates the quality of the protected image without compromising on protection efficacy.

1 Introduction

The groundbreaking advancements in large-scale diffusion models have transformed media creation workflows [2,27,40,43,45]. These can be further enhanced in usability when integrated with external modules that accept multi-modal inputs [18, 21, 28, 38, 62]. Such innovations have also been pivotal in the realm of art creation [1, 22, 50]. Nevertheless, generative AI, while undoubtedly beneficial, brings concerns about its potential misuse. When someone exploits these to replicate artworks without permission, it introduces significant risks of copyright infringement. This *style imitation* becomes a serious threat to artists [49, 60].

To counteract style imitation, previous studies have introduced adversarial perturbation [13,32] to artwork, transforming it into an adversarial example that can resist few-shot generation or personalization methods [29,30,46,52,61,65,66]. Specifically, building on the Stable Diffusion (SD) model [43], they iteratively optimize the protected (or perturbed) image to fool the SD network, guided by the gradients from the image encoder or denoising UNet.

While existing studies are effective in preventing style imitation, they do not prioritize the protected image's quality. They leave discernible traces (or artifacts) on the protected images due to the inherent nature of adversarial perturbations. Moreover, we observed that compared to adversarial attacks on classifiers [7, 31], style protection requires more intense and globally dispersed perturbations. Consequently, despite the commendable protection performance,

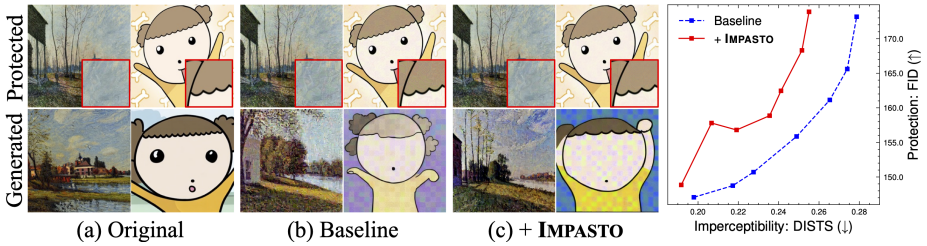


Fig. 1: IMPASTO preserves the authenticity of artworks by impeding style imitation while reducing visual artifacts in the protected images. (**Right**) IMPASTO shows a significant enhancement in the balance between protection efficacy and imperceptibility.

prior works run the risk of severely degrading the original artwork's fidelity, making it less practical for real-world applications. While moderating the strength of protection could mitigate this, it introduces a trade-off, often compromising protection performance and making it challenging to achieve satisfactory results.

To alleviate this, we propose IMperceptible Protection Against STyle imitatiOn (IMPASTO; Fig. 2). We design this upon the principle of *perception-aware protection*, which focuses on perturbing regions less discernible to humans. Although many adversarial attack methods restrict perturbations to small areas to maximize imperceptibility [6,7,37], they are ineffective for style protection since personalization methods can exploit the references from non-perturbed textures. To circumvent this, we instead adopt a *soft restriction* strategy by relaxing the harsh condition of sparse constraints. This applies protection to the entire image but with modulated intensities. To implement this, it is crucial to identify which areas are perceptually more noticeable when perturbations are introduced. For this purpose, we analyze various perceptual maps that are suitable for the style protection task. Then, we propose a method that combines such perceptual maps and refines them in an image-specific manner. Such an *instance-wise refinement* sets itself apart from previous protection methods as it offers a way to strike the optimal balance between imperceptibility and protection performance.

To further enhance imperceptibility, we use a *perceptual constraint bank*. We delve into multiple feature spaces, examining the pixel and the latent spaces of both LPIPS [63] and CLIP [42]. Previous methods have also adopted perceptual constraints [49,61]. However, their approach, typically limited to employing only one or two constraints, does not fully capitalize on the potential of perceptual models. In contrast, we employ a perceptual constraint bank to effectively steer towards more enhanced imperceptibility. Additionally, we integrate a soft restriction within these constraints. Diverging from previous methods that apply constraints uniformly via spatial averaging, we modulate the spatial influence of constraints in areas less perceptible to humans, thereby achieving closer alignment with the nuances of the human visual system. Intuitively, it may not be surprising that imperceptibility can be improved with a constraint bank. However, surprisingly, our work is the first investigation to apply constraints across multiple spaces, thereby enhancing fidelity without sacrificing robustness.

To the best of our knowledge, IMPASTO is the pioneering approach that prioritizes the protected image's quality in the style protection task. We show the

Fig. 2: Model overview. Given an artwork, we construct a perceptual map $\mathcal{M}$ by integrating multiple JNDs, $\mathbf{M}$. IMPASTO optimizes imperceptible perturbation δ using $\mathcal{M}$ and concurrently updates the refinement parameters ω, all while being steered by SD's guidance. Consequently, the protected artwork impedes malicious users from fine-tuning or generating credible imitations. Blue dashed arrows denote the gradient flow.

effectiveness of IMPASTO through a broad range of experiments. It achieves robust protection against style imitation and maintains visual fidelity in protected images (Fig. 1). IMPASTO also significantly improves the trade-off balance between protection efficacy and image quality (Fig. 1, right). The flexibility of IMPASTO is demonstrated by its successful applications to existing protection frameworks [29,30,46,52]. IMPASTO also maintains resilience and generalizes well against a range of countermeasures and personalization techniques, performing on par with the baseline methods. Our key contributions are highlighted as:

- We propose IMPASTO, which applies human visual perception principles to achieve subtle and effective style protection in diffusion models.
- IMPASTO incorporates perception-aware protection and a perceptual constraints bank to realize imperceptible but effective style protection.
- We validate the efficacy of IMPASTO through various experiments, despite its straightforward modulo design. It can be effortlessly adopted in any existing protection method, confirming its ability to be deployed in real applications.

2 Background

2.1 Diffusion Models

Diffusion models have risen to prominence for their capacity to produce high-quality images [2, 9, 16, 39, 45]. In AI-assisted art production, Stable Diffusion (SD) [43] is widely used because of its exceptional quality and efficiency. In the SD model, an input image $\mathbf{x}$ is projected into a latent code via an image encoder $\mathcal{E}$ such that $\mathbf{z} = \mathcal{E}(\mathbf{x})$. A decoder $\mathcal{D}$ reverts the latent code to the image domain, represented as $\mathbf{x}' = \mathcal{D}(\mathbf{z}')$. The diffusion model derives a modified latent code $\mathbf{z}'$ by incorporating external factors y, such as text prompt or other modalities [21, 62]. The training objective for SD at timestep t is defined as:

$$\mathcal{L}_{\mathcal{SD}} = \mathbb{E}_{\mathbf{z}\sim E(\mathbf{x}),y,\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c(y))||_2^2]. \quad (1)$$

Here, a denoising UNet ϵ_θ reconstructs the noised latent code $\mathbf{z}_t$, given t and a conditioning vector $c(y)$. Leveraging the power of diffusion models, recent studies have investigated personalizing SD with a given few reference images.

For example, textual inversion-based methods [1, 12, 53] exploit the embedding space of CLIP [42] while freezing the denoising UNet. On the other hand, model optimization-based methods [24, 44, 50] directly update the UNet.

2.2 Protection Against Style Imitation

Previous protection methods introduce adversarial perturbations δ to image $\mathbf{x}$, making protected image $\hat{\mathbf{x}} = \mathbf{x}+\delta$ through projected gradient descent (PGD) [32], a renowned algorithm in the adversarial attack task.

Encoder-based methods [29,46,49,61] update δ under the guidance of the VAE encoder $\mathcal{E}$. They aim to maximize the distance between the encoded feature of the original image $\mathbf{x}$ and protected image $\hat{\mathbf{x}}$. In practice, many methods instead minimize the distance between the protected image $\hat{\mathbf{x}}$ and target image $\mathbf{y}$:

$$\delta = \underset{||\delta||_\infty \leq \eta}{\arg\min}\, \mathcal{L}_{\mathcal{E}}(\mathbf{x}+\delta, \mathbf{y}),\ \ \mathcal{L}_{\mathcal{E}} = ||\mathcal{E}(\mathbf{x}+\delta) - \mathcal{E}(\mathbf{y})||_2^2. \tag{2}$$

While L_∞ norm ($||\delta||_\infty \leq \eta$; η is a protection budget) is widely used constraints, GLAZE [49] adopts LPIPS [63] and DUAW [61] employs SSIM [55].

UNet-based methods [29, 30, 46, 52, 65] update δ under the guidance of the denoising UNet, ϵ_θ, maximizing diffusion loss, $\mathcal{L}_{\mathcal{SD}}$ as:

$$\delta = \underset{||\delta||_\infty \leq \eta}{\arg\max}\, \mathcal{L}_{\mathcal{SD}}(\mathcal{E}(\mathbf{x}+\delta)). \tag{3}$$

Upon this, Anti-DreamBooth [52] integrates DreamBooth training and Mist [29] merges Eq. 2 and 3, enhancing performance and robustness in various scenarios.

Since Impasto is versatile and can be effortlessly integrated into any existing protection methods, we generalize the style protection as below formulation.

$$\delta = \underset{||\delta||_\infty \leq \eta}{\arg\max}\, \mathcal{L}_{\mathcal{SP}}(\mathbf{x}+\delta, \mathbf{y}), \tag{4}$$

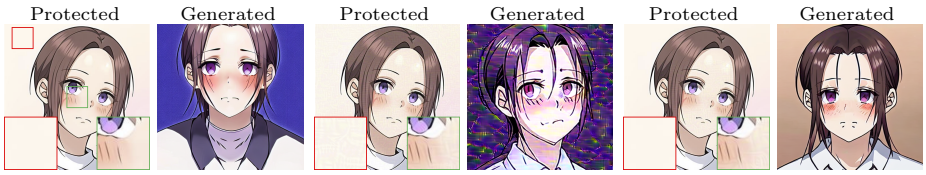
where $\mathcal{L}_{\mathcal{SP}} = -\lambda_{\mathcal{E}}\mathcal{L}_{\mathcal{E}}(\mathbf{x}+\delta, \mathbf{y}) + \lambda_{\mathcal{SD}}\mathcal{L}_{\mathcal{SD}}(\mathcal{E}(\mathbf{x}+\delta))$. Then, we employ PGD [32] to get a protected image $\hat{\mathbf{x}}$. Let $\mathbf{x}^{(0)}$ denote the original artwork. The protected image of i-th optimization step is generated by a signed gradient ascent with step function sgn and step length α as given by:

$$\mathbf{x}^{(i)} = \Pi_{\mathcal{N}_\eta(\mathbf{x})}\Big[\mathbf{x}^{(i-1)} + \alpha \mathrm{sgn}(\nabla_{\mathbf{x}^{(i)}}\mathcal{L}_{\mathcal{SP}}(\mathbf{x}^{(i-1)}, \mathbf{y})\Big], \tag{5}$$

where $\Pi_{\mathcal{N}_\eta(\mathbf{x})}$ is the projection onto the L_∞ neighborhood around $\mathbf{x}$ with radius η. This process is repeated N steps as $\hat{\mathbf{x}} = \mathbf{x}^{(N)}$.

2.3 Imperceptible Adversarial Examples

The concept of imperceptibility is an actively investigated subject in adversarial attacks; some studies target specific elements; *e.g.* the low-frequency components [14,31]. Another approach use advanced constraints [48]; color distance [64]



(a) Original (b) Full protection (c) Partial protection

Fig. 3: Sparse restriction in style protection. We compare the full image protection with partial one that simulates sparse restriction. For partial protection, we only apply perturbation in the facial region (no perturbation in red inlet). Generated result shows that partial protection is inadequate for style protection, as personalization methods can capitalize on unprotected areas when learning the artwork's style.

or quality assessment [54] are adopted. Several methods focus on restricting perturbation regions; leveraging L_0 norm to produce sparse perturbation [6, 37] or limiting perturbations to tiny salient regions [7]. Our research draws inspiration from these studies. However, attacking discriminative models is fundamentally distinct from that of targeting generative models. Hence, we employ a specialized strategy designed specifically for disrupting style imitation.

3 Method— IMPASTO

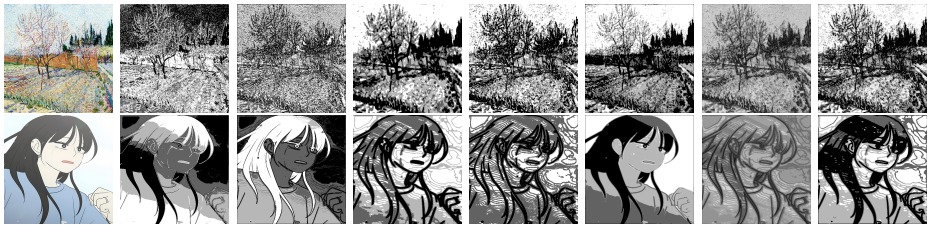
3.1 Perception-Aware Protection (PAP)

Naive approach— Sparse restriction. In adversarial attacks, perturbations are often confined to a sparse region to increase imperceptibility [6,7,37]. Yet, we observed that such restriction does not sufficiently prevent style imitation (Fig. 3). In this analysis, we compare the protection encompassing the entire image (Fig. 3b) against the one that applies perturbation to facial region only (Fig. 3c). It is shown that the partial protection cannot protect original artwork against DreamBooth [44] and we argue that this is because the personalization method can leverage textures from unprotected regions. Overall, sparse restriction may confine the perturbation region too aggressively to be applied in style protection. Therefore, we relax such assumptions to better align with this task.

Soft restriction. To address the limitation of sparse restriction, we instead employ a soft restriction strategy. It protects the entire image but with varying intensities across different regions. To this end, we introduce a perceptual map, $\mathcal{M} \in \mathbb{R}^d$, where d is the number of pixels of an image. This map reflects the human sensitivity to subtle alterations; a value near 1.0 indicates a region with the highest perceptibility, while a value close to 0.0 signifies a region where changes are hardest to notice. With this map $\mathcal{M}$, we define perception-aware protection loss $\mathcal{L}_{\mathcal{PAP}}(\mathbf{x}, \delta, \mathbf{y}, \mathcal{M})$, which use soft restriction-based L_p norm constraint as:

$$\mathcal{L}_{\mathcal{PAP}} = \mathcal{L}_{\mathcal{SP}}(\mathbf{x} + \delta, \mathbf{y}) + \left(\sum_{i=1}^{d} |\mathcal{M}_i * \delta_i|^p\right)^{1/p}, \tag{6}$$

where $*$ denotes element-wise multiplication. Employing a soft restriction-based L_p norm controls perturbations to be suppressed in regions with high percep-



(a) Input (b) LA (c) CM (d) CSF (e) Std. (f) Entropy (g) Avg. (h) IWR

Fig. 4: Examples of perceptual maps. We employ multiple JND models in our analysis. Darker region corresponds to increased protection intensity. The perceptual maps include luminance adaptation (LA), contrast masking (CM), contrast sensitivity function (CSF), standard deviation (Std.), and entropy. IMPASTO constructs perceptual map $\mathcal{M}$ by spatially averaging these estimations (g) or an learnable manner (h).

tual visibility to humans and amplified in areas with lower perceptibility. The subsequent objective is now to quantify perceptual sensitivity to distortions.

Perceptual map analysis. To build a perceptual map $\mathcal{M}$ in a simple yet effective manner, we investigate the just noticeable difference (JND) concept [58] inspired by the human visual system. JND represents the minimum intensity of stimulus (perturbation in our context) required to produce a noticeable change in visual perception. The intent behind the JND estimation model is to determine this perceptual threshold for every image pixel. Given that the fundamental premise of JND— to quantify human sensitivity to subtle changes— aligns with our objective of perception-aware protection, we focus on analyzing 1) the effectiveness of JNDs in our formulation, and 2) which JND models yield the best results. To this end, as depicted in Fig. 4, we compare the following JND models. Detailed explanations for each model are described in Suppl.

- **Luminance adaptation (LA)**: Perturbations are less visible in regions of very low or high luminance and more noticeable in moderate lighting conditions [19]. Hence, we modulate protection strength based on pixel luminance with a fixed adaptation model (Fig. 4b).
- **Contrast masking (CM):** Perturbations can seamlessly blend into regions with intricate textures, while they leave distinct traces on flat surfaces. To simulate this, we utilize the luminance contrast (or change) of a region to measure the complexity of the pixel [26, 57] (Fig. 4c).
- **Contrast sensitivity function (CSF):** Given the band-pass characteristics of the human visual system, we utilize a frequency-based JND model [56]. The human eye is receptive to signals at modulated frequencies while exhibiting insensitivity to high-frequency components. Consequently, perturbations overlaid on high-frequency signals (*e.g.* edges) are less perceptible (Fig. 4d).
- **Standard deviation:** To assess the spatial structure of an image, we calculate the standard deviation of local image blocks, inspired by SSIM [55]. This measures the image's structural complexity, which correlates with the sensitivity to subtle perturbations (Fig. 4e).
- **Entropy:** The entropy of an image block is computed to quantify the amount of information or complexity within a local region [59] (Fig. 4f).

Table 1 presents the protection performance when adopting the JND-based perceptual map. Image quality is assessed through DISTS [10], and protection performance is evaluated using FID [15]. The baseline is trained with Eq. 5 and other models are trained via Eq. 6 with corresponding JNDs. For all the JNDs, we inverse them *i.e.* $M^k = 1 - \text{JND}^k$, since a higher JND threshold represents lower sensitivity to changes. When compared to the baseline, perceptual map improves image quality (DISTS) across the board, but it also leads to a compromise in protection performance (FID). Among the JNDs, LA demonstrates the best protection performance. We conjecture that in many artworks, the majority of areas fall high or low-luminance, thereby maintaining perturbations strength high across extensive regions. However, as illustrated in Fig. 4b, even simple textures like the sky can have strong perturbations, placing LA at the lower image quality. On the other hand, CSF, Std, and Entropy generally apply high perturbations only to specific areas, such as edges, resulting in most regions being not fully protected and consequently, causing a huge degradation in protection performance. CM's protection intensity is also determined by spatial changes but this covers more detailed local regions (see *fields* in Fig. 4c) and also being based on contrast, leading to both high quality and effective protection.

Perceptual map. In our earlier investigation, we observed that the JND model fits surprisingly well with soft restriction approach, offering an effective way for the imperceptible style protection. Consequently, IMPASTO employs a perceptual map constructed using JND estimates, considering its simplicity and effectiveness. Among JNDs we analyzed (Table 1), LA and CM emerge as superior in the quality-protection trade-off. However, these results represent the average scores across a dataset and we note that some images exhibit better trade-off results with different JND models. Especially, since artworks have diverse styles, the best combinations can differ; for example, in Fig. 4, Std. and entropy demonstrate best performance for top image while bottom image shows a preference for CM and CSF. In addition, in practical scenarios, users seeking to protect their artwork are soley concerned with their specific pieces. Hence, it is crucial to ensure that the protection method is effectively applied to every individual artwork.

Table 1: Perceptual map analysis. DISTS measures image quality, while FID evaluates protection performance.

Map	DISTS (↓)	FID (↑)
Baseline	0.212	**299.1**
LA	0.175	<u>284.7</u>
CM	**0.169**	280.8
CSF	0.172	273.2
Std.	<u>0.171</u>	277.7
Entropy	0.174	277.3

Therefore, we apply all the JNDs listed in Table 1 simultaneously to create the perceptual map, since relying on a single JND could result in decreased performance for some artworks. Formally, for an artwork $\mathbf{x}$ requiring protection, we initially generate a corresponding perceptual map $\mathcal{M}$ with a collection of JNDs, $\mathbf{M} = \{M^i, \ldots, M^K\}$, where K is the number of JNDs. To integrate multiple JNDs, the simplest method involves using a spatial average: $\mathcal{M} = \frac{1}{K}\sum_{k=1}^{K} M^k$. As we will demonstrate in the experimental results, this straightforward approach proves to be a surprisingly effective universal algorithm for creating perceptual map. Considering its simplicity and efficacy, we adopt spatially averaged JNDs to construct the initial perceptual map.

Algorithm 1: Optimization of IMPASTO

Data: Image $\mathbf{x}$, target image $\mathbf{y}$, all perceptual maps $\mathbf{M}$, step length α
Result: Protected image $\hat{\mathbf{x}}$

1 Initialize $\mathbf{x}^{(0)} \leftarrow \mathbf{x}$
2 Initialize $\mathcal{M} \leftarrow \frac{1}{K}\sum_{k=1}^{K} M^k, \;\; \omega \leftarrow \frac{1}{K} \cdot \mathbf{1}^K$
3 **for** $i = 1$ **to** N **do**
4 $\quad \delta^{(i)} \leftarrow \alpha \text{sgn}(\nabla_{\delta^{(i)}} \mathcal{L}(\mathbf{x}^{(i-1)}, \delta^{(i-1)}, \mathbf{y}, \mathcal{M}))$
5 $\quad \mathbf{x}^{(i)} \leftarrow \Pi_{\mathcal{N}_\eta(\mathbf{x})}(\mathbf{x}^{(i-1)} + \delta^{(i)})$
6 **end**
7 $\mathcal{M}' \leftarrow \mathcal{M}$
8 **for** $j = 1$ **to** P **do** // instance-wise refinement
9 $\quad \omega \leftarrow \omega - \nabla_\omega \mathcal{L}_\mathcal{M}(\mathbf{x}^{(N)}, \delta^{(N)}, \mathbf{y}, \mathcal{M}', \mathcal{M}(\omega))$
10 **end**
11 $\hat{\mathbf{x}} \leftarrow \mathbf{x}^{(0)} + \mathcal{M}(\omega) \odot \delta^{(N)}$

Instance-wise refinement. Although the above method is more effective in many scenarios than the single JND, applying a uniform averaged map across *all artworks* may still be suboptimal for some artworks. Moreover, the optimal perceptual map corresponding to an artwork varies depending on both artwork's structure and the applied protective perturbations. Even for identical artwork, the detectability of perturbations is affected by changing these as human sensitivity varies with distortion type [11]. To this end, we propose an instance-wise refinement (IWR), customizing the perceptual map $\mathcal{M}$ for each artwork. During optimization, the perceptual map is refined through a weighted sum: $\mathcal{M}(\omega) = \sum_{k=1}^{K} \text{softmax}(\omega)^k * M^k$, where ω is a set of learnable parameters that adjust the contributions of each JND to $\mathcal{M}$. The refinement parameters ω is optimized using the objective function below:

$$\mathcal{L}_\mathcal{M} = ||\mathcal{L}_{\mathcal{SP}}(\mathbf{x} + \mathcal{M}' \odot \delta) - \mathcal{L}_{\mathcal{SP}}(\mathbf{x} + \mathcal{M}(\omega) \odot \delta)||_2^2 + \left(\sum_{i=1}^{d} |\mathcal{M}(\omega)_i * \delta_i|^p\right)^{\frac{1}{p}} \quad (7)$$

with $\mathcal{M}'$ being the initial perceptual map before the refinement steps. In Eq. 7, the former term enforces the consistency of the refined perceptual map $\mathcal{M}$ with the initial map $\mathcal{M}'$ by minimizing the discrepancy in protection loss between them. The latter term enhances the perception-aware protection for a given specific image. Algorithm 1 (L7-10) details the procedure for this refinement. As demonstrated in Fig. 4h, instance-wise refinement has a pronounced effect as compared to the naive averaging approach (Fig. 4g), better capturing the nuances of image-specific perceptual sensitivity.

3.2 Perceptual Constraint Bank

To further enhance the imperceptibility, we employ a bank of perceptual constraints across multiple feature spaces:

Masked LPIPS. LPIPS [63] is a widely used constraint and is also utilized in GLAZE [49]. Our approach distinguishes itself by applying a *masked* LPIPS

constraint, modulating the LPIPS influence using a perceptual map $\mathcal{M}$. Let ϕ_l be a l-th layer of the LPIPS network, with a corresponding feature map resolution d_l, the masked LPIPS is calculated as:

$$\mathcal{L}_{\mathcal{L}} = \sum_l \frac{1}{d_l} \sum_{i=1}^{d_l} \mathcal{M}_i * ||w_l * (\phi_l(\mathbf{x})_i - \phi_l(\mathbf{x}+\delta)_i)||_2^2, \tag{8}$$

where w_l is the channel-wise scale parameters. By focusing on perceptually significant regions with a mask, IMPASTO can achieve better protection performance.

Masked low-pass. In line with our motivation for perceptual protection, we apply a pixel-domain constraint that focuses solely on the low-frequency components. Inspired by Luo et al. [31], we implement a discrete wavelet transform (DWT) to enforce a low-pass filter-based constraint. To tailor this constraint further closely to human perception, we adjust the loss impact using a perceptual map. The associated loss function is formulated as in below.

$$\mathcal{L}_{\mathcal{LP}} = \frac{1}{d} \sum_{i=1}^{d} \mathcal{M}_i * ||\mathrm{LP}(\mathbf{x})_i - \mathrm{LP}(\mathbf{x}+\delta)_i||_2^2, \tag{9}$$

where $\mathrm{LP}(\mathbf{x})$ is the reconstructed image from the low-frequency component only. For more details, please refer to Suppl. This constraint mimics observing a painting from a distance, where perturbations in smooth regions are perceptible, while detailed textures hide these until we closely inspect the artwork. It reflects practical viewing where visibility depends on spatial detail and observer distance.

CLIP. We also leverage the CLIP [42] space which benefits from training on a vast and varied dataset of image-text pairs. This extensive training enables CLIP to evaluate image quality independently of the original image, offering a novel approach to quality assessment within perceptual constraints. With the prompt C = "Noise-free image", the CLIP constraint aims to maximize the feature distance between the protected image and the descriptive prompt.

$$\mathcal{L}_{\mathcal{C}} = -cos(\mathrm{CLIP}_I(\mathbf{x}+\delta),\ \mathrm{CLIP}_T(C)), \tag{10}$$

where $\mathrm{CLIP}_I, \mathrm{CLIP}_T$ are image and text encoders. The final protection loss, $\mathcal{L}(\mathbf{x}^{(i)}, \delta, \mathbf{y}, \mathcal{M})$ combines all the losses, weighted by their respective λs as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{PAP}} + \lambda_L \mathcal{L}_{\mathcal{L}} + \lambda_{LP} \mathcal{L}_{\mathcal{LP}} + \lambda_C \mathcal{L}_{\mathcal{C}}. \tag{11}$$

Algorithm 1 overviews the protection process of IMPASTO. It is designed to be versatile, allowing the integration of existing protection frameworks. Such adaptability will be discussed in the next section as well as in Suppl.

4 Experiment

Implementation details. We optimize for $N = 100$ steps and IWR updates $\mathcal{M}$ for $P = 25$ steps after perturbation optimization. Other settings are aligned with that of baseline models; please refer to Suppl. for more details.

Table 2: Quantitative comparison of protection methods w/ and w/o IMPASTO, both selected for their comparable protection performance. IMPASTO markedly elevates the protected images' quality while maintaining comparable levels of protection efficacy.

Dataset	Method	Protected Image Quality			Protection Performance		
		DISTS (↓)	PieAPP (↓)	TOPIQ (↑)	NIQE (↑)	BRISQUE (↑)	FID (↑)
Painting	PhotoGuard	0.181 (+0.000)	0.364 (+0.000)	0.896 (+0.000)	4.306	20.99	277.6
	+ IMPASTO	0.159 (+0.022)	0.315 (+0.049)	0.912 (+0.016)	4.479	20.74	279.3
	AdvDM	0.167 (+0.000)	0.730 (+0.000)	0.846 (+0.000)	3.761	12.45	269.0
	+ IMPASTO	0.136 (+0.031)	0.531 (+0.199)	0.895 (+0.049)	3.897	12.54	271.6
Cartoon	PhotoGuard	0.249 (+0.000)	0.782 (+0.000)	0.797 (+0.000)	5.037	10.19	155.9
	+ IMPASTO	0.207 (+0.042)	0.709 (+0.073)	0.886 (+0.089)	5.632	10.90	157.8
	AdvDM	0.241 (+0.000)	0.776 (+0.000)	0.775 (+0.000)	4.802	10.95	153.5
	+ IMPASTO	0.231 (+0.010)	0.774 (+0.002)	0.797 (+0.022)	4.793	11.87	154.0

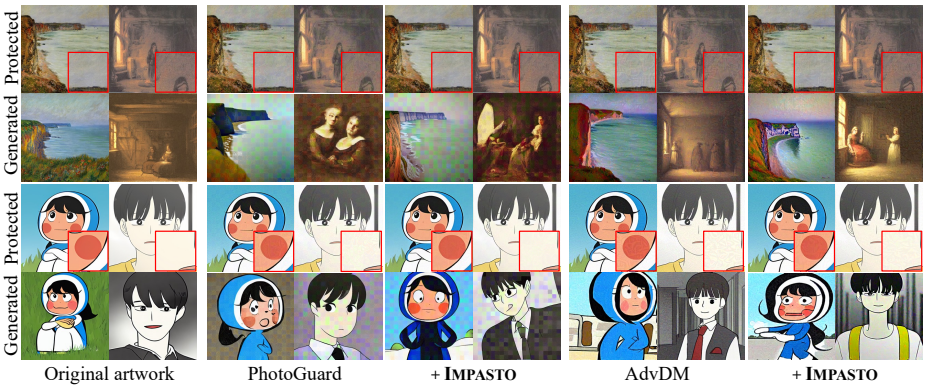
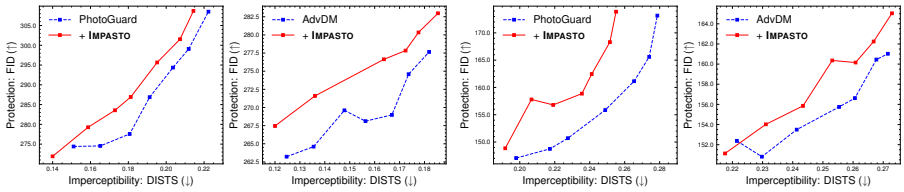


Fig. 5: Qualitative comparison of baseline methods with and without IMPASTO. While maintaining comparable style protection efficacy (artifacts in the generated results), IMPASTO significantly enhances the protected images' quality.

Datasets. We utilize two art domain datasets: painting and cartoon. The painting dataset is curated from WikiArt [51] with a selection of 15 artists, 10 works per artist. The cartoon dataset is a collection of 15 cartoons with 10 cartoon face images. Further details are in Suppl.

Evaluation. To evaluate protected image quality, we use DISTS [10], PieAPP [41], and TOPIQ [4]. Protection performance is measured with NIQE [36], BRISQUE [35], and FID [15]. In protection metrics, worse scores indicate more effective protection, as our objective is to prevent style mimicry. It is important to note that some protection assessments, such as NIQE and BRISQUE, are non-reference-based, leading to somewhat inconsistent scores. These tend to fluctuate instead of showing a consistent progression with varying protection strengths. FID, although potentially inconsistent due to a limited number of evaluation samples, aligns more closely with human preferences in this task. Therefore, we further validated IMPASTO's effectiveness through human evaluation. For more detailed settings on evaluation metrics, including the user study, please refer to Suppl.

Baseline. IMPASTO can be integrated into any methods generally formulated in Eq. 4. Based on this, in our benchmark, we incorporate IMPASTO into Photo-



(a) Painting dataset (b) Cartoon dataset

Fig. 6: Style protection comparison. Protection performance is evaluated with FID [15] and imperceptibility via DISTS [10]. Adaptation of IMPASTO to both PhotoGuard [46] and AdvDM [30] ensures superior performance.

Guard [46] (encoder-based) and AdvDM [30] (UNet-based). In Suppl., applications to Mist [29] and Anti-DreamBooth [52] are also presented.

4.1 Model Comparison

Table 2 presents a quantitative comparison of visual quality under a comparable protection performance of the models with and without IMPASTO. Across all the scenarios, IMPASTO substantially enhances the fidelity of the protected images. Fig. 5 also supports the superior efficacy of IMPASTO; it successfully minimizes artifacts, in contrast to baselines that leave discernible traces on the artwork. It is particularly pronounced in the cartoon dataset (bottom), where both vanilla methods introduce noticeable artifacts in facial areas, potentially disrupting user immersion when reading cartoons. In contrast, IMPASTO reduces artifacts significantly, rendering them nearly invisible unless examined closely and meticulously. User evaluation further confirms the effectiveness of IMPASTO (Fig. 7).

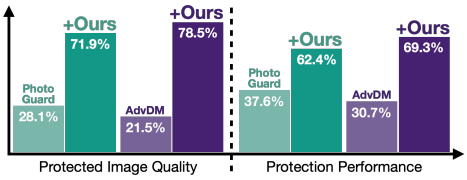


Fig. 7: User preference study (via A/B test) of PhotoGuard [46] and AdvDM [30] methods with and without IMPASTO.

Varying protection strengths. In Fig. 6, we manipulate the protection strengths (budgets) to delineate an imperceptibility-protection trade-off curve. On both painting and cartoon datasets, employing IMPASTO into the baselines (PhotoGuard and AdvDM) considerably improves the trade-off dynamics; IMPASTO consistently achieves higher imperceptibility with comparable protection performance, or delivers enhanced protection without compromising image quality.

4.2 Model Analysis

Ablation study. In Table 3, we dissect components of IMPASTO. The proposed PAP markedly improves image fidelity over the base model (Base*), albeit with a slight reduction in protection efficacy. However, compared to the base model with lower protection (Base**), PAP achieves superior preservation of protection performance with enhanced image quality. We observed that when $\mathcal{M}$ is formed from a random mask, not JND, (w/o JND), the protection performance is degraded.

Table 3: Component analysis. We incrementally attach proposed components to assess their contribution. PAP: perception-aware protection. w/o JND: initialize $\mathbf{M}$ with random masks, not from JNDs. w/o Mask: Constraints without a perceptual map $\mathcal{M}$. Base*: the baseline model (PhotoGuard [46]) with equal protection strength. Base**: the baseline model with a similar level of protection performance.

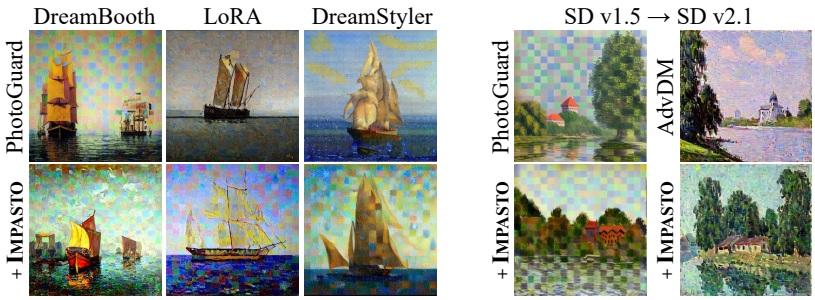
Method	Image Quality		Protection Performance	
	DISTS (↓)	TOPIQ (↑)	BRISQUE (↑)	FID (↑)
Base*	0.212	0.845	17.66	**299.1**
+ PAP	0.171	0.890	20.05	<u>286.8</u>
w/o JND	0.170	0.892	20.45	278.1
+ LPIPS	<u>0.163</u>	0.910	<u>21.24</u>	280.4
+ Low-pass	<u>0.163</u>	<u>0.911</u>	**21.25**	277.5
w/o Mask	<u>0.163</u>	<u>0.911</u>	19.74	272.6
+ CLIP	**0.159**	**0.912**	20.74	279.2
Base**	0.181	0.896	20.99	277.6

Table 4: Perception-aware protection. The efficacy of individual JNDs (LA and CM since they show superior results) is presented along with an averaged perceptual map and with IWR. IWR†: IWR is conducted before perturbation optimization.

Method	Image Quality		Protection Performance	
	DISTS (↓)	TOPIQ (↑)	BRISQUE (↑)	FID (↑)
Baseline	0.212	0.845	17.66	**299.1**
LA	0.175	0.879	18.29	284.7
CM	**0.169**	0.882	17.44	280.8
Average	<u>0.170</u>	<u>0.891</u>	17.42	277.9
IWR†	0.171	**0.894**	<u>19.26</u>	282.5
IWR	0.171	0.890	**20.05**	<u>286.8</u>

One might expect that the JND maintains protection scores while improving image quality. Indeed, the random mask constrains the protection intensity in a similar level to the JND-based map. This is because both maps are normalized between 0 and 1. As a result, they yield similar perturbation magnitudes, leading to a comparable image quality. Nonetheless, the JND-based perceptual map prioritizes less sensitive regions for perturbation while reducing it in highly sensitive areas. Despite the similar protection magnitudes, the JND-based PAP achieves more effective protection. This hints at the possibility that the JND maps are helpful in finding the areas that are important for style protection. For instance, in areas with complex textures, the PAP applies more perturbations than its non-JND counterpart as shown in Fig. 4e. However, these perturbations are typically imperceptible to humans, thereby ensuring perceptually acceptable image quality. Employing a constraint bank significantly improves image quality with comparable protection scores to the baseline (Base**). The omission of the perceptual map $\mathcal{M}$ in the constraints (w/o Mask) leads to a decline in protection performance, akin to the observations in the PAP case; as we confine the influence of constraints in a perceptual manner, it improves protection efficacy.

Perceptual map. We also analyze the proposed perception-aware protection in Table 4. LA and CM, which are the best JNDs as demonstrated in Table 1, enhance the fidelity of the protected images but at the cost of significantly



(a) Against personalization methods (b) Against diffusion models

Fig. 8: Generalization. IMPASTO does not impede baselines' generalization abilities on **(a)** diverse personalization methods; DreamBooth [44], LoRA [17], and DreamStyler [1] or **(b)** model black-box scenario; optimize on SD v1.5 and test on SD v2.1.

compromising protection performance. Creating a perceptual map with multiple JNDs through averaging leads to better image quality compared to scenarios with a single JND. However, this approach also substantially reduces the effectiveness of protection. We speculate that such a straightforward averaging method may not adequately capture the unique structural elements of the image, resulting in overly smooth perturbations that could weaken the protection performance. On the other hand, IWR enhances all protection scores while preserving satisfactory image quality, as it can adapt to the specific textures and structures of a given artwork. It's noteworthy that applying IWR prior to perturbation optimization (IWR†), where it does not consider the perturbations, slightly diminishes protection performance, accentuating the importance of modeling the interplay between artwork and applied perturbation to finalization mask $\mathcal{M}$.

Countermeasures. To analyze the robustness of IMPASTO, we conduct countermeasure experiments involving JPEG compression ($q = 40$), Gaussian blur (3×3 kernel, $\sigma = 0.02$), and Gaussian noise ($\sigma = 0.02$). Results indicate a performance degradation of all protection methods when these countermeasures are applied, as they tend to remove the protective perturbations (Fig. 9). Nonetheless, IMPASTO demonstrates comparable robustness against such countermeasures. For the blur effect, there is a slight performance degradation with IMPASTO, as subtle perturbations are particularly weakened to this; we also observed that baseline with low-budget protection is also vulnerable to blur operation. However, the blurring effect also renders the artwork less plausible, making it an impractical choice for malicious users. The most commonly used method is likely JPEG compression, as it preserves the artwork's fidelity while being readily applicable.

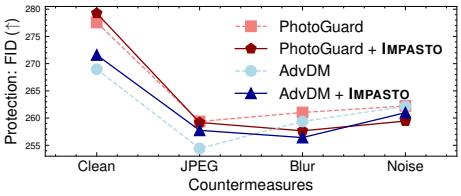


Fig. 9: Evaluation on robustness. Methods with IMPASTO exhibit comparable protection performance to baselines.

Black-box scenario. We extend to evaluating the impact of IMPASTO on the performance of protection when trained with other personalization methods.

Table 5: Quantitative comparison on facial datasets.

Dataset	Method	Protected Image Quality			Protection Performance		
		DISTS (↓)	PieAPP (↓)	TOPIQ (↑)	NIQE (↑)	BRISQUE (↑)	FID (↑)
CelebA-HQ	PhotoGuard	0.280 (+0.000)	0.379 (+0.000)	0.878 (+0.000)	5.031	15.36	233.1
	+ Impasto	0.266 (+0.014)	0.376 (+0.003)	0.880 (+0.002)	5.161	14.13	234.6
	AdvDM	0.213 (+0.000)	0.573 (+0.000)	0.832(+0.000)	4.051	10.05	278.8
	+ Impasto	0.195 (+0.018)	0.522 (+0.051)	0.854(+0.022)	4.080	10.59	273.6
VGGFace2	PhotoGuard	0.270 (+0.000)	0.413 (+0.000)	0.870 (+0.000)	5.916	19.73	253.9
	+ Impasto	0.255 (+0.015)	0.413 (+0.000)	0.868 (-0.002)	5.746	19.39	257.0
	AdvDM	0.206 (+0.000)	0.589 (+0.000)	0.821 (+0.000)	3.977	8.43	292.9
	+ Impasto	0.187 (+0.019)	0.543 (+0.043)	0.846 (+0.025)	3.986	9.86	289.8

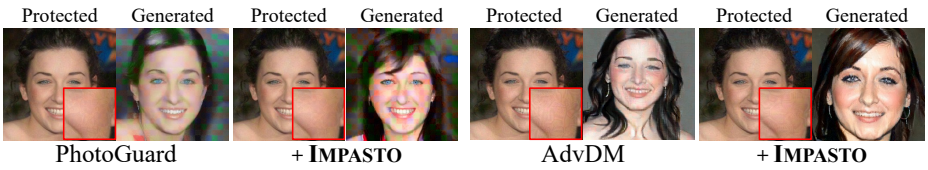


Fig. 10: Qualitative comparison on facial dataset.

We adopt LoRA [17], a standard personalization technique in the art creation community, and DreamStyler [1], known for its effectiveness in style adaptation using a textual inversion approach. As illustrated in Fig. 8a, Impasto maintains robust protection effectiveness even with these personalization methods.

In style protection, generalization robustness against unknown models is also a crucial aspect. To examine this, we compare by optimizing on SD v1.5 and testing on SD v2.1, following the setups of *Van et al.* [52]. As illustrated in Fig. 8b, Impasto successfully preserves the robustness of PhotoGuard and AdvDM. Overall, in both facets of the black-box scenario; unknown personalization methods and diffusion models, Impasto not only maintains the generalization ability but also effectively reduces the artifact on the protected images.

4.3 Impasto in Other Domain

Although Impasto is initially proposed to prevent style imitation, its applicability can be extended beyond other domains or applications. To validate this, we conduct protection on two facial datasets, CelebA-HQ [20] and VGGFace2 [3]. As demonstrated in Table 5 and Fig. 10, adopting Impasto in these natural domains can also enhance the quality of protected images without compromising the protection performance of baseline models. This broad applicability highlights Impasto's versatility and potential as a universal tool for protecting users' copyright and preventing serious threats of deepfakes.

5 Conclusion

We have introduced Impasto to prevent style imitation in perceptual orientation. With a proposed perceptual map, Impasto markedly improves the quality

of the protected images. A perceptual constraint bank further boosts performance, establishing our method as a versatile and superior protector of artwork.

Limitations. Current protection methods mostly adopt adversarial perturbations, which can hamper usability due to the extensive time required for the optimization. Even accessible software [49] takes 30-60 mins to protect a 512^2 image on an M1 Max CPU. Addressing this time constraint is a challenge that future research should aim to overcome.

References

1. Ahn, N., Lee, J., Lee, C., Kim, K., Kim, D., Nam, S.H., Hong, K.: Dreamstyler: Paint by style inversion with text-to-image diffusion models. arXiv preprint arXiv:2309.06933 (2023)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
4. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. arXiv preprint arXiv:2308.03060 (2023)
5. Chou, C.H., Li, Y.C.: A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. IEEE Transactions on circuits and systems for video technology **5**(6), 467–476 (1995)
6. Croce, F., Hein, M.: Sparse and imperceivable adversarial attacks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4724–4732 (2019)
7. Dai, Z., Liu, S., Li, Q., Tang, K.: Saliency attack: Towards imperceptible black-box adversarial attack. ACM Transactions on Intelligent Systems and Technology **14**(3), 1–20 (2023)
8. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. IEEE transactions on image processing **9**(4), 636–650 (2000)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
11. Dodge, S., Karam, L.: A study and comparison of human and deep learning recognition performance under visual distortions. In: 2017 26th international conference on computer communication and networks (ICCCN). pp. 1–7. IEEE (2017)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Guo, C., Frank, J.S., Weinberger, K.Q.: Low frequency adversarial perturbation. arXiv preprint arXiv:1809.08758 (2018)

15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
18. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
19. Jarsky, T., Cembrowski, M., Logan, S.M., Kath, W.L., Riecke, H., Demb, J.B., Singer, J.H.: A synaptic mechanism for retinal adaptation to luminance and contrast. Journal of Neuroscience **31**(30), 11003–11015 (2011)
20. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
21. Kim, S., Lee, J., Hong, K., Kim, D., Ahn, N.: Diffblender: Scalable and composable multimodal text-to-image diffusion models. arXiv preprint arXiv:2305.15194 (2023)
22. Ko, H.K., Park, G., Jeon, H., Jo, J., Kim, J., Seo, J.: Large-scale text-to-image generation models for visual artists' creative works. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. pp. 919–933 (2023)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
25. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging **19**(1), 011006–011006 (2010)
26. Legge, G.E., Foley, J.M.: Contrast masking in human vision. Josa **70**(12), 1458–1471 (1980)
27. Li, W., Xu, X., Xiao, X., Liu, J., Yang, H., Li, G., Wang, Z., Feng, Z., She, Q., Lyu, Y., et al.: Upainting: Unified text-to-image diffusion generation with cross-modal guidance. arXiv preprint arXiv:2210.16031 (2022)
28. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
29. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models. arXiv preprint arXiv:2305.12683 (2023)
30. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Zhengui, X., Ma, R., Guan, H.: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples (2023)
31. Luo, C., Lin, Q., Xie, W., Wu, B., Xie, J., Shen, L.: Frequency-driven imperceptible adversarial attack on semantic similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15315–15324 (2022)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
33. Mannos, J., Sakrison, D.: The effects of a visual fidelity criterion of the encoding of images. IEEE transactions on Information Theory **20**(4), 525–536 (1974)

34. Mitsa, T., Varkur, K.L.: Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. In: 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 5, pp. 301–304. IEEE (1993)
35. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR). pp. 723–727. IEEE (2011)
36. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
37. Modas, A., Moosavi-Dezfooli, S.M., Frossard, P.: Sparsefool: a few pixels make a big difference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9087–9096 (2019)
38. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
39. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
41. Prashnani, E., Cai, H., Mostofi, Y., Sen, P.: Pieapp: Perceptual image-error assessment through pairwise preference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1808–1817 (2018)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
45. 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.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
46. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing. arXiv preprint arXiv:2302.06588 (2023)
47. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural computation **12**(5), 1207–1245 (2000)
48. Sen, A., Zhu, X., Marshall, L., Nowak, R.: Should adversarial attacks use pixel p-norm? arXiv preprint arXiv:1906.02439 (2019)
49. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models. arXiv preprint arXiv:2302.04222 (2023)
50. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)

51. Tan, W.R., Chan, C.S., Aguirre, H.E., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing **28**(1), 394–409 (2018)
52. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2116–2127 (2023)
53. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
54. Wang, Y., Wu, S., Jiang, W., Hao, S., Tan, Y.a., Zhang, Q.: Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. arXiv preprint arXiv:2107.01396 (2021)
55. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
56. Wei, Z., Ngan, K.N.: Spatio-temporal just noticeable distortion profile for grey scale image/video in dct domain. IEEE Transactions on Circuits and Systems for Video Technology **19**(3), 337–346 (2009)
57. Wu, J., Lin, W., Shi, G., Wang, X., Li, F.: Pattern masking estimation in image with structural uncertainty. IEEE Transactions on Image Processing **22**(12), 4892–4904 (2013)
58. Wu, J., Shi, G., Lin, W.: Survey of visual just noticeable difference estimation. Frontiers of Computer Science **13**, 4–15 (2019)
59. Wu, J., Shi, G., Lin, W., Liu, A., Qi, F.: Just noticeable difference estimation for images with free-energy principle. IEEE Transactions on Multimedia **15**(7), 1705–1710 (2013)
60. Xiang, C.: Artists are revolting against ai art on artstation. Vice `https://www.vice.com/en/article/ake9me/artists-are-revolt-against-ai-art-on-artstation`
61. Ye, X., Huang, H., An, J., Wang, Y.: Duaw: Data-free universal adversarial watermark against stable diffusion customization. arXiv preprint arXiv:2308.09889 (2023)
62. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
64. Zhao, Z., Liu, Z., Larson, M.: Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1039–1048 (2020)
65. Zhao, Z., Duan, J., Hu, X., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., Chen, Y.: Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. arXiv preprint arXiv:2306.01902 (2023)
66. Zheng, B., Liang, C., Wu, X., Liu, Y.: Understanding and improving adversarial attacks on latent diffusion model. arXiv preprint arXiv:2310.04687 (2023)

A Method Details

In this section, we describe the details of IMPASTO— the computation and discussion of JND estimations (Sec. A.1), perceptual constraints (Sec. A.2), and the implementation specifics (Sec. A.3).

A.1 JND Estimations

In our study, we investigated five JND models, and here we we provide a comprehensive elucidation of the fundamental rationale of these estimations and delve into the specifics of their computations:

Luminance adaptation (LA): The visibility thresholds within the human visual system vary to luminance levels [19]. For instance, our eye is more sensitive under moderate lighting conditions, while discrimination is challenging in a very dark light. Similarly, we observed that perturbations are less noticeable in regions with extremely low or high luminance, prompting an increase in protection strength within these areas. To implement this concept, luminance adaptation is computed following Chou and Li [5] as in the below equation.

$$LA(\mathbf{x}) = \begin{cases} 17 \times (1 - \sqrt{\frac{B(\mathbf{x})}{127}}) + 3, & \text{if } B(\mathbf{x}) \leq 127 \\ \frac{3}{128} \times (B(\mathbf{x}) - 127) + 3, & \text{otherwise,} \end{cases} \tag{12}$$

where $B(\mathbf{x})$ is the background luminance, which is calculated as the mean luminance of a local 3×3 block. With this formulation, the sensitivity (the inverse of LA) peaks at luminance ranging from 100 to 200. As luminance approaches an extremely low value (darker area), human perception fails to discern minor variations. Conversely, at high luminance levels, the detection ability becomes progressively more difficult, albeit less so compared to the darker regions.

Contrast masking (CM): It is obvious that stimuli become less perceptible against patterned, non-uniform backgrounds. In our context, perturbations can seamlessly blend into regions with intricate textures, while they leave distinct traces on flat surfaces. To simulate this effect, we first compute the luminance contrast [26] by convolving an input image with four directional filters. This process highlights the complexity of a region by contrasting it with its immediate surroundings. Subsequently, contrast masking is established by mapping the luminance contrast onto a logarithmic curve [57].

$$CM(\mathbf{x}) = 0.115 \times \frac{16 \times LC(\mathbf{x})^{2.4}}{LC(\mathbf{x})^2 + 26^2}, \tag{13}$$

where luminance contrast $LC(\mathbf{x})$ is obtained via four directional filters as $LC(\mathbf{x}) = \frac{1}{16} \times \max_{k=1,\ldots,4} |\mathbf{x} * \nabla_k|$ following Wu et al. [58].

Contrast sensitivity function (CSF): The human visual system exhibits a band-pass response to spatial frequency and CSF represents a function of how our eye is sensitive to the contrast of signals at various spatial frequencies. High

spatial frequencies correspond to rapid changes in image details, and contrast sensitivity is optimal at moderate spatial frequencies. Beyond a certain frequency threshold, known as the resolution limit, our eyes are unable to detect changes. Upon this concept, several studies have introduced and modified the CSF [8, 25, 33, 34]; the CSF model $H(f, \theta)$ is given below.

$$H = \begin{cases} 2.6 \times (0.0192 + 0.114 f_\theta) e^{-(0.114 f_\theta)^{1.1}}, & \text{if } f \geq 7.8909 \\ 0.981, & \text{otherwise.} \end{cases} \tag{14}$$

Here, f is the radial spatial frequency, measured in cycles per degree of visual angle (c/deg). The variable θ, ranging from $[-\pi, \pi]$, denotes the orientation. Additionally, $f_\theta = f/[0.15 \times \cos(4\theta) + 0.85]$ accounts for the oblique effect [25].

The CSF model is now applied to an image in the frequency domain as $\mathbf{x}_{csf} = \mathcal{F}^{-1}[H(u, v) \times \mathcal{F}(\check{\mathbf{x}})]$, where $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ represent the DFT and its inverse, respectively. Here, $H(u, v)$ is the DFT-version of $H(f, \theta)$, with u, v being the DFT indices. The transformation from $H(f, \theta)$ to $H(u, v)$ is elaborated in Larson et al. [25]. Prior to applying the CSF model, an input image $\mathbf{x}$ is converted into a perception-adjusted luminance form to reflect the non-linear relationship between digital pixel values and physical luminance [25]. This is achieved by calibrating $\mathbf{x}$ to the settings of an sRGB display and converting it into perceived luminances, which indicates the relative lightness: $\check{\mathbf{x}} = \sqrt[3]{(0.02874\mathbf{x})^{2.2}}$.

Standard deviation: It has served as a crucial metric for quantifying the structural information of an image, both in pixel space [55] and feature space [10]. Given that perturbations tend to be less noticeable in areas exhibiting high levels of change, (as discussed in contrast masking), we focus on these sudden changes, interpreting them as structural components. For this purpose, we compute the block-wise standard deviation in pixel space using a 9×9 local block.

Entropy: Similar to standard deviation, block-wise entropy is calculated using a 9×9 local window. This approach is based on the concept that the complexity of a local region influences the detectability of subtle perturbations. The higher the local entropy, the more intricate the region, potentially rendering minor perturbations less perceptible.

Post-processing: All the JND estimations are min-max normalized and then inverted by subtracting from one. When constructing a perceptual map $\mathcal{M}$ with JND estimations, we encountered issues with some JNDs displaying extremely skewed distributions. Additionally, the distributions of JNDs often do not align with each other, posing challenges in their combination. While standardization aligns the distributions, we found that the continuous JND values provide weak signals as masks (*e.g.* on average, the protection strength drops to 50% compared to the original). To give a more distinct signal depending on the JND values, we discretize the scores by quantizing the JND values. This involves calculating JND quantiles, where the first quantile is set to 1.0. For each subsequent quantile, we multiply by a factor of β to decrease the value in a discrete manner. We use $\beta = 0.85$, resulting in quantized JND values across four quantiles as [1.0, 0.85,

0.7225, 0.6141]. This allows for a more nuanced adjustment of the protection intensity while accommodating the varied distributions of JND estimations.

A.2 Perceptual Constraints

Masked LPIPS. We use AlexNet [23] as the backbone network for the LPIPS loss, adhering to the parameter settings in the official LPIPS documentation.

Masked low-pass. In this constraint, we utilize $\mathrm{LP}(\cdot)$, a reconstruction function focusing on the low-frequency component. Inspired by Luo et al. [31], we implement a DWT-based reconstruction module. Given an input image $\mathbf{x}$, DWT decomposes it into one low-frequency component and three high-frequency components, as expressed by the following equations.

$$\mathbf{x}_{ll} = \mathbf{L}\mathbf{x}\mathbf{L}^T, \;\; \mathbf{x}_{lh} = \mathbf{H}\mathbf{x}\mathbf{L}^T, \;\; \mathbf{x}_{hl} = \mathbf{L}\mathbf{x}\mathbf{H}^T, \;\; \mathbf{x}_{hh} = \mathbf{H}\mathbf{x}\mathbf{H}^T. \tag{15}$$

$\mathbf{L}$ and $\mathbf{H}$ represent the low-pass and high-pass filters of an orthogonal wavelet, respectively. To reconstruct an image using only its low-frequency component, we input $\mathbf{x}_{ll}$ alone into the inverse DWT function. Thus, $\mathrm{LP}(\mathbf{x})$ is defined as:

$$\mathrm{LP}(\mathbf{x}) \;=\; \mathbf{L}^T\mathbf{x}_{ll}\mathbf{L} \;=\; \mathbf{L}^T(\mathbf{L}\mathbf{x}\mathbf{L}^T)\mathbf{L}. \tag{16}$$

CLIP. As outlined in Eq. 10, IMPASTO focuses on maximizing the feature distance between the protected image and the prompt C = "Noise-free image". While it is conceivable to minimize the distance using a 'bad' prompt (*e.g.* C = "noisy image") or to integrate both 'good' and 'bad' prompts, we observed that all these show comparable performance. Therefore, we opted to employ the 'good' prompt only in the CLIP-based constraint.

A.3 Implementation Details

When we implement IMPASTO into the existing protection methods, we adhere to their respective settings for optimizing perturbations. Regarding the IMPASTO's own components, we use following hyperparameters depicted in Eq. 11 as: $\lambda_L = 5.0, \lambda_{LP} = 10.0, \lambda_C = 0.1$. For our PAP (Eq. 6), we utilize the L_∞ norm. On the other hand, for the IWR loss (Eq. 7), we employ the L_2 norm. When we calculate IWR loss, since the distance between $\mathcal{L}_{\mathcal{SP}}$ respect to $\mathcal{M}'$ and $\mathcal{M}(\omega)$ (first term) is significantly smaller than the L_2 constraint (second term), we amplify the former term by a factor of 5×10^7. In addition, during our experiments, it was observed that the magnitudes of $\mathcal{L}_{\mathcal{E}}$ and $\mathcal{L}_{\mathcal{SD}}$ differ considerably, with the SD loss exhibiting much smaller values. Therefore, when integrating IMPASTO into UNet-based protection methods (*e.g.* AdvDM [30], Anti-DreamBooth [52], Mist [29]), we further scale down all the $\lambda_L, \lambda_{LP}, \lambda_C$ by a factor of 0.05. This adjustment is made to balance the influence of our proposed components across both protection methods, ensuring that the impact of IMPASTO is consistent and effective in enhancing image protection.

B Experimental Settings

Datasets. To create the "painting" dataset, we gather artworks from the WikiArt dataset [51]. This dataset includes selections from 15 artists, 10 artworks per each. The "cartoon" dataset is sourced from the NAVER WEBTOON platform; images are cropped to contain only character depictions. This comprises 15 cartoons with 10 cartoon character images each.

Evaluation. To assess the quality of the protected images, we utilize three deep learning-based full-reference quality assessments: DISTS [10], PieAPP [41], and TOPIQ [4]. These measures compare the protected images against the original artworks. DISTS is designed to be sensitive to structural changes while exhibiting explicit tolerance to texture resampling. It achieves this by evaluating both the mean and correlation of feature maps from the compared images. PieAPP measures the perceptual error of a distorted image with respect to a reference and its training dataset. TOPIQ adopts a top-down approach, leveraging high-level semantics to direct the quality assessment network's focus towards semantically significant local distortion regions.

For evaluating protection performance, we employ NIQE [36], BRISQUE [35], and FID [15]. NIQE and BRISQUE are blind image quality assessment methods that only use generated images for evaluation. BRISQUE operates on the premise that distortions in natural images disrupt pixel distributions as well. It processes an input image and then extracts features, which are finally mapped to a mean opinion score using an SVM regressor [47]. NIQE, in contrast, does not rely on opinion scores but instead assesses the quality of an image by comparing its statistical properties with those of a clean image dataset. Despite their widespread use, these methods have limitations in our context: 1) They are designed primarily for natural image domains, making their measurements potentially unreliable in artistic domains, particularly for cartoons. 2) Traditional non-reference assessments focus on common distortions such as JPEG compression, blurring, and Gaussian noise, which may not effectively capture the unique artifacts in generated images in our task. FID, conversely, measures the distance between two distributions—the original artwork and the generated image—providing more reliable assessments. For example, as protection strength increases, FID is the only metric where scores worsen.

User study. We conducted a user study in the form of an A/B test with a reference image as a benchmark. A total of 44 participants were involved, tasked with determining the better-quality protected image (given the reference original artwork) and identifying the lower-quality generated image (given a reference DreamBooth generated image). Each participant was asked to vote on 8 questions for PhotoGuard [46] and another 9 for AdvDM [30].

C Additional Analyses and Results

Ablation study. In Table 6, we conduct additional component analysis by omitting certain components from the full Impasto framework. Specifically, we

Table 6: Component analysis. Within IMPASTO, we conduct an analysis by excluding the JND-based perceptual map generation, resulting in the creation of the perceptual map from random masks (denoted as w/o JND). Additionally, we evaluate the impact of omitting the perception-aware protection (denoted as w/o PAP).

Method	Protected Image Quality		Protection Performance	
	DISTS (↓)	TOPIQ (↑)	BRISQUE (↑)	FID (↑)
IMPASTO (Full model)	0.159	0.912	20.74	279.3
w/o JND	0.158	0.912	19.30	268.1
w/o PAP	0.199	0.890	17.57	281.1

Table 7: Evaluating the impact of varying refinement Step P. In this analysis, we explore the effects on performance when adjusting the number of refinement steps P within the instance-wise refinement process.

Method	Protected Image Quality		Protection Performance	
	DISTS (↓)	TOPIQ (↑)	BRISQUE (↑)	FID (↑)
$P = 100$	0.159	0.913	19.29	278.4
$P = 75$	0.159	0.913	19.06	278.2
$P = 50$	0.158	0.913	18.19	279.7
$P = 25$	0.159	0.912	20.74	279.3
$P = 10$	0.159	0.913	18.49	274.3

assess the impact of 1) removing the JND-based perception map, which results in initializing $\mathbf{M}$ with random masks (w/o JND), and 2) excluding the proposed perception-aware protection (PAP), thereby relying solely on the perceptual constraint bank (w/o PAP). For the w/o JND case, in contrast to Table 3, we start with a full-component model and remove only the JND part. The results indicate that discarding the JNDs leads to a decrease in protection performance while maintaining similar image quality, corroborating the findings in Table 3. When PAP is not applied, there is a slight increase in protection performance (FID: $279.3 \rightarrow 281.1$) with significant compromise on image quality. These observations suggest that the integration of PAP alongside a JND-based perceptual map is crucial for achieving an imperceptible style protection framework that effectively balances protection performance with image quality.

Varying refine step in IWR. In this analysis, we investigate the effect of varying the refining step P during the instance-wise refinement (IWR) process (Algorithm 1 L8-10) and assess the changes in protection performance (Table 7). We observe that though P is decreased from 100 to 25, both the image quality and protection capability remain relatively consistent. This performance stability can be attributed to the initial utilization of a JND-based perceptual map $\mathcal{M}$, which necessitates only minor adjustments during the IWR step. However, a notable reduction in protection performance is shown when P is reduced to 10, indicating that such a small step is insufficient for achieving optimal refinement. Consequently, we set $P = 25$ as the default setting, balancing both performance efficacy and the number of optimization steps.

IMPASTO with other protection methods. In addition to PhotoGuard [46] and AdvDM [30], we extend the application of IMPASTO to recent protection techniques, Anti-DreamBooth [52] and Mist [29]. Anti-DreamBooth, a variant

Table 8: Quantitative comparison of protection methods w/ and w/o IMPASTO, both selected for their comparable protection performance. IMPASTO markedly elevates the protected images' quality while maintaining comparable levels of protection efficacy.

Dataset	Method	Protected Image Quality			Protection Performance		
		DISTS (↓)	LPIPS (↓)	TOPIQ (↑)	NIQE (↑)	BRISQUE (↑)	FID (↑)
Painting	Anti-DB [52]	0.151 (+0.000)	0.081 (+0.000)	0.876 (+0.000)	3.865	13.63	266.1
	+ IMPASTO	0.138 (+0.013)	0.008 (+0.073)	0.889 (+0.013)	3.780	12.60	270.9
	Mist [29]	0.167 (+0.000)	0.100 (+0.000)	0.846(+0.000)	4.052	13.96	272.2
	+ IMPASTO	0.144 (+0.023)	0.077 (+0.023)	0.879(+0.033)	3.830	11.11	273.0
Cartoon	Anti-DB [52]	0.260 (+0.000)	0.154 (+0.000)	0.700(+0.000)	4.437	15.48	160.6
	+ IMPASTO	0.234 (+0.026)	0.027 (+0.127)	0.782(+0.092)	4.589	12.35	161.0
	Mist [29]	0.256 (+0.000)	0.160 (+0.000)	0.709(+0.000)	4.597	10.86	158.7
	+ IMPASTO	0.238 (+0.018)	0.077 (+0.083)	0.772(+0.063)	4.693	11.23	158.3

of AdvDM, incorporates DreamBooth tuning into its optimization process. Mist utilizes a hybrid approach, combining both Encoder (*e.g.* PhotoGuard) and UNet-based (*e.g.* AdvDM) protection methods (similar to Eq. 4). We adhere to the hyperparameters specified in their official implementations. Regarding IMPASTO, we employ the same settings used in our application to AdvDM (Sec. A.3), for both Anti-DreamBooth and Mist. As shown in Table 8, IMPASTO notably enhances the quality of the protected images with comparable protection performances. Fig. 11 and 12 also support the efficacy of our method; IMPASTO significantly diminishes visual artifacts in the protected images while preserving the original methods' style protection ability. These experiments highlight the adaptability and efficacy of IMPASTO across various protection frameworks.

Additional qualitative results. Fig. 13 and 14 provide additional visual comparisons, illustrating the impact of applying IMPASTO in conjunction with PhotoGuard [46] and AdvDM [30], respectively.

We also compare the robustness on countermeasure of IMPASTO in Fig. 15, including JPEG compression, blurring, and noise addition. Across all these countermeasures, our method demonstrates a protection capability comparable to that of the baseline. Furthermore, in Fig. 16, we present additional qualitative results demonstrating the generalization capability of IMPASTO. When applied to three different personalization methods—DreamBooth, LoRA, and DreamStyler—IMPASTO maintains performance levels akin to those observed in their respective non-IMPASTO implementations. This highlights the effectiveness and adaptability of IMPASTO across diverse personalization techniques.

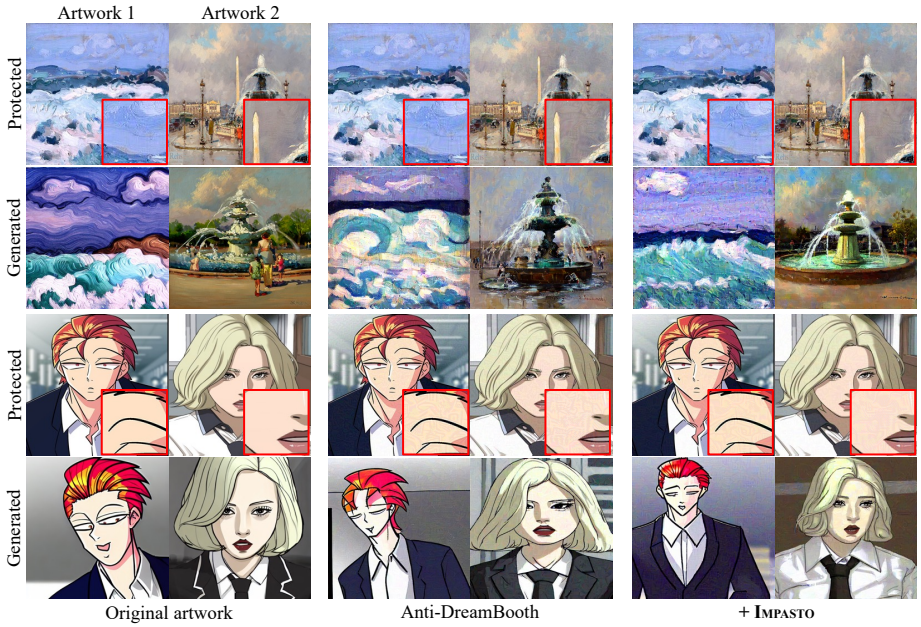


Fig. 11: Qualitative comparison of Anti-Dreambooth [52] with and without Impasto.

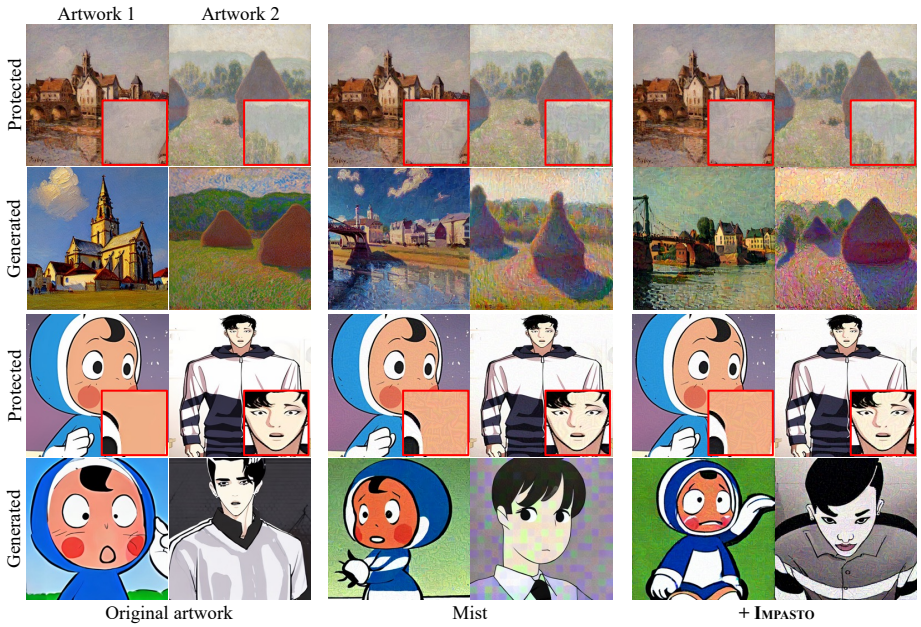


Fig. 12: Qualitative comparison of Mist [29] with and without Impasto.

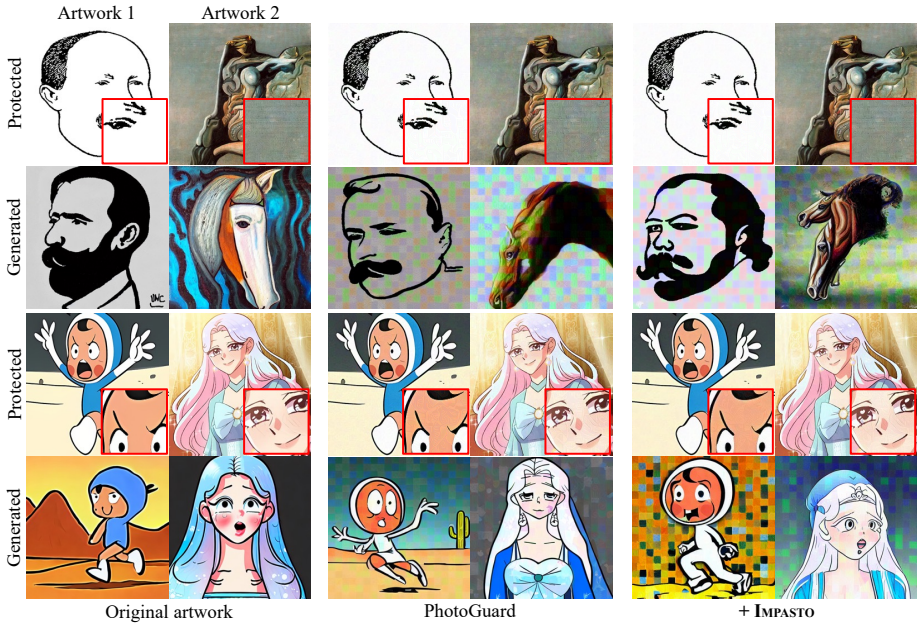


Fig. 13: Qualitative comparison of PhotoGuard [46] with and without IMPASTO.

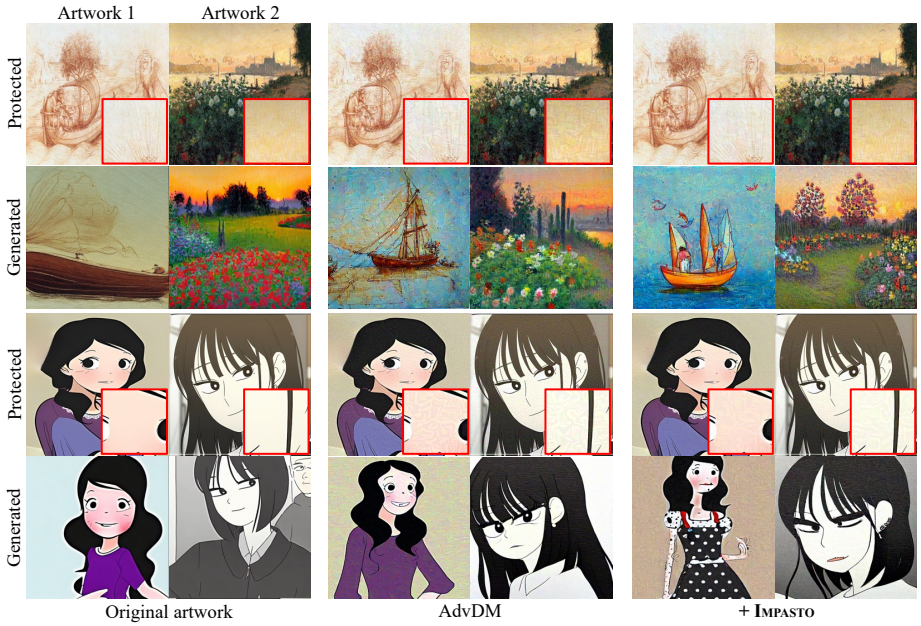


Fig. 14: Qualitative comparison of AdvDM [30] with and without IMPASTO.

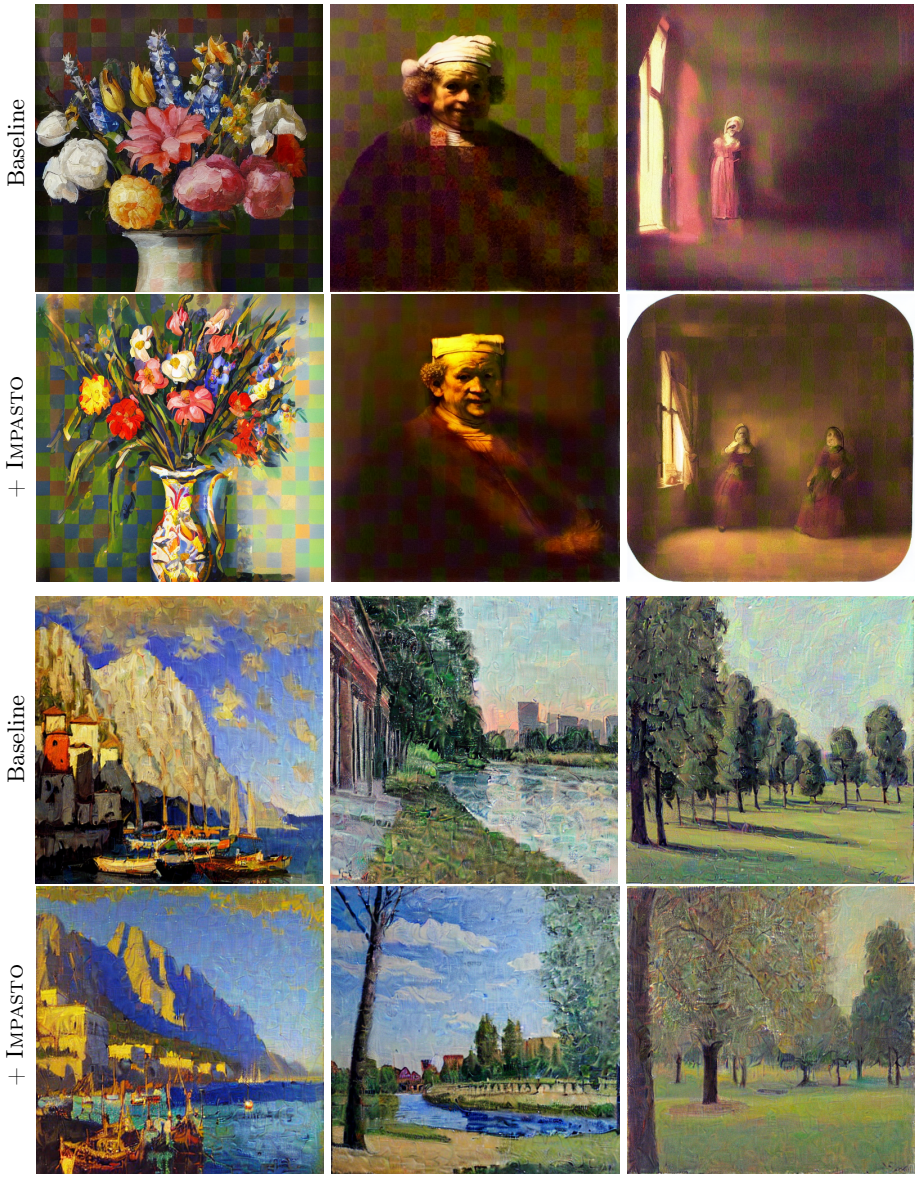


Fig. 15: Qualitative comparison on robustness. Methods with Impasto exhibit comparable protection to the baselines.

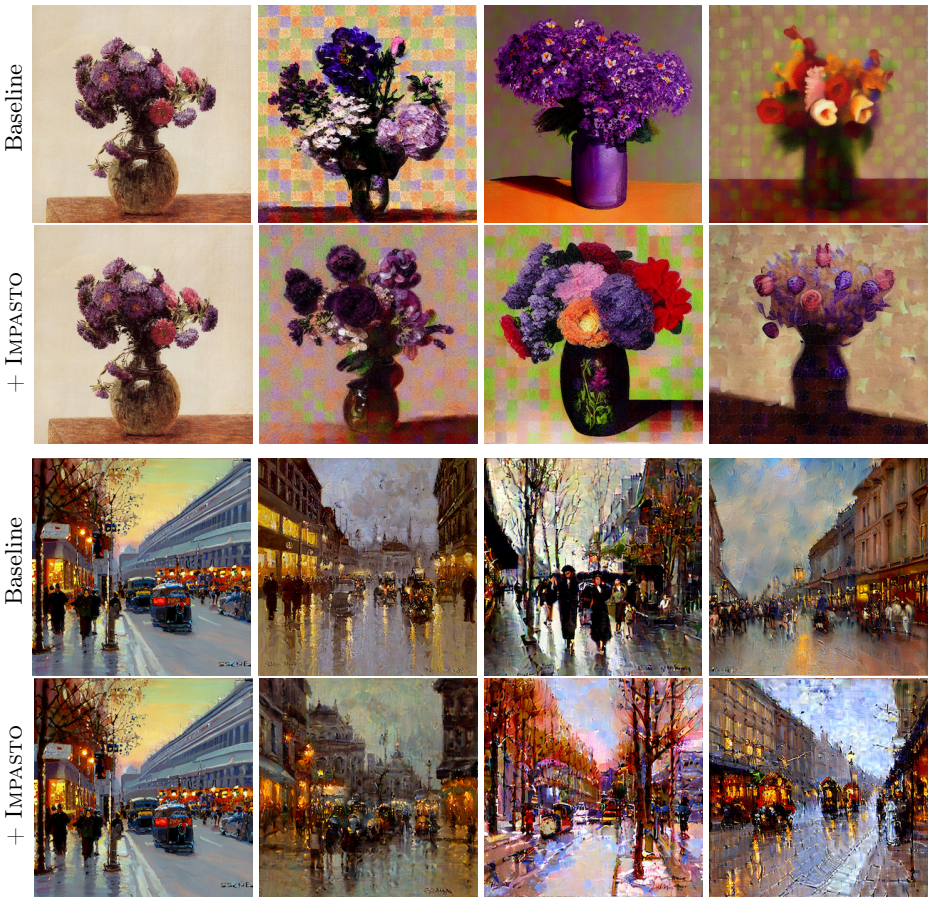


Fig. 16: Qualitative comparison on generalization. Impasto does not impede protection methods' generalization abilities across diverse personalization methods; DreamBooth [44], LoRA [17], and DreamStyler [1]