Evaluating Adversarial Protections for Diffusion Personalization: A Comprehensive Study

Kai Ye [1] **Tianyi Chen** [2] **Zhen Wang** [1]

Abstract

With the increasing adoption of diffusion models for image generation and personalization, concerns regarding privacy breaches and content misuse have become more pressing. In this study, we conduct a comprehensive comparison of eight perturbation-based protection methods—AdvDM, ASPL, FSGM, MetaCloak, Mist, PhotoGuard, SDS, and SimAC—across both portrait and artwork domains. These methods are evaluated under varying perturbation budgets, using a range of metrics to assess visual imperceptibility and protective efficacy. Our results offer practical guidance for method selection. Code is available at: https://github.com/vkeilo/DiffAdvPerturbationBench.

1. Introduction

In recent years, generative models based on Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have achieved remarkable progress in image synthesis. Unlike traditional Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), which rely on adversarial training, diffusion models learn data distributions via forward and reverse sampling, enabling high-fidelity image generation. Building on this, researchers have introduced personalized generation techniques that let models quickly learn individual visual or artistic styles from only a few samples (Ruiz et al., 2023; Kumari et al., 2023; Gal et al., 2022), expanding the possibilities of customized content creation.

While personalized diffusion techniques offer creative potential, fine-tuning models with few samples raises significant privacy and copyright risks (Liu et al., 2024a). Public or covert photos can be exploited to generate harmful content, and artists' styles can be misused without consent, underscoring the need for strong protections against privacy breaches and unauthorized style use. A promising defense involves adversarial perturbations, which disrupt model fine-tuning and hinder the reproduction of specific identities or styles (Yang et al., 2021).

To address the lack of standardized evaluation in current perturbation-based defenses for diffusion models, this paper proposes a unified benchmarking framework that systematically compares eight representative protection methods across two core tasks: identity protection and style imitation prevention. The framework supports multi-level perturbation control and incorporates a diverse set of perceptual and semantic metrics to jointly assess stealthiness and defensive efficacy. We apply this evaluation across two representative domains—portrait (VGGFace2) and artwork (WikiArt)—and report comprehensive experimental results that reveal trade-offs, robustness patterns, and sample-level variability. This study provides actionable insights for selecting appropriate protection strategies under different deployment constraints, and offers a generalizable foundation for future research on privacy-preserving generative models.

2. Related Work

2.1. Diffusion Models and Customization

Diffusion Models (DMs) generate high-quality images by gradually adding and removing noise, learning a reverse process to reconstruct samples (Ho et al., 2020). Compared to VAEs (Kingma, 2013) and GANs (Goodfellow et al., 2020), DMs offer superior image fidelity and training stability. Advances like classifier-free guidance (Ho & Salimans, 2022) and Latent Diffusion Models (LDMs) (Rombach et al., 2022) further enhance expressiveness and efficiency, enabling scalable conditional generation.

To support personalization, methods like Textual Inversion (Gal et al., 2022) optimize embeddings without tuning model weights, while DreamBooth (Ruiz et al., 2023) fine-tunes diffusion models using a few reference images. Custom Diffusion (Kumari et al., 2023) improves conditioning precision, and DiffuseKronA (Marjit et al., 2024) reduces parameter overhead via Kronecker-based adapters. Despite these benefits, techniques raise privacy and copyright risks,

[1]School of Cyberspace, Hangzhou Dianzi University, Hangzhou, China [2]Microsoft, Redmond, USA. Correspondence to: Zhen Wang <wangzhen@hdu.edu.cn>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

as they may be exploited to generate harmful or plagiarized content, emphasizing the need for robust protection mechanisms.

2.2. Perturbation-Based Privacy Protection

Adversarial perturbation methods have emerged to protect users against unauthorized use of personal or artistic images. These approaches introduce small, carefully crafted noise to disrupt fine-tuning or generation.

AdvDM (Liang et al., 2023) maximizes loss during denoising to hinder feature extraction. Anti-DreamBooth (Van Le et al., 2023) combines surrogate models with learned perturbations to defend against DreamBooth-style personalization. PhotoGuard (Salman et al., 2023) applies perturbations in the latent space, while GLAZE (Shan et al., 2023) targets style imitation with subtle visual noise. Mist (Liang & Wu, 2023) introduces texture- and semantics-aware losses for robust protection. SimAC (Wang et al., 2024) leverages frequency awareness and timestep selection to suppress identity features effectively.

MetaCloak (Liu et al., 2024b) uses meta-learning and surrogate models to ensure robust protection under diverse transformations. DisDiff (Liu et al., 2024a) disrupts prompt-based generation by erasing attention via cross-attention and schedule tuning.

Despite progress, existing methods often trade off stealthiness and protection strength, with performance varying across tasks and budgets. A unified evaluation framework is needed to compare these techniques systematically and guide real-world deployment.

3. Method

3.1. Prerequisite Knowledge

Diffusion Models generate high-quality data by denoising noisy samples through a learned reverse process. Starting from an original sample $\mathbf{x}_0$, the forward process adds Gaussian noise to form a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ via:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (1)$$

where β_t controls the noise level. The reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is learned by minimizing:

$$L_{\text{denoise}} = \mathbb{E}_{\mathbf{x}_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right] \quad (2)$$

Here, ϵ is sampled noise, and $\epsilon_\theta(\cdot)$ is the predicted noise. Optimizing this loss enables faithful data generation.

However, when fine-tuned on few target samples, diffusion models can produce high-fidelity imitations, posing privacy and copyright risks. To counter this, adversarial perturbations δ can be added to inputs to hinder feature learning by disrupting the denoising objective during training.

3.2. Perturbation Optimization Objective

The goal of adversarial perturbations is to slightly modify input $\mathbf{x}$ to disrupt the generative process:

$$\delta^* = \arg\min_{\|\delta\|\leq\epsilon} \alpha L_{\text{denoise}}(\mathbf{x}+\delta) \quad (3)$$

where δ^* is optimized via Projected Gradient Descent (PGD). This optimization hinders the model's ability to learn accurate representations, reducing output fidelity. The overall protection pipeline is illustrated in Figure 1.

Different methods adopt distinct strategies for optimizing δ^* in this setting: **Mist** adds semantic-aware losses (e.g., texture loss), **SimAC** applies frequency-aware filtering and timestep control, while **MetaCloak** leverages surrogate feedback for robust adaptation under diverse conditions.

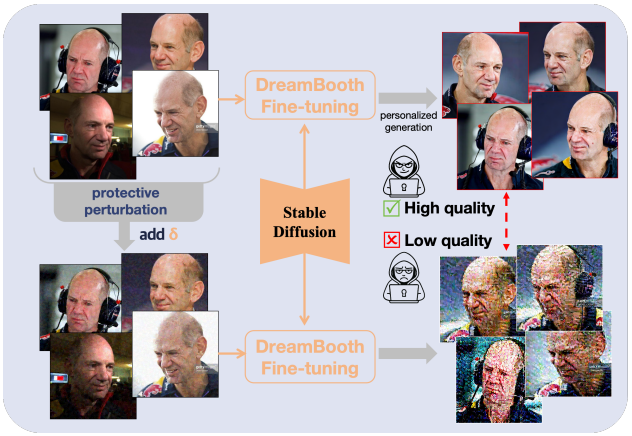

Figure 1. Overview of protective perturbation against diffusion-based personalization. Adding perturbation δ to training images degrades output quality after DreamBooth fine-tuning, preventing unauthorized identity replication.

3.3. Evaluation Framework Design

To enable fair and consistent evaluation of protective perturbations for diffusion models, we introduce a unified framework that assesses both perturbation effectiveness and impact on customization. It supports batch processing across datasets and standardized metric-based comparisons, and consists of three modules:

Perturbation Sample Generation: We unify existing open-source implementations into a batch-compatible interface that supports custom datasets and standardized control over budgets and parameters.

Customized Model Training: Perturbed samples are used to train personalized models using a shared training pipeline to ensure consistent customization procedures.

Sample Generation and Evaluation: Customized models regenerate samples with consistent prompts, and evaluation uses diverse standardized metrics for comparability.

Our unified framework standardizes the evaluation of protective perturbations across tasks and metrics, enabling fair benchmarking of current methods and supporting future extensions. It reveals how performance varies with perturbation strength and task type, offering insights into the robustness and applicability of each method.

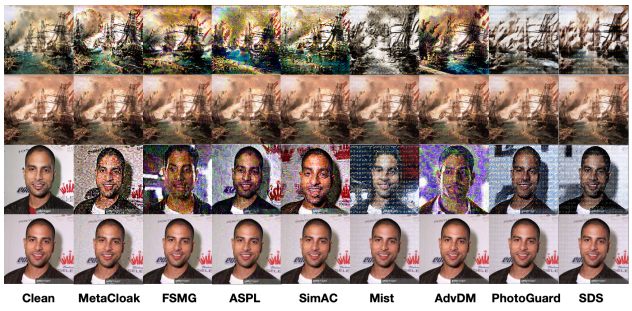

Figure 2. Comparison of perturbations and DreamBbooth results on WikiArt (top) and VGGFace2 (bottom). Each column: perturbed input (bottom) and generation (top); clean sample on the left.

4. Experiments

4.1. Selection of Datasets, Models, and Algorithms

We select the **VGGFace2** (Cao et al., 2018) and **WikiArt** (Saleh & Elgammal, 2015) datasets as our primary evaluation benchmarks. **VGGFace2** is a diverse face dataset with identity labels, commonly used for identity-preserving and privacy-related generation tasks. **WikiArt** contains artworks from various artists and styles, making it a standard benchmark for evaluating stylistic fidelity and cross-domain consistency.

For each dataset, we sample 50 homogeneous groups (same identity or artist), each with 8 images: 4 for clean reference and 4 for perturbation. Methods requiring pretraining are given 4 extra support images. All images are center-cropped and resized to 512×512.

We evaluated eight perturbation methods on a workstation equipped with 4 × RTX 4090 GPUs: ASPL (Van Le et al., 2023), FSGM (Van Le et al., 2023), SimAC (Wang et al., 2024), AdvDM (Liang et al., 2023), Mist (Liang & Wu, 2023), PhotoGuard (Salman et al., 2023), MetaCloak (Liu et al., 2024b), and SDS (Xue et al., 2023). The step size is fixed at $1/255$ per iteration, with all other hyperparameters following default settings in their respective papers. See Figure 2 for visual examples of all methods.

4.2. Evaluation Metrics

We categorize the evaluation metrics into two major groups: **perturbation perceptibility metrics** and **generated image quality metrics**.See Appendix A for detailed descriptions of each metric.

Perturbation perceptibility is evaluated using PSNR, LPIPS, SSIM, and CIEDE2000, which measure low-level visual differences and perceptual similarity.

Generated image quality is assessed via FID, PSNR, BRISQUE, LIQE, CLIP-IQA, and CLIP-IQAC, covering both perceptual fidelity and semantic consistency.

4.3. Results and Discussion

4.3.1. Algorithm Performance Analysis

Table 1 reports perceptibility scores on VGGFace2 under overall, low, and high perturbation levels. **MetaCloak** excels in pixel-level and structural similarity for identity protection, while **FSGM** performs best for artistic styles. **SimAC** consistently yields the lowest LPIPS, indicating strong deep-feature stealth. Stealthiness varies with budget: **MetaCloak** leads at $4/255$, but **FSGM** and **SimAC** excel at $16/255$, stressing the importance of budget-aware method selection.

Protective perturbations aim to degrade personalized outputs post-finetuning by disrupting input semantics. Table 2 shows generation quality on VGGFace2 and WikiArt. On VGGFace2, **MetaCloak** yields the highest FID, and **SimAC** the lowest CLIP-IQAC, indicating strong disruption to structure and semantics. **Mist**, **PhotoGuard**, and **SDS** lead on distortion-based metrics (BRISQUE, PSNR), reflecting greater visual degradation. **SimAC** also scores well in LIQE, underscoring its broad disruptive impact.

Across both datasets, method effectiveness shows consistent patterns: **SimAC** and **MetaCloak** excel in semantic disruption, while **PhotoGuard**, **SDS**, and **Mist** are stronger in low-level distortion. This highlights the need to align method choice with protection goals—semantic vs. visual degradation. Similar to perceptibility trends, performance varies with perturbation level: **SimAC** dominates at $4/255$, whereas **MetaCloak** leads at $16/255$ on most quality metrics. Notably, **FSGM** retains high generation quality despite good stealth at high budgets, suggesting limited disruption. Detailed results are provided in Appendix B.

4.3.2. Analysis of Sample-Level Variability

In our experiments, we found that the performance of perturbation methods fluctuates significantly across individual identity samples. This variability is largely due to the strong sensitivity of downstream personalization models (e.g., DreamBooth) to training image properties, rather than the inherent robustness of the perturbation methods themselves. Correlation analysis (Figure 3) reveals that higher consistency among training images improves output PSNR, LIQE, and FID—indicating enhanced structural fidelity and stable identity learning. Additionally, a strong BRISQUE correlation ($r = 0.615$) suggests that output quality is closely tied to the quality of the input samples.

Table 1. Comparison of perceptibility metrics across different datasets and perturbation budgets. Columns show results on the VGGFace2 datasets under average, low ($\epsilon = 4/255$), and high ($\epsilon = 16/255$) perturbation strengths. Higher PSNR and SSIM, and lower LPIPS and CIEDE2000 indicate better visual stealth.

Method	VGGFace2 Avg				VGGFace2 Low ($\epsilon = 4/255$)				VGGFace2 High ($\epsilon = 16/255$)			
	PSNR ↑	LPIPS ↓	SSIM ↑	CIEDE2000 ↓	PSNR ↑	LPIPS ↓	SSIM ↑	CIEDE2000 ↓	PSNR ↑	LPIPS ↓	SSIM ↑	CIEDE2000 ↓
AdvDM	**33.141**	0.274	0.785	2597.298	40.688	0.080	0.959	930.766	27.866	0.430	0.619	4132.730
ASPL	31.821	0.311	0.757	2942.567	38.410	0.121	0.932	1263.338	27.137	0.463	0.605	4410.034
FSGM	32.751	0.289	0.797	2479.031	38.596	0.120	0.934	1224.997	**28.975**	0.412	**0.692**	**3413.226**
MetaCloak	33.119	0.274	**0.799**	**2444.655**	**41.460**	**0.064**	**0.968**	**862.729**	27.451	0.444	0.640	3880.745
Mist	32.716	0.264	0.773	2511.432	40.314	0.067	0.955	866.704	27.363	0.430	0.603	4084.828
PhotoGuard	32.553	0.266	0.767	2546.989	40.142	0.068	0.953	882.197	27.224	0.433	0.595	4128.659
SDS	32.547	0.275	0.776	2501.789	40.144	0.071	0.953	885.339	27.243	0.447	0.619	3960.694
SimAC	32.667	**0.255**	0.787	2570.826	39.246	0.089	0.943	1123.767	28.116	**0.389**	0.650	3774.565

Table 2. Comparison of generation quality metrics across VGGFace2 and WikiArt datasets. Strong protection corresponds to higher BRISQUE and FID, and lower LIQE, CLIP-based scores, and PSNR.

Method	VGGFace2						WikiArt					
	LIQE ↓	BRISQUE ↑	FID ↑	PSNR ↓	CLIPIQA ↓	CLIP IQAC ↓	LIQE ↓	BRISQUE ↑	FID ↑	PSNR ↓	CLIPIQA ↓	CLIP IQAC ↓
clean	2.922	9.810	199.949	9.181	0.860	0.269	2.069	19.989	320.613	10.017	0.721	0.231
AdvDM	1.041	3.221	344.216	8.521	0.623	-0.323	1.105	9.665	379.464	9.312	0.542	-0.251
ASPL	**1.015**	6.247	356.409	8.561	0.562	**-0.411**	**1.065**	9.798	393.336	**9.223**	0.487	-0.350
FSGM	1.030	-3.199	328.851	8.472	0.635	-0.312	1.068	6.613	379.033	9.228	0.529	-0.283
MetaCloak	1.191	10.462	**381.347**	8.587	0.597	-0.373	1.382	18.120	395.469	9.952	0.465	-0.348
Mist	1.756	16.898	379.146	8.801	0.762	-0.159	1.845	23.552	410.557	10.006	0.489	-0.279
PhotoGuard	1.986	**19.858**	362.558	8.885	0.767	-0.112	2.095	**24.982**	405.753	9.747	0.515	-0.243
SDS	1.806	17.968	374.678	8.916	0.784	-0.136	1.888	23.952	**413.541**	9.620	0.535	-0.258
SimAC	1.048	12.779	362.534	**8.217**	**0.526**	-0.406	1.134	14.626	409.888	9.421	**0.415**	**-0.409**

However, excessive consistency may harm semantic generalization. We observed a significant negative correlation between CLIP-IQAC and training consistency (e.g., $r = -0.41$ with PSNR), suggesting that overly redundant samples can cause overfitting to local features, leading to semantic drift. To mitigate this, we recommend a "structurally mixed training set" strategy—combining both consistent and diverse subsets in each identity's training set. This hybrid design balances structural fidelity and semantic robustness, offering practical guidance for data selection in perturbation-based protection.

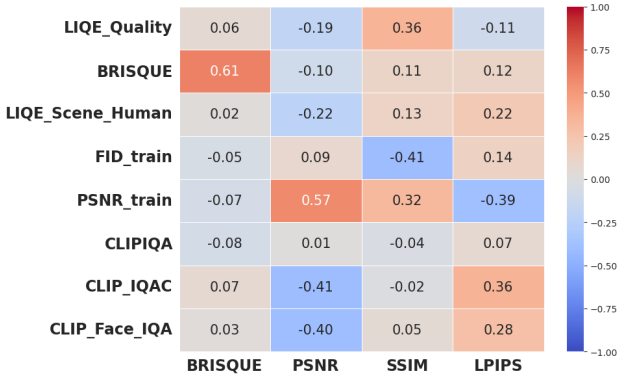

Figure 3. Pearson correlations between output quality (rows) and training image quality over 50 clean VGGFace2 identities.

5. Conclusion

With the increasing use of diffusion models for personalized image generation, protecting user privacy has become a critical and urgent concern. This paper presents a unified evaluation framework and systematically compares eight representative perturbation methods across two key dimensions: perceptibility and generation quality. Experimental results show that no single method dominates across all metrics or perturbation budgets. For instance, **SimAC** excels in perceptual stealth at low budgets, while **MetaCloak** proves more effective at degrading output quality under stronger perturbations, emphasizing the need to match strategies to specific deployment constraints.

Further analysis reveals that the effectiveness of protection methods varies significantly across individual samples, largely due to the sensitivity of personalization models to training image structure. Correlation studies on VGGFace2 suggest that training image consistency improves output fidelity but may hinder semantic generalization when overly excessive. We propose a "structurally mixed training set" strategy to better balance these factors. Experiments on WikiArt confirm consistent performance trends across domains, supporting the general applicability of these proposed methods. Future work will extend the framework to broader tasks and explore scalable privacy-preserving techniques for multimodal diffusion systems.

References

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and 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.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and 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.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and 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.

Liang, C. and Wu, X. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

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

Liu, Y., An, J., Zhang, W., Wu, D., Gu, J., Lin, Z., and Wang, W. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3587–3596, 2024a.

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

Marjit, S., Singh, H., Mathur, N., Paul, S., Yu, C.-M., and Chen, P.-Y. Diffusekrona: A parameter efficient fine-tuning method for personalized diffusion model. *arXiv preprint arXiv:2402.17412*, 2024.

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and 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.

Saleh, B. and Elgammal, A. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., and Madry, A. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

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

Sharma, G., Wu, W., and Dalal, E. N. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(1):21–30, 2005.

Van Le, T., Phung, H., Nguyen, T. H., Dao, Q., Tran, N. N., and 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.

Wang, F., Tan, Z., Wei, T., Wu, Y., and Huang, Q. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12047–12056, 2024.

Wang, J., Chan, K. C., and Loy, C. C. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

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

Xue, H., Liang, C., Wu, X., and Chen, Y. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2023.

Yang, C., Ding, L., Chen, Y., and Li, H. Defending against gan-based deepfake attacks via transformation-aware adversarial faces. In *2021 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and 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.

Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14071–14081, 2023.

A. Metric Descriptions

To ensure a comprehensive and fair evaluation of perturbation-based protection methods, we adopt a diverse set of metrics that reflect both perceptual quality and feature-level fidelity. These metrics are divided into two main categories:

Perturbation Perceptibility Metrics:

- **PSNR (Peak Signal-to-Noise Ratio)**: Measures the pixel-level similarity between the original and perturbed images. Higher PSNR indicates lower distortion.
- **SSIM (Structural Similarity Index)** (Wang et al., 2004): Evaluates structural similarity between two images, emphasizing luminance, contrast, and structure. Higher values indicate better visual similarity.
- **LPIPS (Learned Perceptual Image Patch Similarity)** (Zhang et al., 2018): A deep feature-based distance that correlates well with human perception. Lower scores indicate better perceptual similarity.
- **CIEDE2000** (Sharma et al., 2005): A perceptual color difference metric based on the CIE Lab color swpace. In our evaluation, we compute the L2 norm across all per-pixel CIEDE2000 values to obtain a global color distortion measure. Lower values indicate higher perceptual similarity in color.

Generated Image Quality Metrics:

- **BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator)** (Mittal et al., 2012): A no-reference quality score based on natural scene statistics. Lower values denote higher perceived quality.
- **LIQE (Learning-based Image Quality Evaluator)** (Zhang et al., 2023): A learned quality model trained on subjective ratings to predict perceptual quality. Lower scores are better.
- **FID (Fréchet Inception Distance)** (Heusel et al., 2017): Measures the distributional distance between generated and reference images using deep features. Lower values reflect better alignment with the target domain.
- **CLIP-IQA** (Wang et al., 2023): Computes perceptual similarity using CLIP features, capturing high-level alignment between generated and reference content. Lower values indicate greater degradation.
- **CLIP-IQAC (CLIP Image-Query Alignment Consistency)** (Liu et al., 2024b): Evaluates the consistency between a prompt (text) and generated images via CLIP similarity. Lower scores imply more effective protection.

B. Additional Evaluation Metrics

B.1. Generation Quality Results at Varying Perturbation Strengths

Table 3. Generation quality metrics of each method under different perturbation strengths $r \in \{4, 8, 12, 16\}/255$, on VGGFace2

r	Method	LIQE_Quality	BRISQUE	LIQE Scene Human	FID	PSNR	CLIPIQA	CLIP-IQAC	CLIP Face IQA
$0/255$	Clean Avg	1.646	8.838	0.810	283.106	8.760	0.725	-0.109	0.279
$4/255$	AdvDM	1.149	4.929	0.918	269.781	8.712	0.712	-0.105	0.256
	ASPL	1.052	6.066	0.900	289.036	8.791	0.652	-0.252	0.258
	FSGM	1.097	-0.691	0.936	279.614	8.559	0.702	-0.193	0.267
	MetaCloak	1.499	9.157	0.849	274.512	8.741	0.725	-0.116	0.206
	Mist	2.386	10.538	0.714	284.934	8.855	0.805	0.026	0.357
	PhotoGuard	2.430	13.953	0.746	265.942	8.961	0.797	0.058	0.369
	SDS	2.399	13.002	0.751	277.158	8.875	0.814	0.037	0.342
	SimAC	1.154	13.747	0.666	323.874	8.584	0.593	-0.329	0.179
$8/255$	AdvDM	1.011	2.183	0.864	330.384	8.526	0.631	-0.316	0.253
	ASPL	1.004	2.827	0.876	339.068	8.552	0.566	-0.409	0.242
	FSGM	1.015	-4.390	0.923	312.659	8.427	0.632	-0.300	0.279
	MetaCloak	1.153	11.021	0.491	372.615	8.523	0.639	-0.385	0.225
	Mist	1.642	17.696	0.389	376.225	8.977	0.768	-0.191	0.357
	PhotoGuard	1.934	19.669	0.413	354.655	8.826	0.787	-0.115	0.394
	SDS	1.647	18.096	0.489	367.428	8.906	0.787	-0.152	0.380
	SimAC	1.020	13.588	0.581	356.480	8.295	0.535	-0.414	0.187
$12/255$	AdvDM	1.003	3.555	0.820	373.292	8.486	0.586	-0.400	0.236
	ASPL	1.001	7.378	0.814	380.787	8.392	0.532	-0.476	0.227
	FSGM	1.004	-4.248	0.884	360.779	8.476	0.619	-0.364	0.276
	MetaCloak	1.079	11.466	0.286	429.366	8.543	0.546	-0.466	0.186
	Mist	1.495	18.461	0.248	416.673	8.616	0.759	-0.217	0.359
	PhotoGuard	1.768	22.404	0.350	407.268	8.861	0.746	-0.186	0.374
	SDS	1.583	19.586	0.313	411.580	9.044	0.789	-0.175	0.355
	SimAC	1.015	12.368	0.601	376.901	8.057	0.505	-0.439	0.180
$16/255$	AdvDM	1.001	2.219	0.769	403.408	8.359	0.563	-0.469	0.211
	ASPL	1.001	8.716	0.811	416.747	8.507	0.496	-0.507	0.209
	FSGM	1.003	-3.465	0.881	362.353	8.424	0.586	-0.394	0.265
	MetaCloak	1.032	10.204	0.226	448.894	8.542	0.477	-0.523	0.097
	Mist	1.503	20.896	0.205	438.753	8.757	0.718	-0.255	0.330
	PhotoGuard	1.812	23.406	0.278	422.369	8.891	0.739	-0.204	0.342
	SDS	1.595	21.188	0.271	442.548	8.838	0.745	-0.253	0.309
	SimAC	1.005	11.414	0.538	392.880	7.931	0.470	-0.443	0.156