中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

多模态人工智能系统
全国重点实验室
State Key Laboratory of
Multimodal Artificial Intelligence Systems

中国科学院自动化研究所
Institute of Automation
Chinese Academy of Sciences

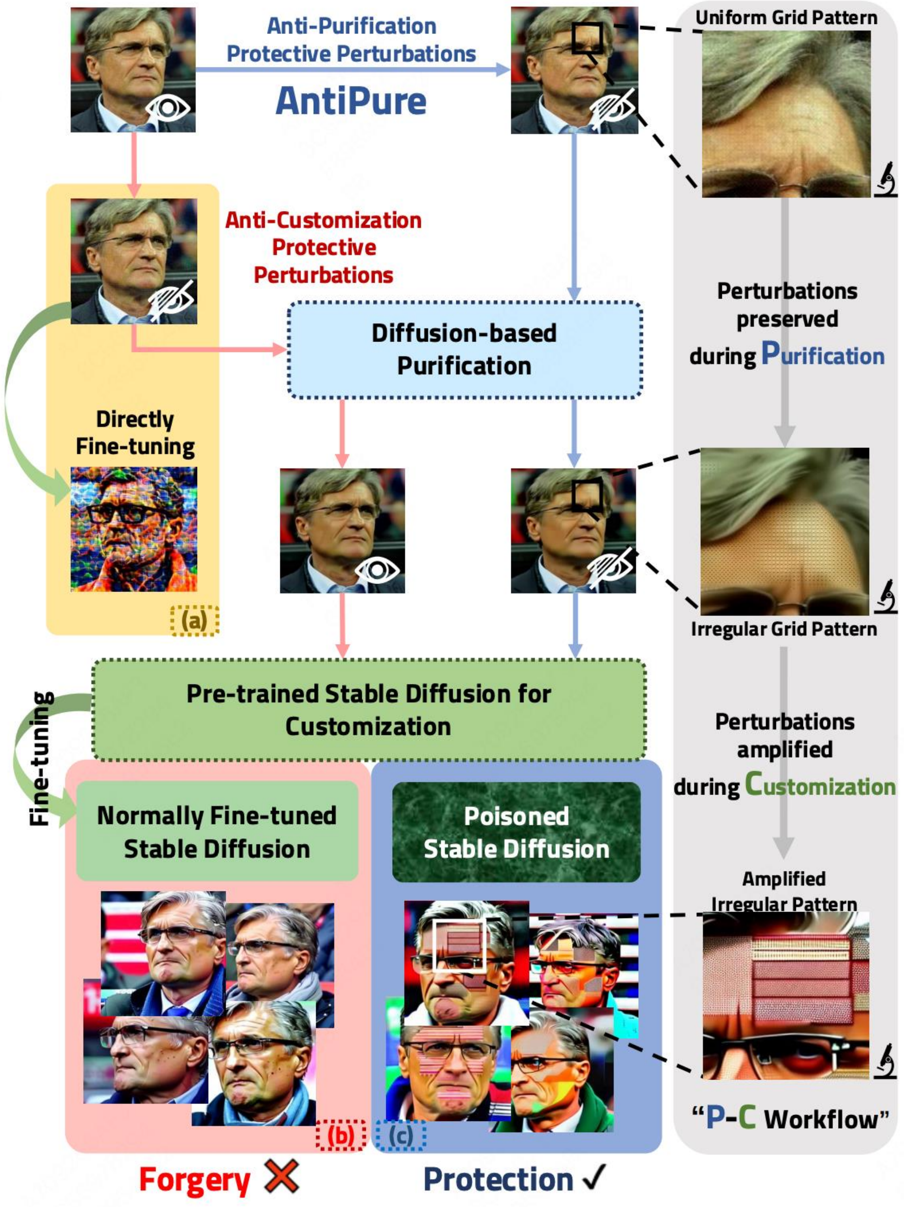
Towards Robust Defense against Customization via Protective Perturbation Resistant to Diffusion-based Purification

Wenkui Yang, Prof. Jie Cao & Prof. Ran He

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Towards Robust Defense against Customization via Protective Perturbation Resistant to Diffusion-based Purification

- □ a). Protective perturbations use small noises to distort the outputs of fine-tuned diffusion models.
- □ b). However, existing methods can be removed by diffusion-based purification.
- □ c). We propose a simple diagnostic method called **AntiPure**, which achieves protective perturbations resistant to purification and makes customization outputs more distinguishable.

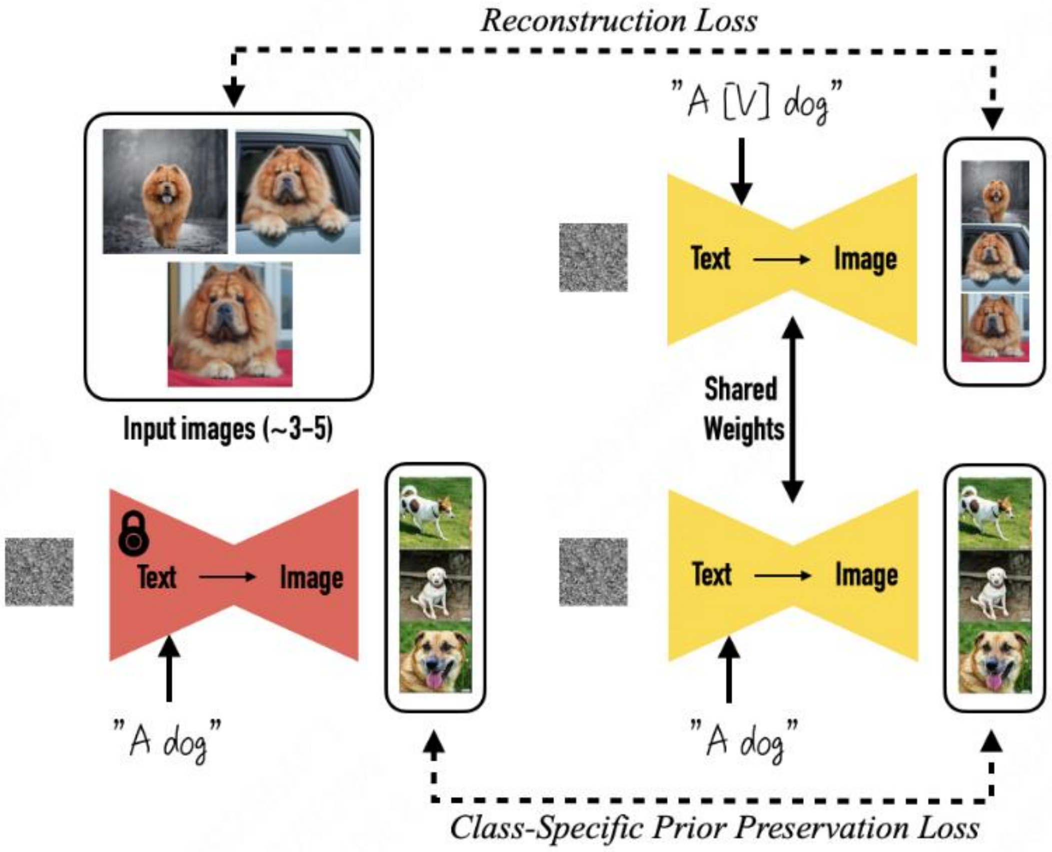

Anti-Purification Protective Perturbations
AntiPure
Uniform Grid Pattern
Anti-Customization Protective Perturbations
Diffusion-based Purification
Perturbations preserved during Purification
Directly Fine-tuning
(a)
Irregular Grid Pattern
Pre-trained Stable Diffusion for Customization
Fine-tuning
Perturbations amplified during Customization
Normally Fine-tuned Stable Diffusion
Poisoned Stable Diffusion
Amplified Irregular Pattern
(b)
(c)
"P-C Workflow"
Forgery ✗
Protection ✓

Preliminaries – Customization/Personalization

- **Let's revisit a commonly used few-shot fine-tuning method: DreamBooth**
  - The essence of customization is to fine-tune a model, pretrained on large-scale data, on a smaller, concept-specific set to capture that unseen concept.
  - Specifically, DreamBooth is optimized via:

$$\mathcal{L}_{ldm}(x_0;\theta_c) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(1,T)} \|\epsilon - \epsilon_{\theta_c}(z_t, t, \tau_{\theta_c}(y))\|_2^2 , \tag{1}$$

$$\begin{aligned}\mathcal{L}_{db}(x_0;\theta_c) &= \mathcal{L}_{ldm}(x_0;\theta_c)\\ &+ \lambda \underbrace{\mathbb{E}_{\epsilon',t'}\|\epsilon' - \epsilon_{\theta_c}(z_{t'}^{pr}, t', \tau_{\theta_c}(y^{pr}))\|_2^2}_{\text{Class-Specific Prior Preservation Loss}}, \quad (14)\end{aligned}$$



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR'23

Preliminaries – From Adversarial Attacks to Anti-Customization

□ Let’s revisit the strongest first-order attack method: Projected Gradient Descent (PGD)

- PGD is formalized as:

$$x^{adv}_{t+1} = \Pi_{x_0,\eta}\left(x^{adv}_t + \alpha \cdot \mathrm{sgn}\left(\nabla_{x^{adv}_t}\mathcal{L}(x^{adv}_t, y; \theta)\right)\right), \tag{16}$$

- Anti-customization utilizes adversarial attacks against generation, aiming to distort the concepts learned during fine-tuning by injecting protective perturbation delta. For the optimal solution, this presents a saddle point problem:

$$\delta^{adv} = \underset{\|\delta\|_\infty \le \eta}{\arg\max} \min_{\theta_c} \mathbb{E}_x \mathcal{L}_{ldm}(x_0 + \delta; \theta_c), \tag{2}$$

$$\Downarrow$$

$$\delta^{adv} = \underset{\|\delta\|_\infty \le \eta}{\arg\max} \mathcal{L}_{ldm}(x_0 + \delta; \theta_c), \tag{2}$$

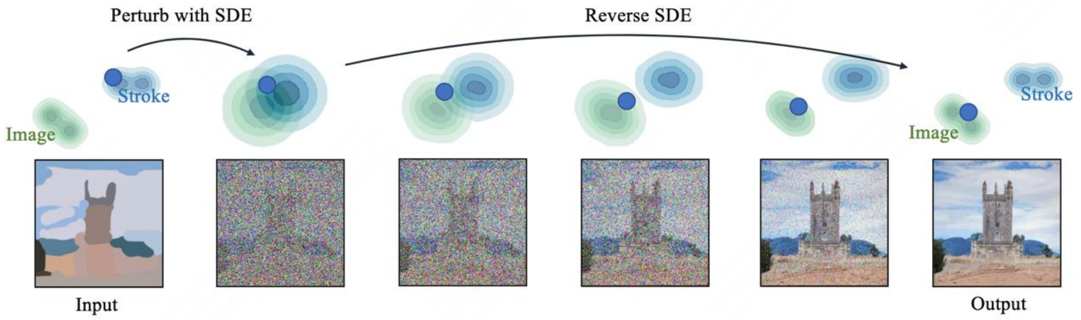
- But this can be simplified. The key lies in the relationship between the model's training data and the adversarial data.
  - ◆ For optimal performance, the training set should encompass adequately trained adversarial samples.
  - ◆ However, this creates a **bootstrap** paradox: fine-tuned theta is needed for optimal delta while delta is needed for optimal theta.
  - ◆ Thus, surrogate models fine-tuned on clean data are frequently employed for simplification.

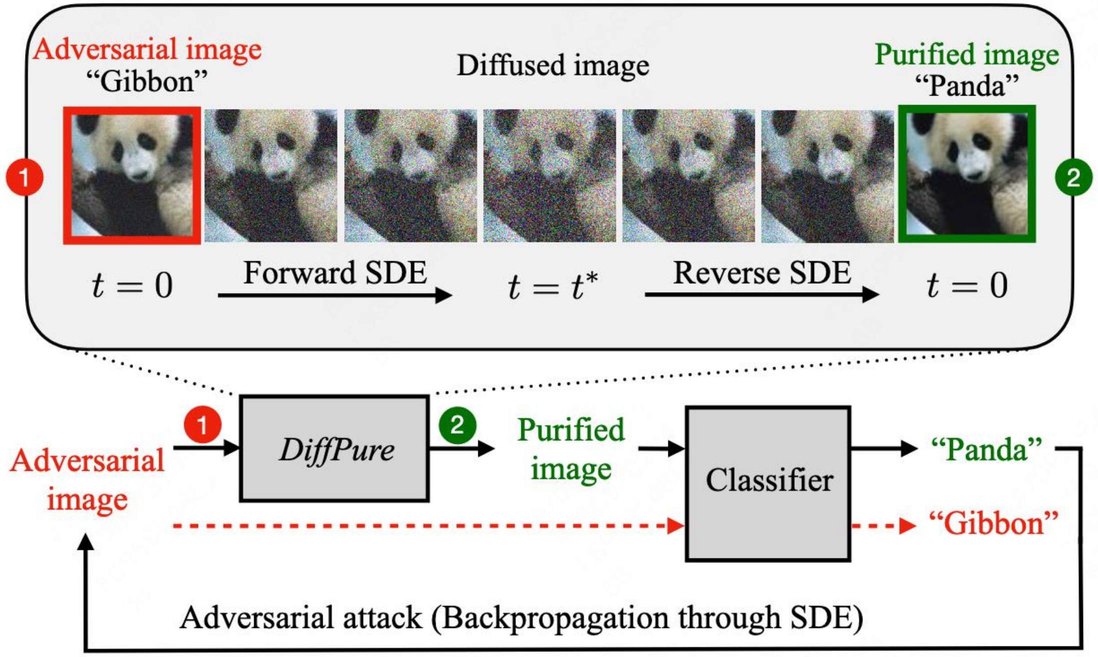
Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR’18
Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. ICCV’23

Preliminaries – Diffusion-based Purification

- **Let's revisit the pioneering diffusion-based purification method: DiffPure**
  - Pretrained unconditional diffusion models, e.g., DDPMs, can be inherently used for purification since the distributions of clean and adversarial samples converge over time during forward diffusion.
  - DiffPure diffuses the input adversarial image at timestep t^p and denoises it back to a purified image. In simplified discrete DDPM form, this can be written as:

$$\mathrm{Pure}(x^{adv}) = \mathrm{Reverse}(\sqrt{\overline{\alpha}_{t^p}}(x^{adv}) + \sqrt{1-\overline{\alpha}_{t^p}}\epsilon, t^p, 0; \theta_p), \quad (3)$$





SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR'22
Diffusion Models for Adversarial Purification. ICML'22

Analysis – Anti-purification: Overall Formulation

- **For ideal perturbations resistant to purification, we first formalize our objective as:**

  - However, direct backpropagation is computationally inefficient here.

$$\delta^{adv} = \arg\max_{\|\delta\|_\infty \le \eta} \min_{\theta_c} \mathbb{E}_x \mathcal{L}_{ldm}(x_0 + \delta; \theta_c), \qquad (2)$$

$$\delta^{adv} = \arg\max_{\|\delta\|_\infty \le \eta} \min_{\theta_c} \mathbb{E}_x \mathcal{L}_{ldm}(\mathrm{Pure}(x_0 + \delta); \theta_c), \qquad (4)$$

- **Alternatively, we decompose into stages. Interestingly, two opposing objectives can accomplish this:**

  - By Eq.5, we can approximate Pure(x) ≈ x, allowing Eq.4 to degenerate into Eq.2 even under purification.

  - By Eq.6, we resort to direct attacks against purification, i.e., **anti-purification**.

$$\delta^{adv\prime}_{min} = \arg\min_{\|\delta\|_\infty \le \eta} \| \mathrm{Pure}(x_0 + \delta) - (x_0 + \delta) \|_\infty, \text{ or} \qquad (5)$$

$$\delta^{adv\prime}_{max} = \arg\max_{\|\delta\|_\infty \le \eta} \| \mathrm{Pure}(x_0 + \delta) - (x_0 + \delta) \|_\infty . \qquad (6)$$

- **We prefer Eq.6 rather than Eq.5, WHY?**

Analysis – Anti-purification: Overall Formulation

- **Eq.5 follows a paradigm called <u>Adaptive Attacks</u>. However, that is unlikely to work in the context of Probabilistic Modeling.**
  - The difference between the clean and adversarial images (which nearly overlap) is far smaller than the range of purified outputs, and the distributions of the purified clean and adversarial images converge as t^p increases.
  - In conclusion, we observe that probabilistic models produce outputs that can become highly **unpredictable at the fine scale required by adversarial attacks**, thereby diminishing the effectiveness of adaptive attacks.

$$\delta^{adv\prime}_{min} = \arg\min_{\|\delta\|_\infty \le \eta} \| \, \mathrm{Pure}(x_0+\delta) - (x_0+\delta)\|_\infty, \text{ or} \qquad (5)$$

$$\delta^{adv\prime}_{max} = \arg\max_{\|\delta\|_\infty \le \eta} \| \, \mathrm{Pure}(x_0+\delta) - (x_0+\delta)\|_\infty . \qquad (6)$$

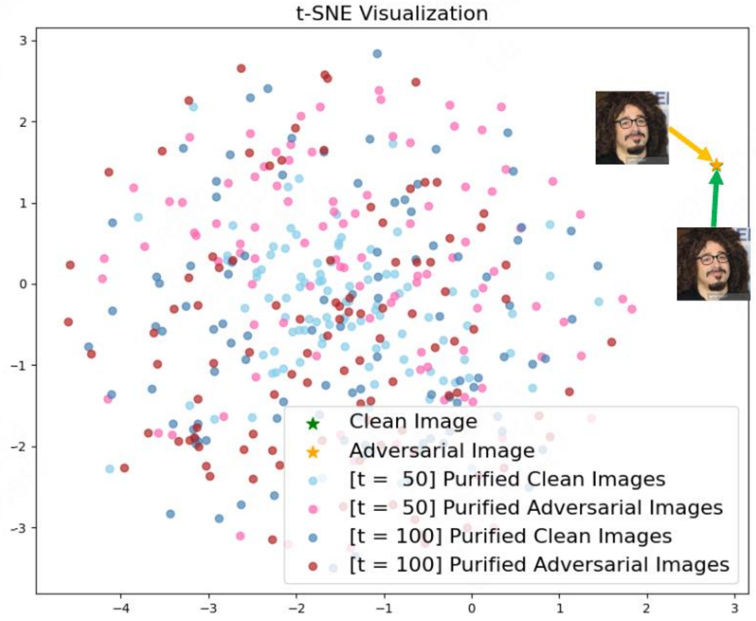


Figure 2. t-SNE [50] visualization (perplexity = 10) of 4×100 purified images obtained using DiffPure [31] with different timesteps for clean and adversarial images.

Analysis – Anti-purification: Overall Formulation

□ Now we choose Eq.6 as our objective, i.e., we want the outputs of purification to be distorted as much as possible.

□ A natural idea is to transfer adversarial attacks from anti-customization to anti-purification.

● **But direct adaptation also fails, ...WHY?**

Anti-Customization

$$\mathcal{L}_{ldm}(x_0;\theta_c) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(1,T)} \left\|\epsilon - \epsilon_{\theta_c}(z_t,t,\tau_{\theta_c}(y))\right\|_2^2, \tag{1}$$

$$\delta^{adv} = \underset{\|\delta\|_\infty\leq\eta}{\arg\max}\,\min_{\theta_c}\mathbb{E}_x\mathcal{L}_{ldm}(x_0+\delta;\theta_c), \tag{2}$$

Anti-Purification

$$\mathcal{L}_{ddpm}(x_0;\theta_p) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(1,T)} \left\|\epsilon - \epsilon_{\theta_p}(x_t,t)\right\|_2^2. \tag{7}$$

$$\delta^{adv} = \underset{\|\delta\|_\infty\leq\eta}{\arg\max}\,\min_{\theta_c}\mathbb{E}_x\mathcal{L}_{ldm}(x_0+\delta;\theta_c), \tag{2}$$

Analysis – Anti-purification: Why Harder?

- Through experiments, we analyze the differences between anti-customization and anti-purification, identifying three core characteristics of purification models that make anti-purification more challenging:
  - 1) lack of vulnerable network components,
  - 2) training-free frozen parameters, and
  - 3) fixed high-timestep denoising.

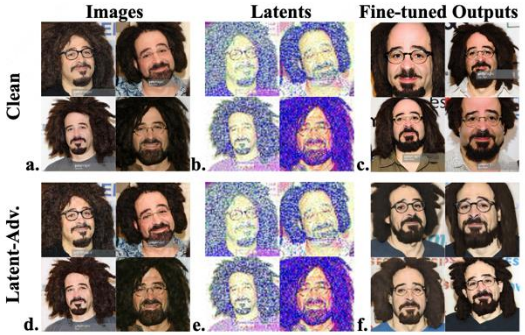


Figure 3. Attacks against DreamBooth [38] on UNet are much harder. Unlike vanilla pixel-space attacks (*a.* → *d.*), latent-space attacks (*b.* → *e.*) cannot target the vulnerable VAE encoder. Here, *d.* (decoded from *e.*) is shown for visualization purposes only; in our experiments, we directly replace *b.* with *e.* during fine-tuning.

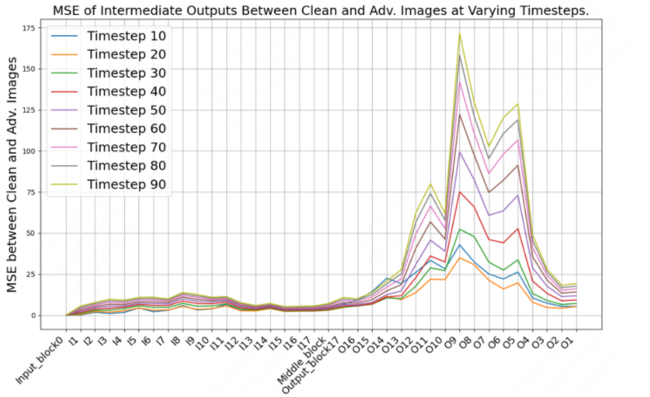


Figure 4. Mean Squared Errors of intermediate outputs between clean and adversarial samples across different UNet blocks at varying timesteps. See Appendix B.2 for details.

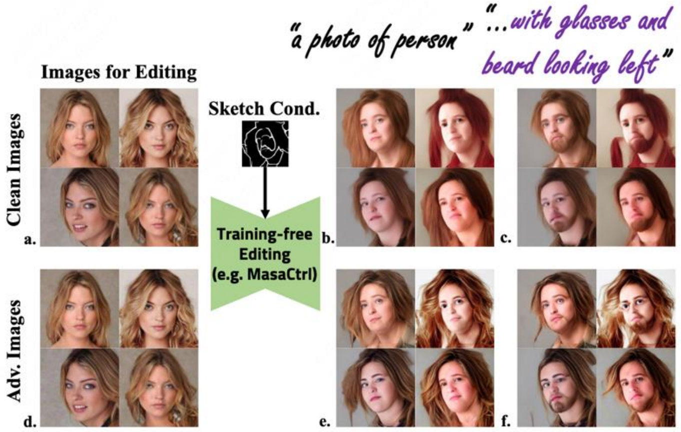


Figure 5. Effectiveness of MasaCtrl [3] on adversarial images. The loss attack makes little difference except that the lower right image of *e.* and *f.* has slight noise in the background.

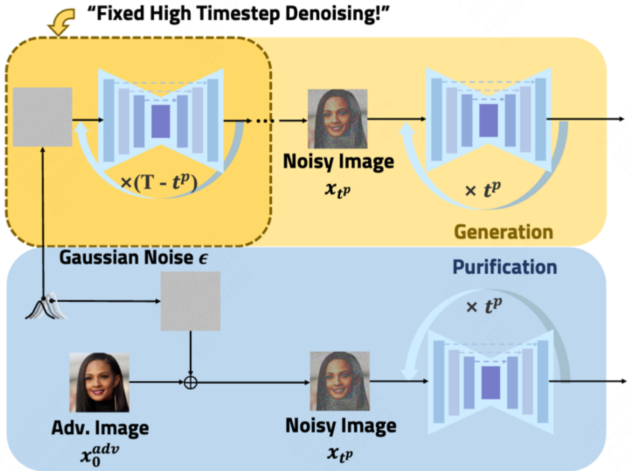
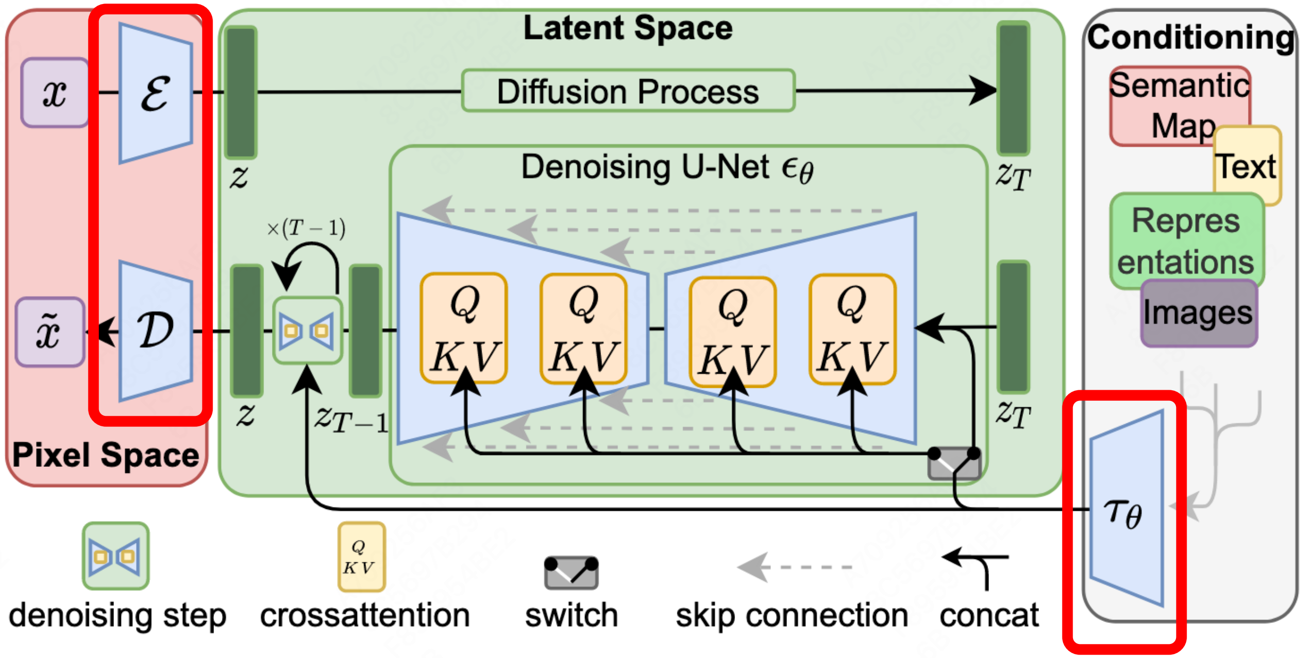


Figure 11. Why "the purification process can be viewed as a generation process where high-timestep denoising is fixed."

Analysis – Anti-purification: Why Harder?

Reason 1: Lack of Vulnerable Components

- Attacks targeting LDMs/SD are easier due to their more vulnerable encoders. In contrast, the only component in DDPMs, the UNet, is extremely robust.



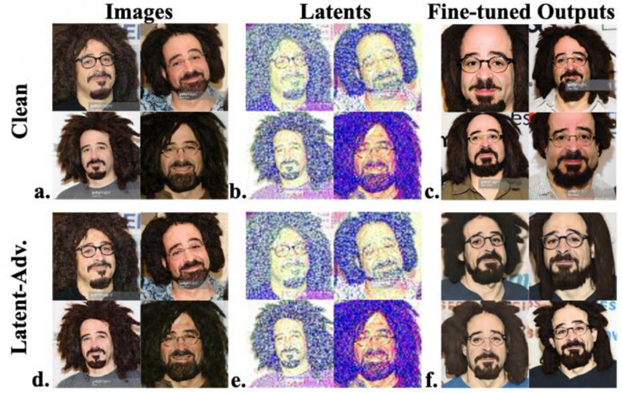


Figure 3. Attacks against DreamBooth [38] on UNet are much harder. Unlike vanilla pixel-space attacks (*a.* → *d.*), latent-space attacks (*b.* → *e.*) cannot target the vulnerable VAE encoder. Here, *d.* (decoded from *e.*) is shown for visualization purposes only; in our experiments, we directly replace *b.* with *e.* during fine-tuning.

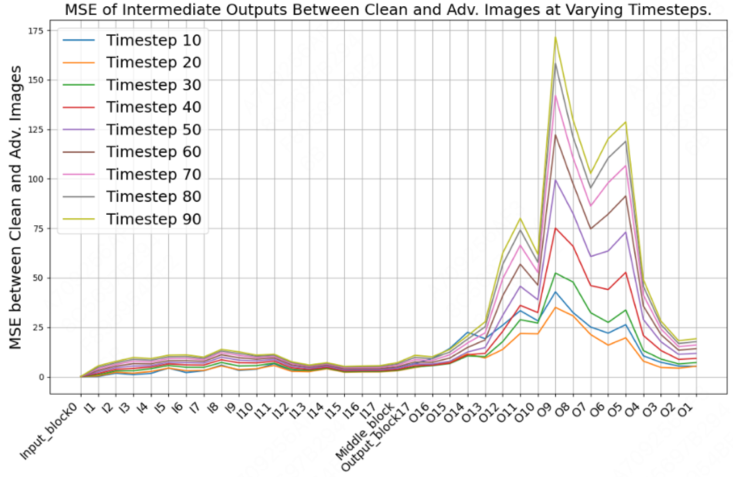


Figure 4. Mean Squared Errors of intermediate outputs between clean and adversarial samples across different UNet blocks at varying timesteps. See Appendix B.2 for details.

Analysis – Anti-purification: Why Harder?

- **Reason 2: Training-free Frozen Parameters**
  - Unlike anti-customization which targets fine-tuning by data poisoning, anti-purification targets a training-free editing task.
- **Reason 3: Fixed High Timestep Denoising**
  - The purification process can be viewed as a generation process where high-timestep denoising is fixed.
  - In cases where vulnerable components are absent and parameters are frozen, conducting a L_{ddpm}-based attack for timesteps beyond t^p is not directly meaningful, and attempting to achieve semantic structural changes by adjusting the input at low timesteps is also unfeasible.

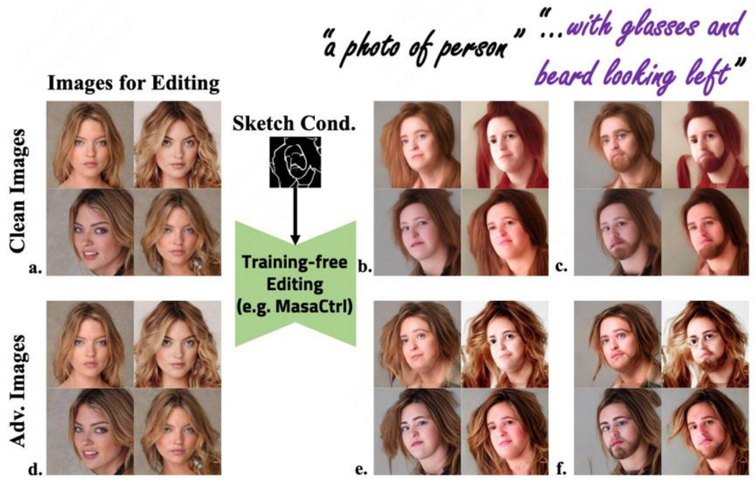


Figure 5. Effectiveness of MasaCtrl [3] on adversarial images. The loss attack makes little difference except that the lower right image of e. and f. has slight noise in the background.

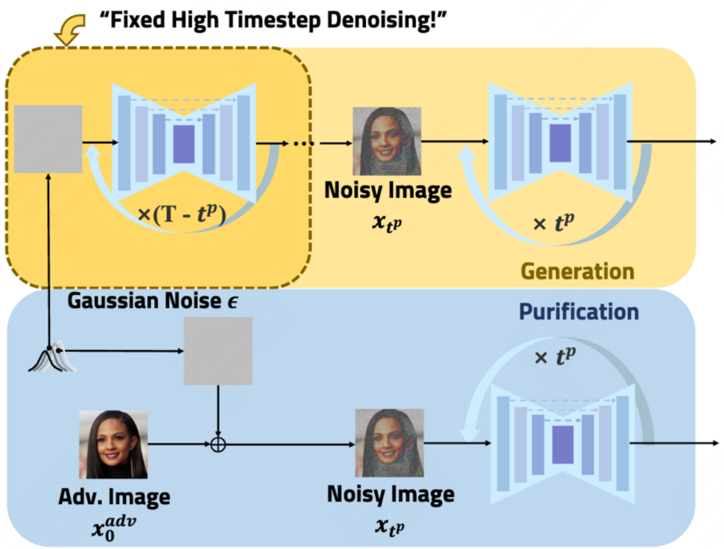


Figure 11. Why "the purification process can be viewed as a generation process where high-timestep denoising is fixed."

Method – AntiPure

Patch-wise Frequency Guidance (PFG)

- Unlike low-frequency semantic structures, consistency in high-frequency components is harder to guarantee, rendering them less controllable during purification.
- PFG aims to enhance the high-frequency components of the purification model's prediction, indirectly reinforcing the adv. perturbation's high-frequency elements.

$$x_t = \sqrt{\overline{\alpha}_t}(x_0 + \delta_i^{adv}) + \sqrt{1 - \overline{\alpha}_t}\epsilon, \tag{8}$$

$$\widehat{x}_0 = (x_t - \sqrt{1 - \overline{\alpha}_t}\epsilon_\theta(x_t, t))/\sqrt{\overline{\alpha}_t}. \tag{9}$$

$$\mathcal{L}_{fre}(x_0; \delta^{adv}) = \sigma(\mathbb{E}_P \frac{4}{s^2} \sum_{m,n=\frac{s}{2}}^{s-1} \mathrm{PatchDCT}(\widehat{x}_0, s)_{m,n}), \tag{10}$$

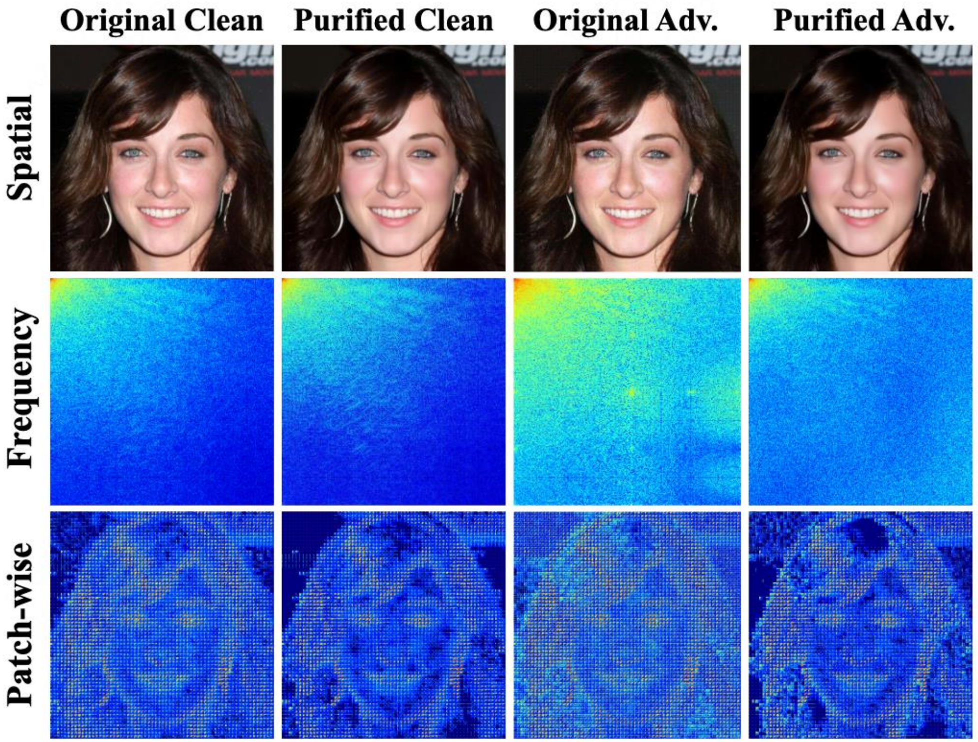


Figure 7. Differences in the spatial and frequency domains before and after DDPM-purification. Pseudocolor transformation is applied to the DCT spectrogram for better visualization.

Method – AntiPure

Erroneous Timestep Guidance (ETG)

- The structure of images cannot be obviously altered because they are fixed during high-timestep denoising.
- However, ETG can identify inputs for which the UNet struggles to select the appropriate actions across timesteps.

$$\mathcal{L}_{err-t}(x_0;\delta^{adv}) = -\left\|\epsilon_\theta(x_t, t_{err}) - \epsilon_\theta(x_t, t)\right\|_2^2 . \quad (11)$$

Overall Attack

$$\mathcal{L}_{pgd}(x_0;\delta^{adv}) = \mathbb{E}_{\epsilon,t}\left(\mathcal{L}_{ddpm} + \lambda_1 e^{\overline{\alpha}_t - 1}\mathcal{L}_{fre} + \lambda_2 e^{\mathcal{L}_{err-t}}\right), \quad (12)$$

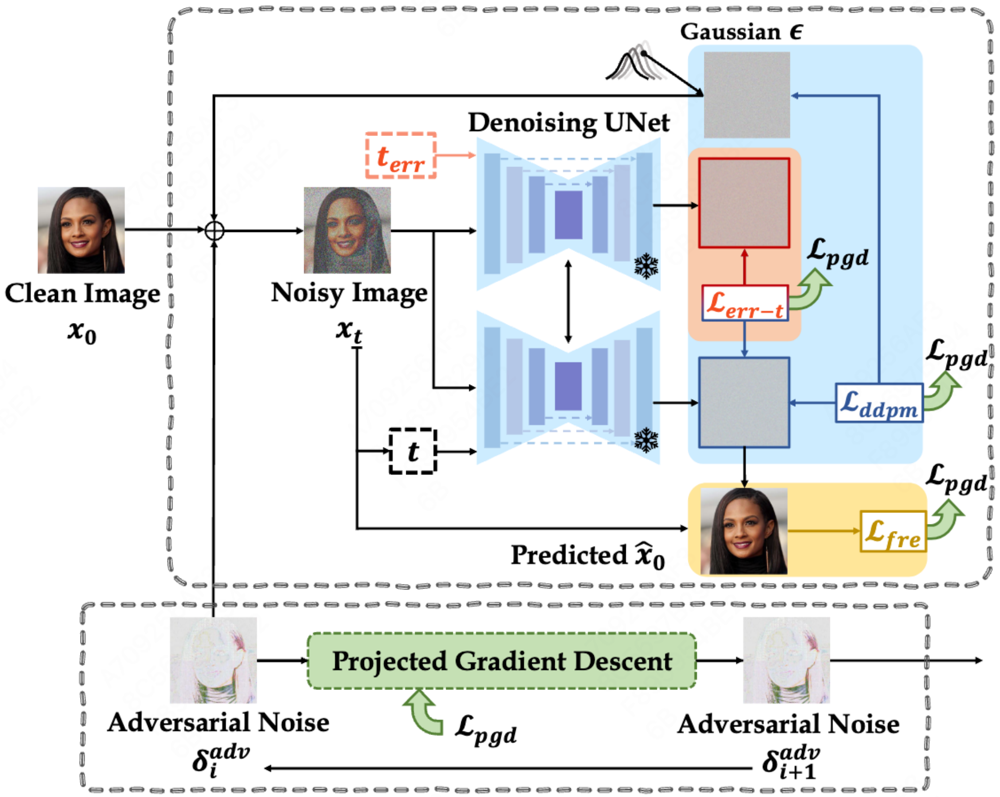
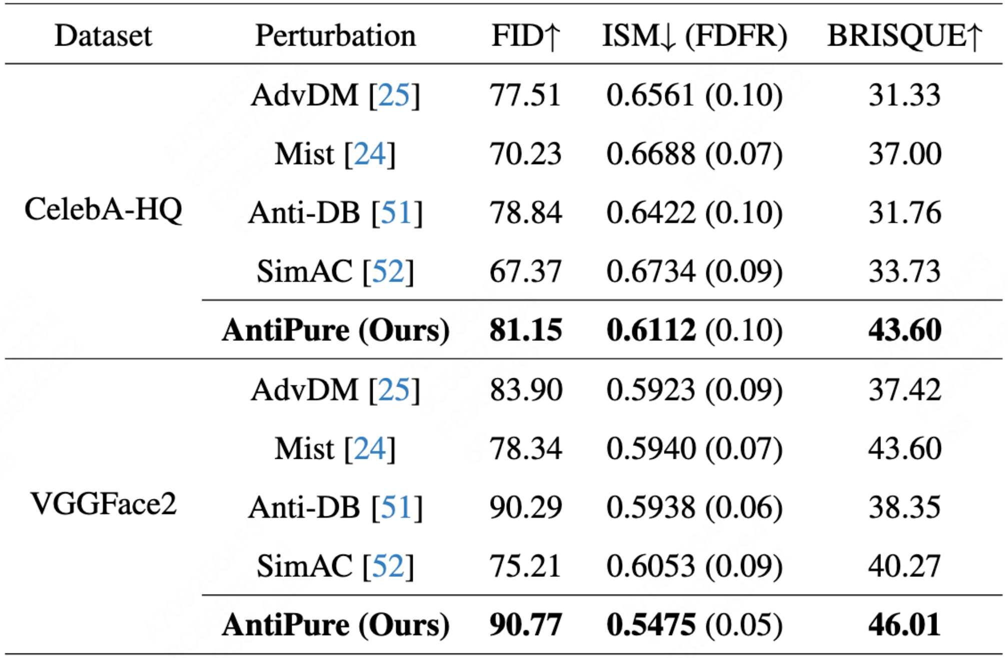


Figure 6. Overview diagram of AntiPure. AntiPure mitigates the negative impact of benign priors (Sec. 4.2.2) and limited timesteps (Sec. 4.2.3) on the attack by introducing Patch-wise Frequency Guidance $\mathcal{L}_{fre}$ and Erroneous Timestep Guidance $\mathcal{L}_{t-err}$.

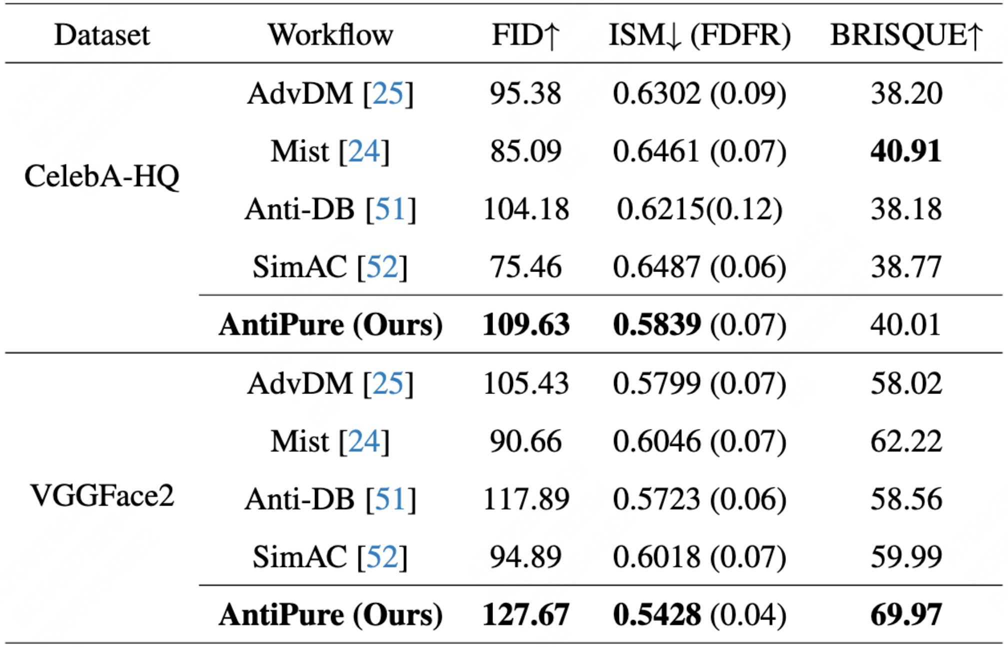
Experiments

Quantitative Results [1]



Dataset	Perturbation	FID↑	ISM↓ (FDFR)	BRISQUE↑
CelebA-HQ	AdvDM [25]	77.51	0.6561 (0.10)	31.33
	Mist [24]	70.23	0.6688 (0.07)	37.00
	Anti-DB [51]	78.84	0.6422 (0.10)	31.76
	SimAC [52]	67.37	0.6734 (0.09)	33.73
	AntiPure (Ours)	**81.15**	**0.6112** (0.10)	**43.60**
VGGFace2	AdvDM [25]	83.90	0.5923 (0.09)	37.42
	Mist [24]	78.34	0.5940 (0.07)	43.60
	Anti-DB [51]	90.29	0.5938 (0.06)	38.35
	SimAC [52]	75.21	0.6053 (0.09)	40.27
	AntiPure (Ours)	**90.77**	**0.5475** (0.05)	**46.01**

Table 1. Comparison of DreamBooth's [38] output quality for different perturbation methods following the P-C workflow.

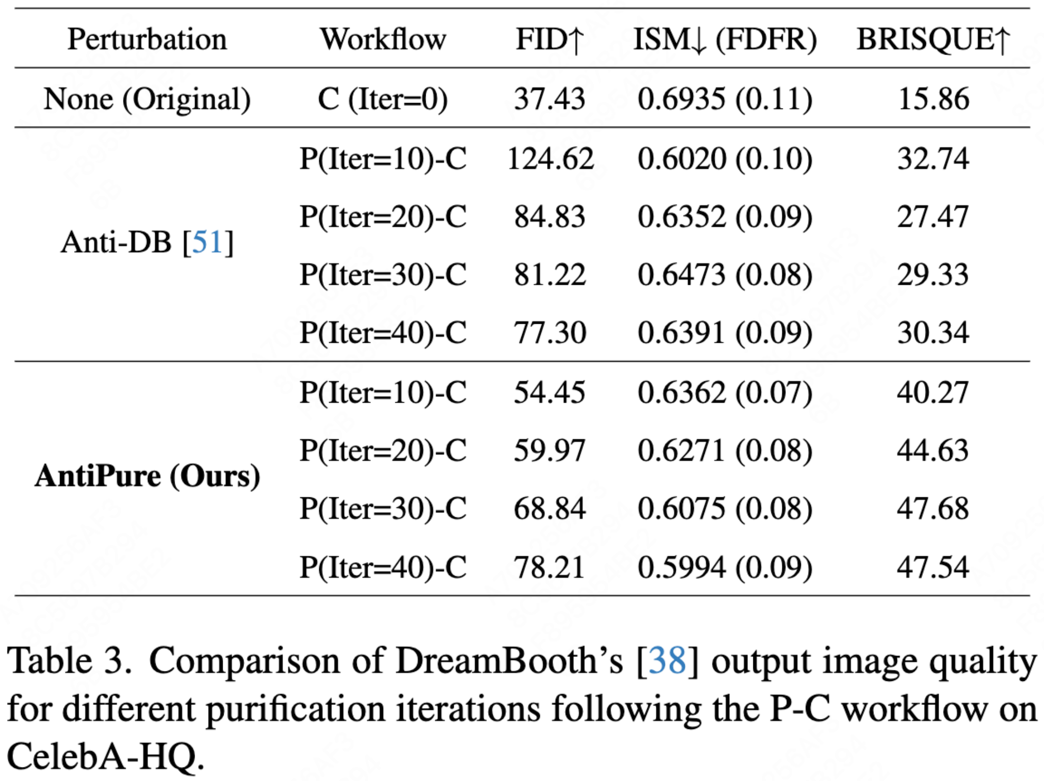


Dataset	Workflow	FID↑	ISM↓ (FDFR)	BRISQUE↑
CelebA-HQ	AdvDM [25]	95.38	0.6302 (0.09)	38.20
	Mist [24]	85.09	0.6461 (0.07)	**40.91**
	Anti-DB [51]	104.18	0.6215(0.12)	38.18
	SimAC [52]	75.46	0.6487 (0.06)	38.77
	AntiPure (Ours)	**109.63**	**0.5839** (0.07)	40.01
VGGFace2	AdvDM [25]	105.43	0.5799 (0.07)	58.02
	Mist [24]	90.66	0.6046 (0.07)	62.22
	Anti-DB [51]	117.89	0.5723 (0.06)	58.56
	SimAC [52]	94.89	0.6018 (0.07)	59.99
	AntiPure (Ours)	**127.67**	**0.5428** (0.04)	**69.97**

Table 2. Comparison of LoRA's [17] output image quality for different perturbation methods following the P-C workflow.

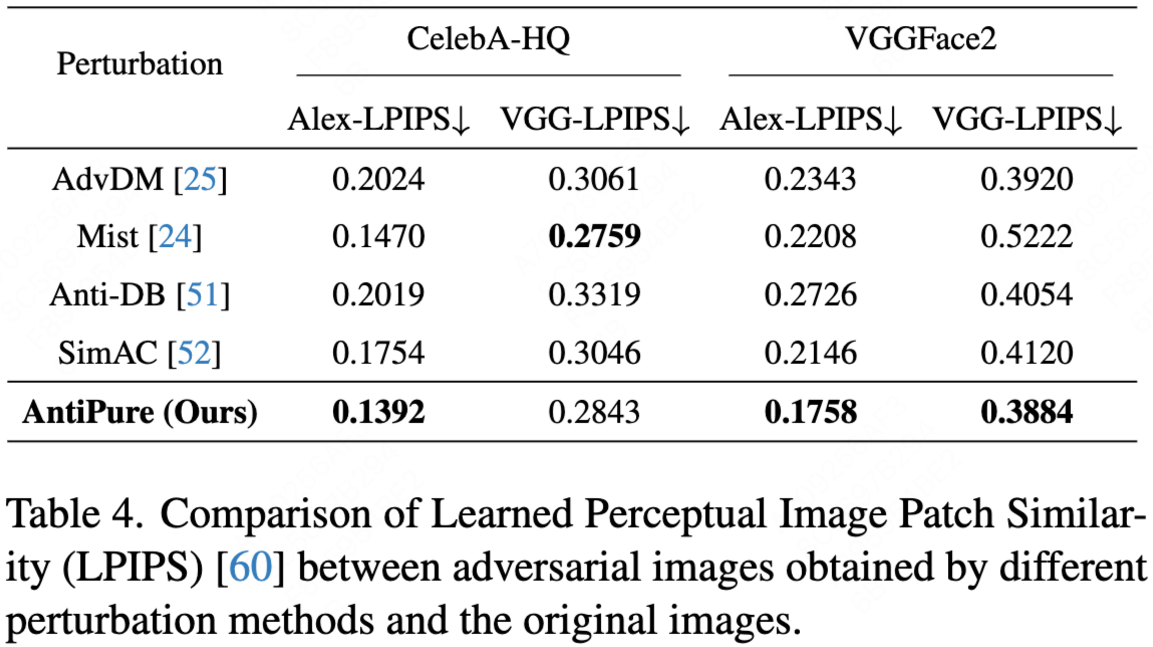
Experiments

Quantitative Results [2]



Perturbation	Workflow	FID↑	ISM↓ (FDFR)	BRISQUE↑
None (Original)	C (Iter=0)	37.43	0.6935 (0.11)	15.86
Anti-DB [51]	P(Iter=10)-C	124.62	0.6020 (0.10)	32.74
	P(Iter=20)-C	84.83	0.6352 (0.09)	27.47
	P(Iter=30)-C	81.22	0.6473 (0.08)	29.33
	P(Iter=40)-C	77.30	0.6391 (0.09)	30.34
AntiPure (Ours)	P(Iter=10)-C	54.45	0.6362 (0.07)	40.27
	P(Iter=20)-C	59.97	0.6271 (0.08)	44.63
	P(Iter=30)-C	68.84	0.6075 (0.08)	47.68
	P(Iter=40)-C	78.21	0.5994 (0.09)	47.54

Table 3. Comparison of DreamBooth's [38] output image quality for different purification iterations following the P-C workflow on CelebA-HQ.

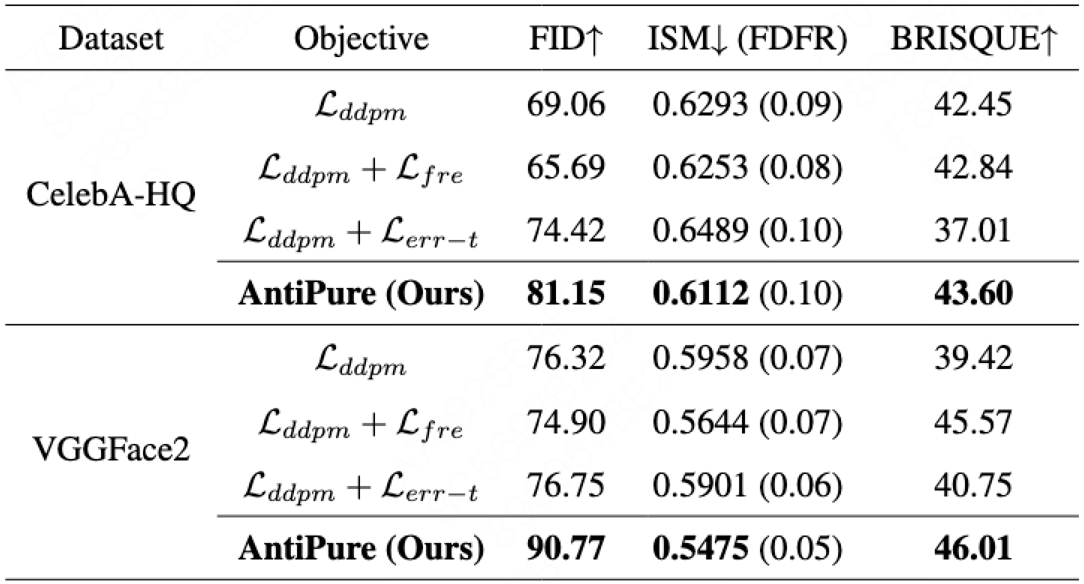


Perturbation	CelebA-HQ		VGGFace2	
	Alex-LPIPS↓	VGG-LPIPS↓	Alex-LPIPS↓	VGG-LPIPS↓
AdvDM [25]	0.2024	0.3061	0.2343	0.3920
Mist [24]	0.1470	**0.2759**	0.2208	0.5222
Anti-DB [51]	0.2019	0.3319	0.2726	0.4054
SimAC [52]	0.1754	0.3046	0.2146	0.4120
AntiPure (Ours)	**0.1392**	0.2843	**0.1758**	**0.3884**

Table 4. Comparison of Learned Perceptual Image Patch Similarity (LPIPS) [60] between adversarial images obtained by different perturbation methods and the original images.

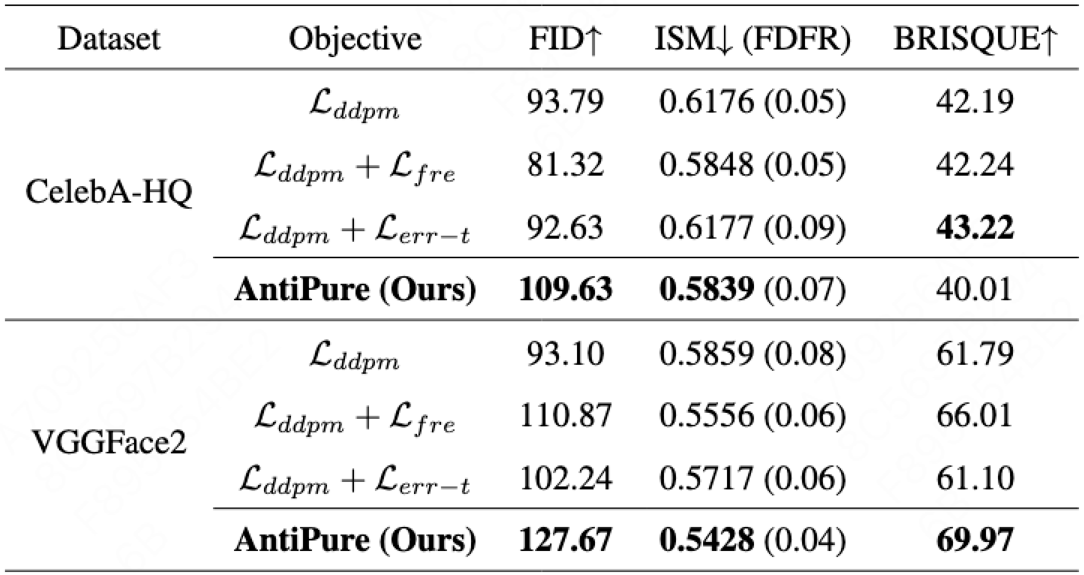
Experiments

Ablation Studies



Dataset	Objective	FID↑	ISM↓ (FDFR)	BRISQUE↑
CelebA-HQ	$\mathcal{L}_{ddpm}$	69.06	0.6293 (0.09)	42.45
	$\mathcal{L}_{ddpm} + \mathcal{L}_{fre}$	65.69	0.6253 (0.08)	42.84
	$\mathcal{L}_{ddpm} + \mathcal{L}_{err-t}$	74.42	0.6489 (0.10)	37.01
	AntiPure (Ours)	**81.15**	**0.6112** (0.10)	**43.60**
VGGFace2	$\mathcal{L}_{ddpm}$	76.32	0.5958 (0.07)	39.42
	$\mathcal{L}_{ddpm} + \mathcal{L}_{fre}$	74.90	0.5644 (0.07)	45.57
	$\mathcal{L}_{ddpm} + \mathcal{L}_{err-t}$	76.75	0.5901 (0.06)	40.75
	AntiPure (Ours)	**90.77**	**0.5475** (0.05)	**46.01**

Table 6. Ablation Study on DreamBooth's [38] output quality for different AntiPure guidance following the Purification-Customization (P-C) workflow.



Dataset	Objective	FID↑	ISM↓ (FDFR)	BRISQUE↑
CelebA-HQ	$\mathcal{L}_{ddpm}$	93.79	0.6176 (0.05)	42.19
	$\mathcal{L}_{ddpm} + \mathcal{L}_{fre}$	81.32	0.5848 (0.05)	42.24
	$\mathcal{L}_{ddpm} + \mathcal{L}_{err-t}$	92.63	0.6177 (0.09)	**43.22**
	AntiPure (Ours)	**109.63**	**0.5839** (0.07)	40.01
VGGFace2	$\mathcal{L}_{ddpm}$	93.10	0.5859 (0.08)	61.79
	$\mathcal{L}_{ddpm} + \mathcal{L}_{fre}$	110.87	0.5556 (0.06)	66.01
	$\mathcal{L}_{ddpm} + \mathcal{L}_{err-t}$	102.24	0.5717 (0.06)	61.10
	AntiPure (Ours)	**127.67**	**0.5428** (0.04)	**69.97**

Table 7. Ablation Study on LoRA's [17] output quality for different AntiPure guidance following the Purification-Customization (P-C) workflow.

Experiments

□ Qualitative Results - Visualization

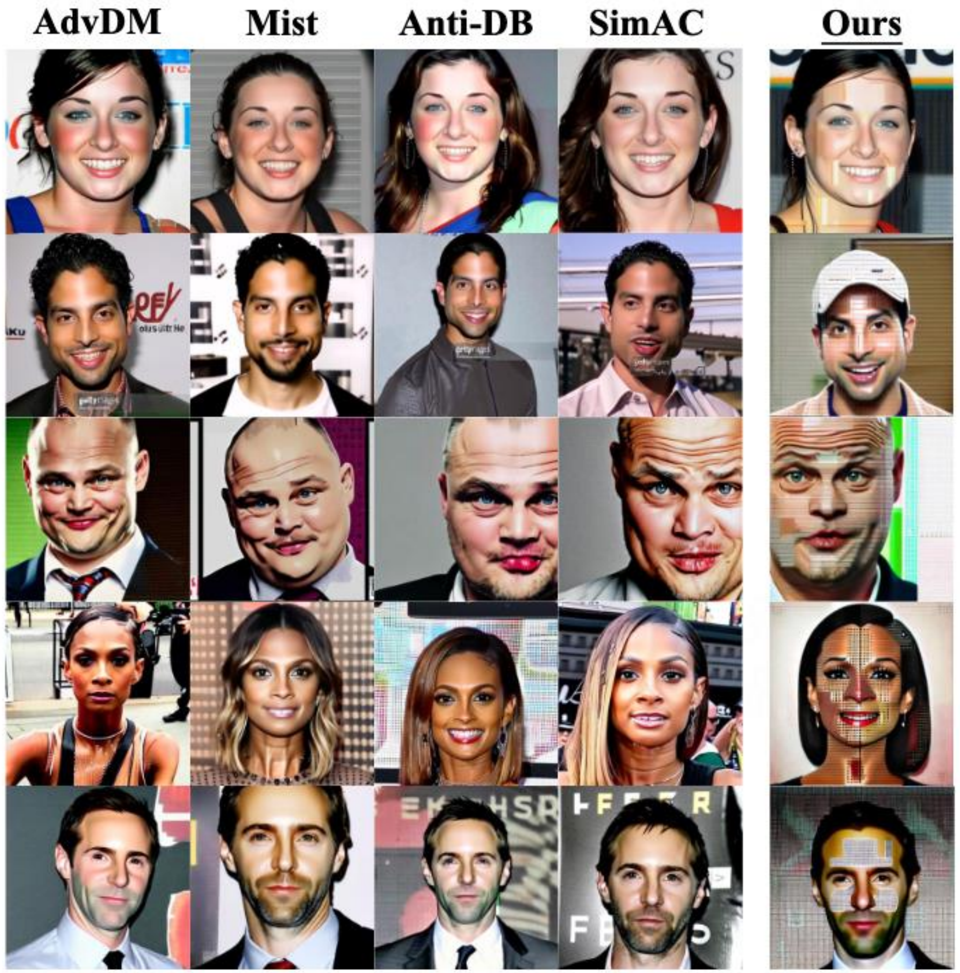


Figure 8. Visualization of DreamBooth's outputs after the P-C workflow.

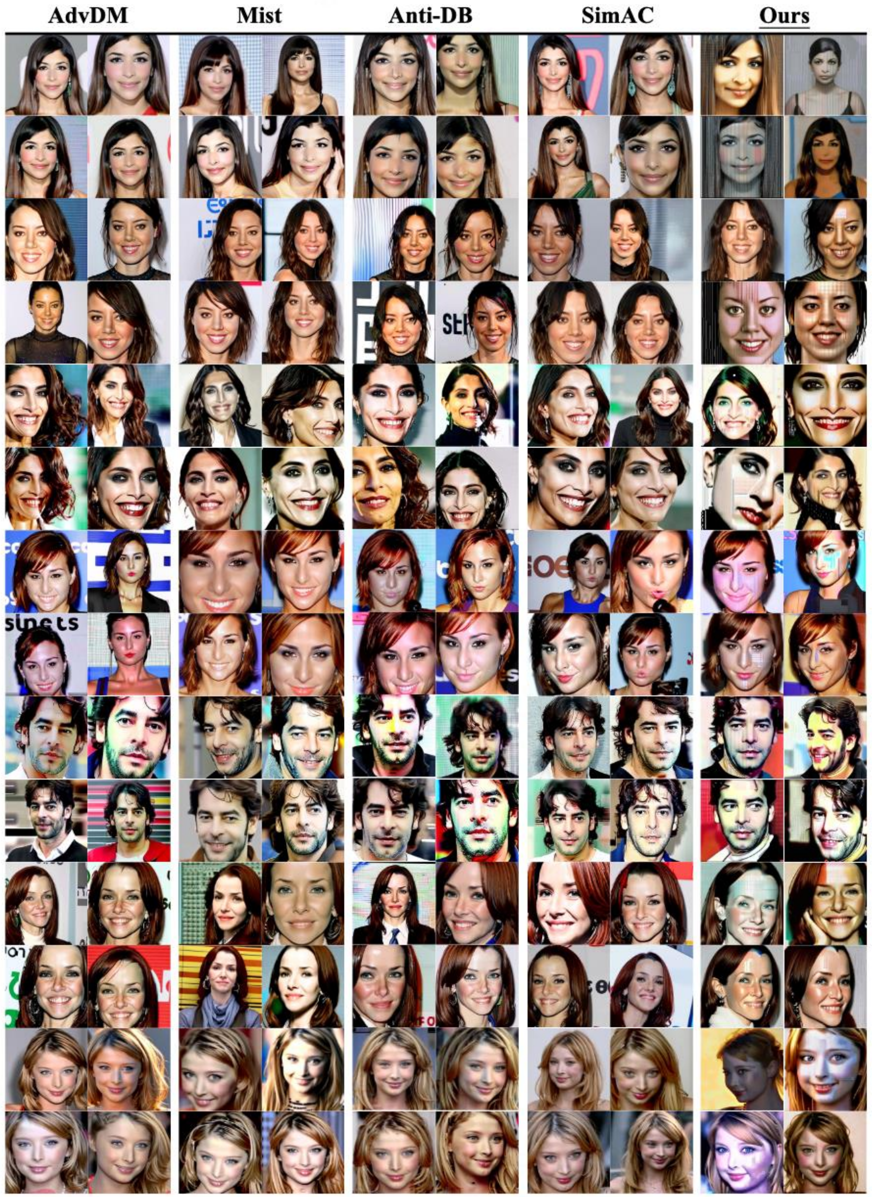
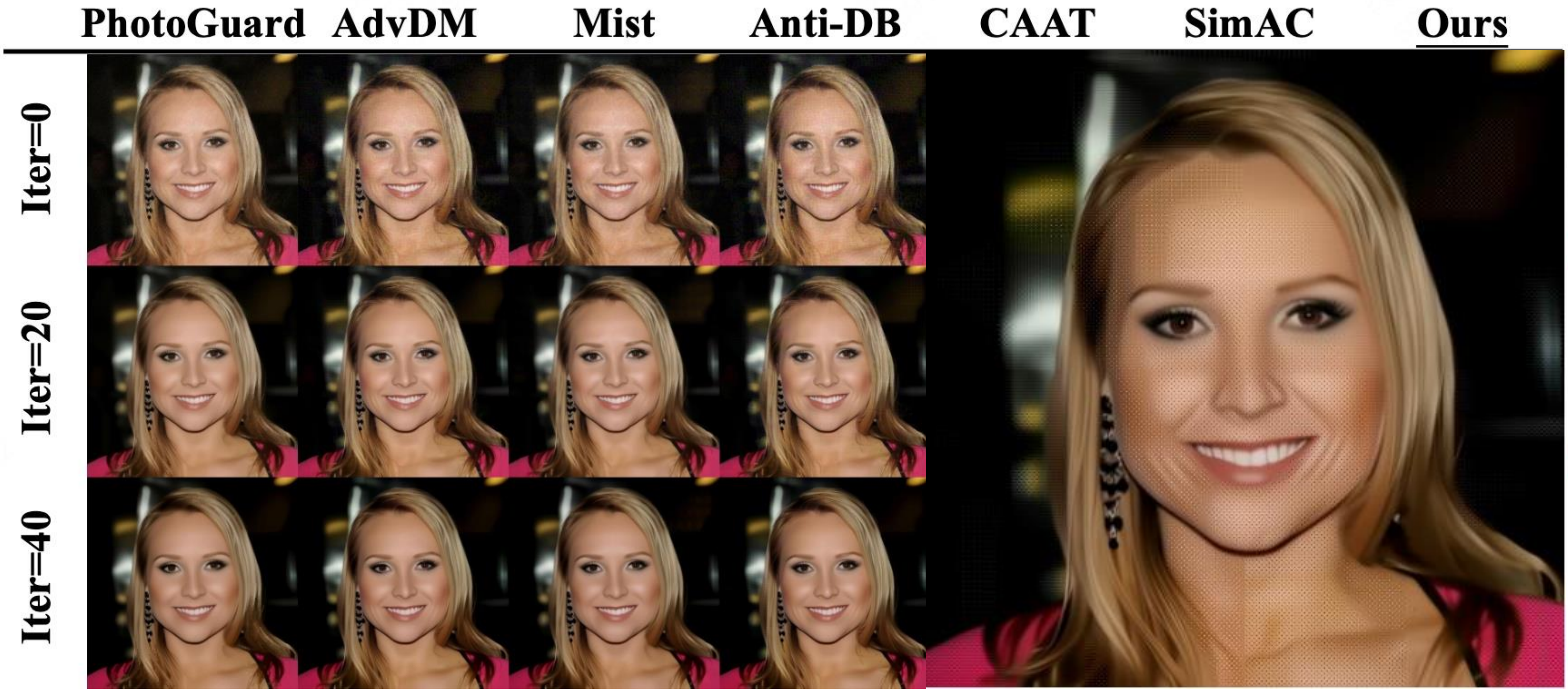


Figure 12. Comparison of DreamBooth's outputs on CelebA-HQ for different perturbation methods following the Purification-Customization (P-C) workflow.

Experiments

□ Qualitative Results - Visualization



Thanks