NDSS SYMPOSIUM/2025

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional

GAP-Diff: Protecting JPEG-Compressed Images from Diffusion-based Facial Customization

Haotian Zhu*, Shuchao Pang*[12], Zhigang Lu†[1], Yongbin Zhou*, and Minhui Xue‡

*
WESTERN SYDNEY UNIVERSITY
†
CSIRO DATA 61
‡

1: equal contribution
2: corresponding author

Outline

1. Background
2. Technical Challenges
   - JPEG Compression Effects
   - Existing Protection Limitations
3. Methodology
   - Generator Module
   - Pre-processing Simulation
   - Fine-tuning T2I-DM
4. Experimental Results
5. Conclusion

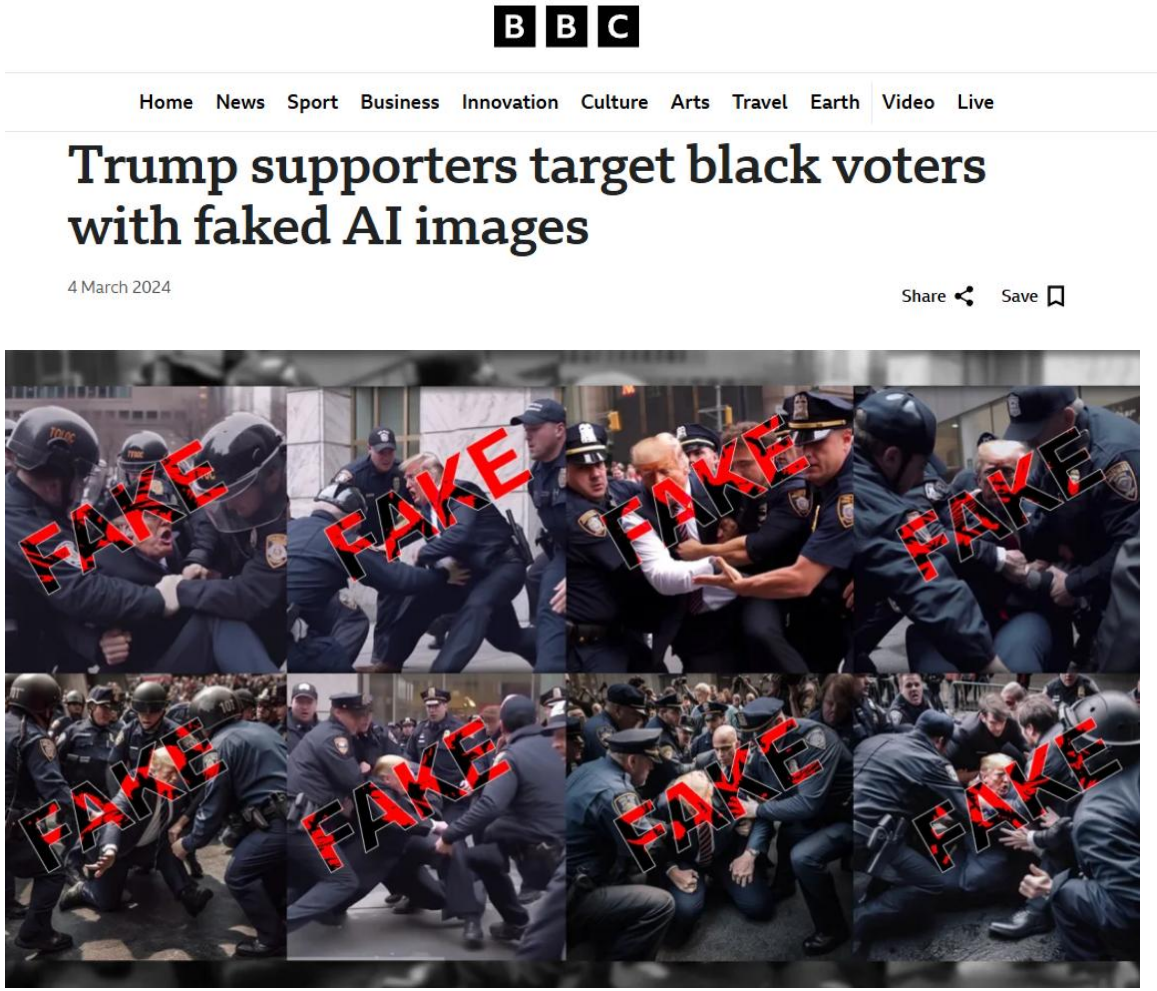
Background

Finetuned Text to Image Diffusion Models (FT-T2I-DMs) can generate customized photos using only 3-5 identity images

Easily obtain customized images through a FT-T2I-DM based on different prompts

- Potential misuse for creating fake content
- Privacy and security risks

Fake images generated by carefully designed customization



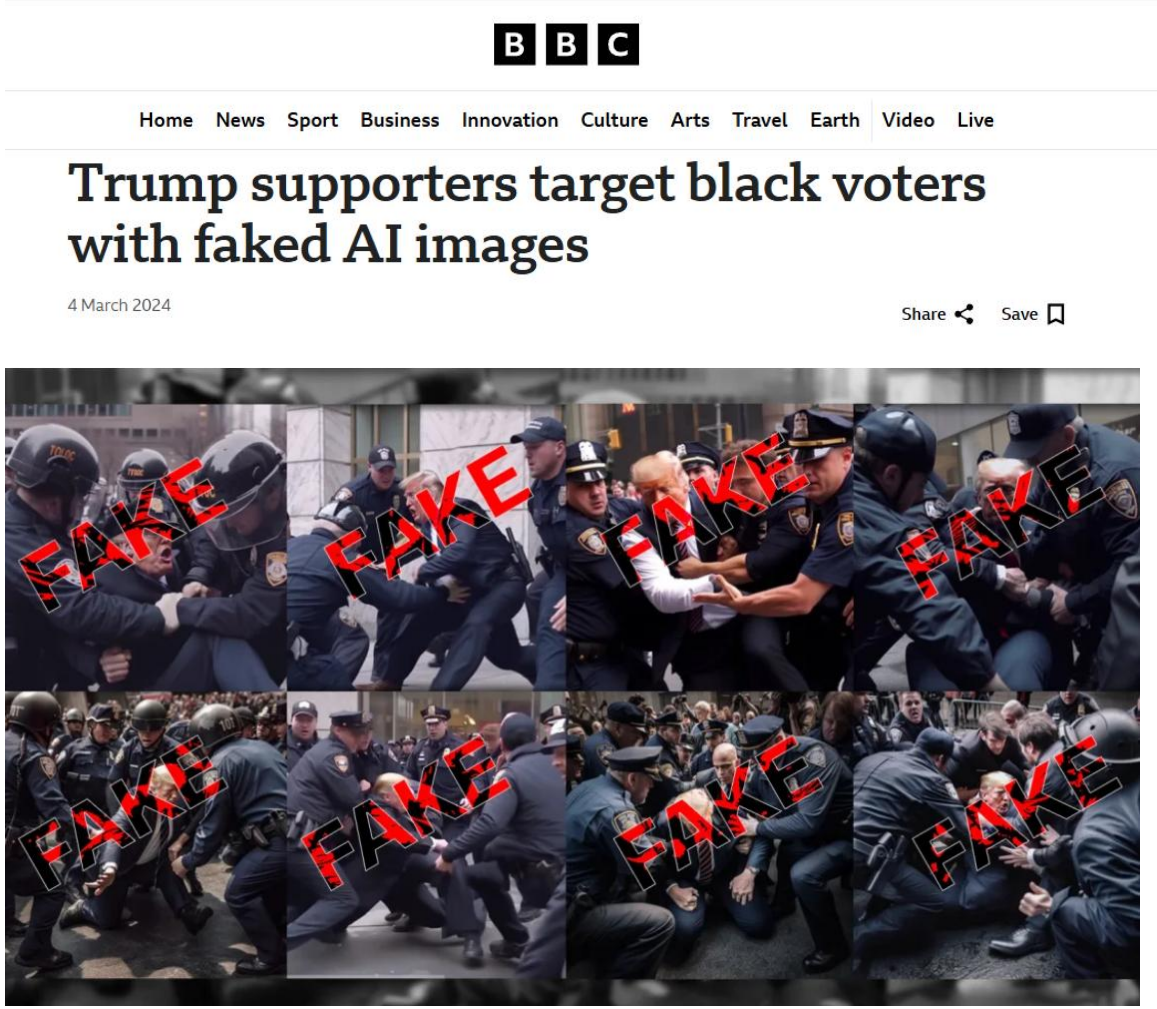
◆ Background

Finetuned Text to Image Diffusion Models (FT-T2I-DMs) can generate customized photos using only 3-5 identity images

Easily obtain customized images through a FT-T2I-DM based on different prompts

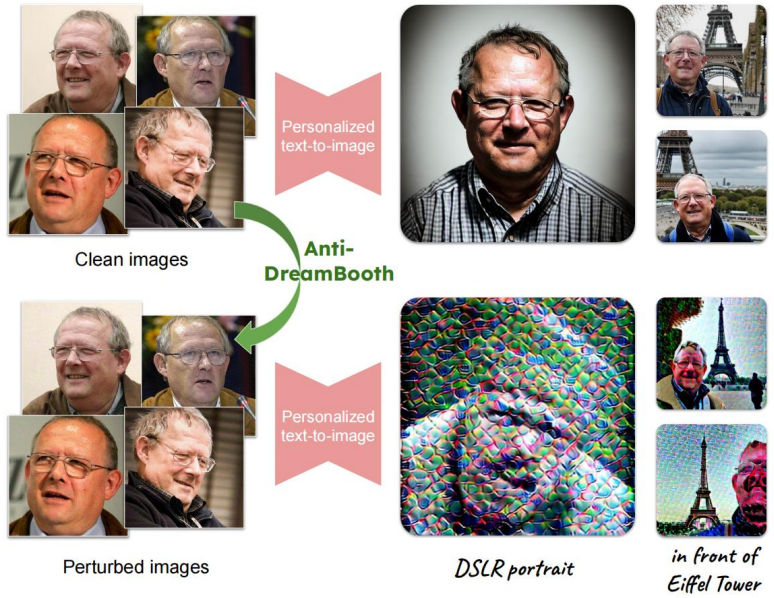
- How to protect facial privacy from diffusion-based customization?

Fake images caused by carefully designed customization



◆ Technical Challenges

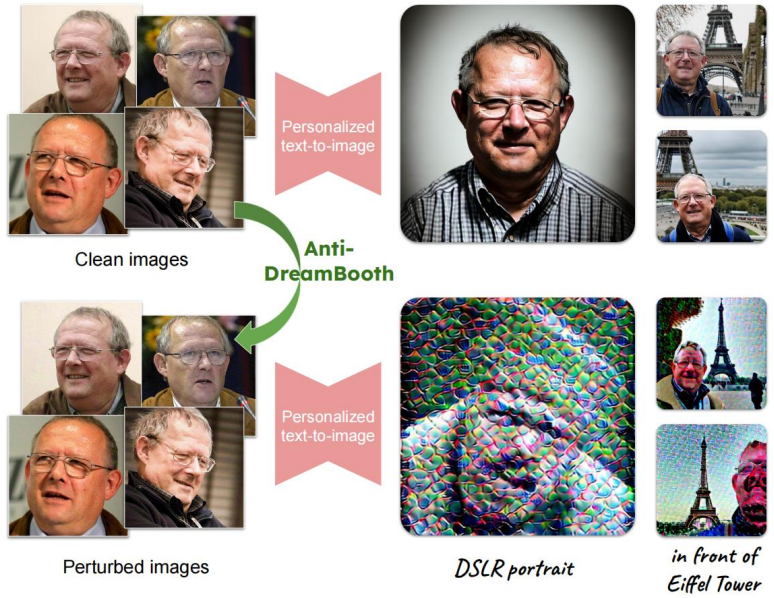
Adding protective noise to images



*Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2116–2127, 2023.

◆ Technical Challenges

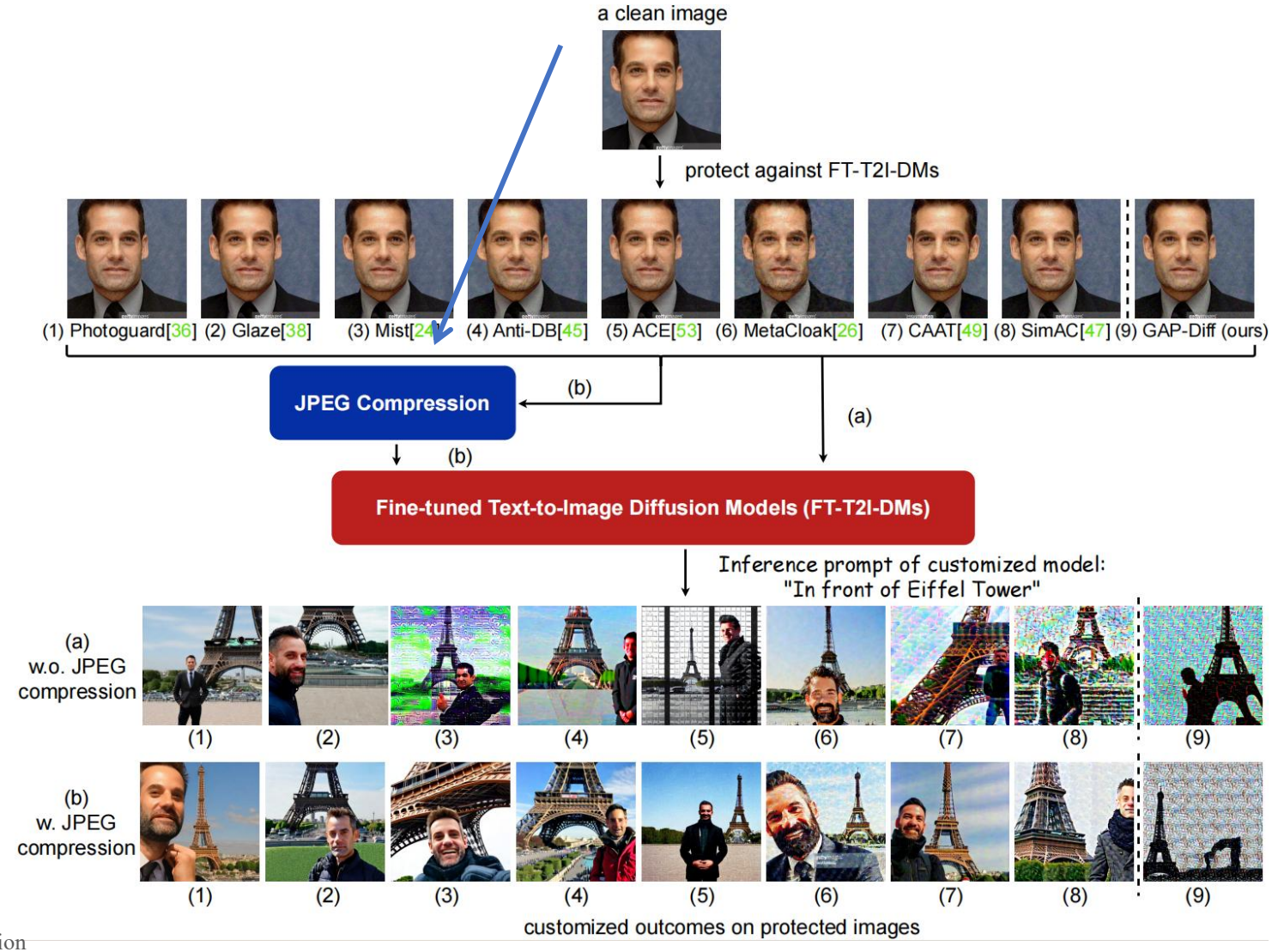
Adding protective noise to images



*Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2116–2127, 2023.

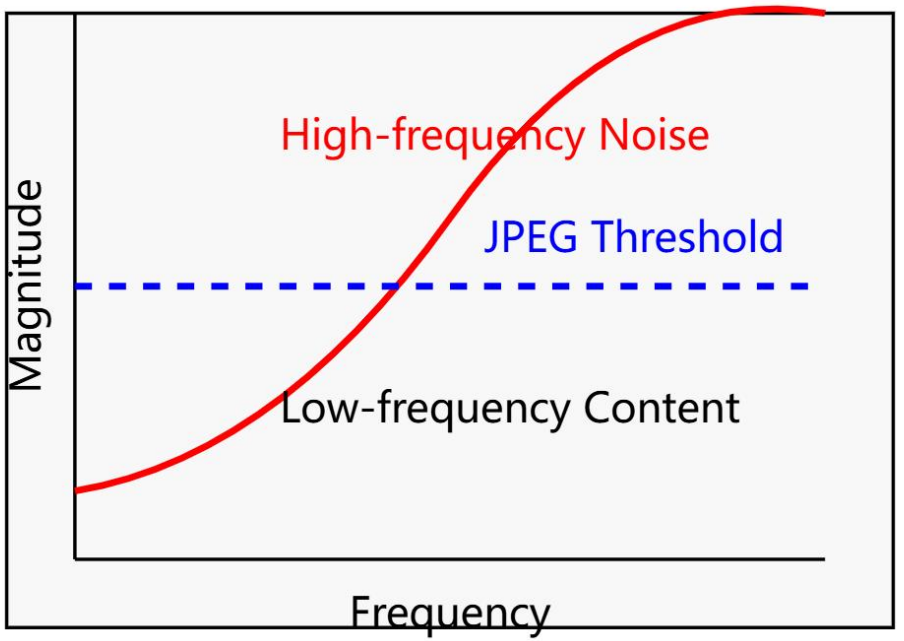
- How to protect facial privacy from model customization while maintaining resistance to JPEG compression?

Existing protections fail after JPEG compression



◆ Technical Challenges

Existing works often use the idea of PGD attacks to iteratively generate adversarial protective noise against customization.



Pixel-level Iterative Process

$$x^{k+1} = \prod_{(x,\eta)} (x^k + \alpha sgn(\nabla_x L(f(x+\delta, y_{true}))))$$

leads to sudden changes and sharp local differences between pixel values
==> a widely distributed high-frequency noise across the entire image

◆ Methodology

Possible approaches to consider:

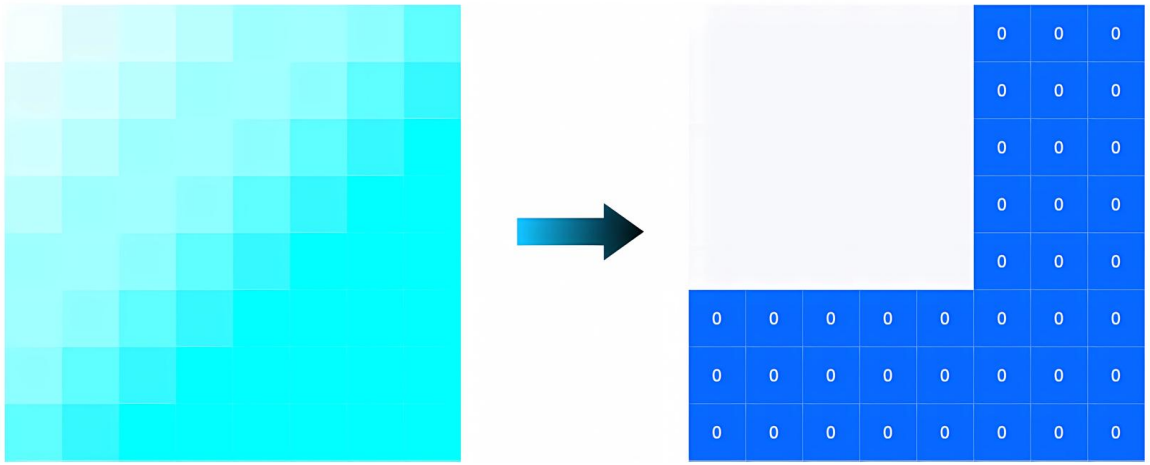
- an adaptive method to generate protective noise iteratively, focusing only on low-frequency regions that are less likely to be eliminated by JPEG compression.
- Use neural networks to learn the generation of noise, allowing the model to generate a protective noise pattern that is difficult to be removed by JPEG compression.

◆ Methodology

Possible approaches to consider:

- an adaptive method to generate protective noise iteratively, focusing only on low-frequency regions that are less likely to be eliminated by JPEG compression.
- Use neural networks to learn the generation of noise, allowing the model to generate a protective noise pattern that is difficult to be removed by JPEG compression.

A proposed differentiable JPEG simulation method:

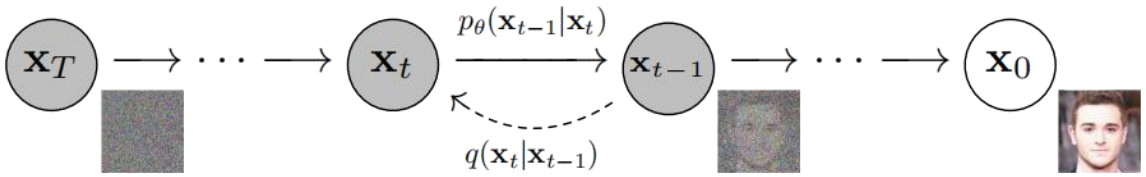


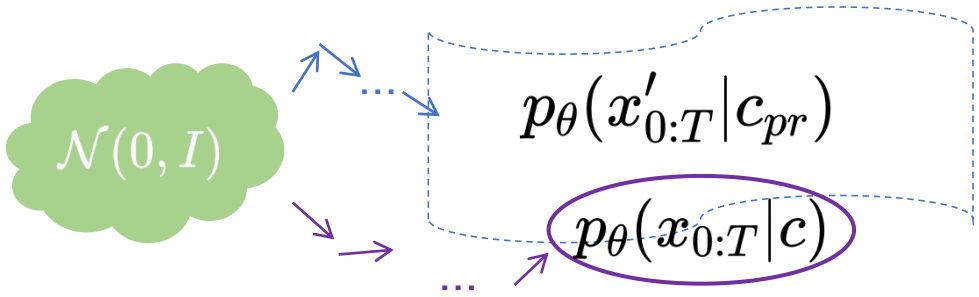
JPEG-Mask* simulates compression by:

- Retaining 5×5 low-frequency region (Y channel)
- Retaining 3×3 low-frequency region (U,V channels)
- Zeroing out other high-frequency coefficients

*Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 657–672, 2018.

◆ Methodology





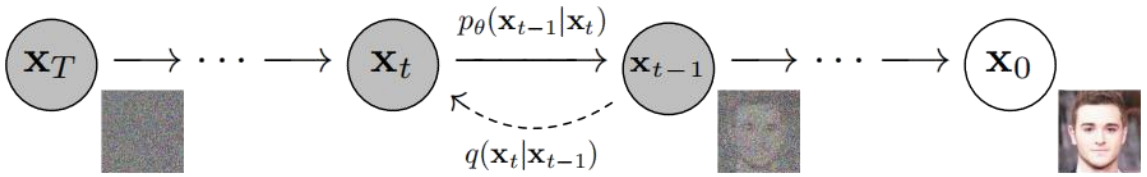
Fine-tune Diffusion models:

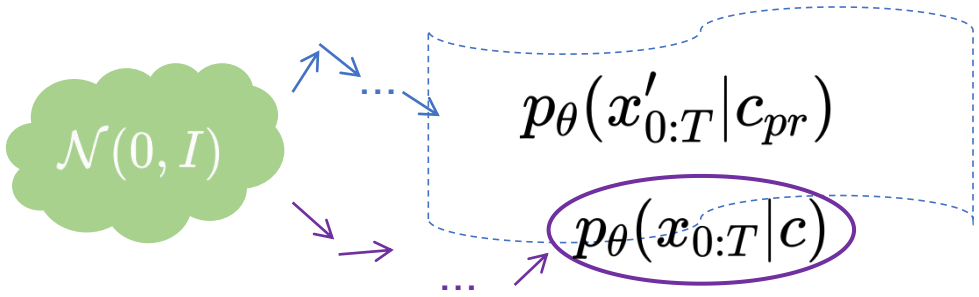
$$\mathcal{L}_{\text{cond}}(\theta, x_0) = \mathbb{E}_{x_0,t,c,\epsilon\sim\mathcal{N}(0,\mathbf{I})}||\epsilon - \epsilon_\theta(x_{t+1}, t, c)||_2^2$$

$$\mathcal{L}_{\text{ft}}(\theta, x_0^n) = \mathbb{E}_{x_0^n,t,t'}||\epsilon - \epsilon_\theta(x_{t+1}^n, t, c)||_2^2 + \lambda||\epsilon' - \epsilon_\theta(x'_{t'+1}, t', c_{pr})||_2^2$$

$$\mathcal{L}_{\text{ft'}}(\theta, x_0^{pre}) = \mathbb{E}_{x_0^{pre},t,t'}||\epsilon - \epsilon_\theta(x_{t+1}^{pre}, t, c)||_2^2 + \lambda||\epsilon' - \epsilon_\theta(x'_{t'+1}, t', c_{pr})||_2^2$$

◆ Methodology





Fine-tune Diffusion models:

$$\mathcal{L}_{\text{cond}}(\theta, x_0) = \mathbb{E}_{x_0,t,c,\epsilon\sim\mathcal{N}(0,\mathbf{I})}||\epsilon - \epsilon_\theta(x_{t+1}, t, c)||_2^2$$

$$\mathcal{L}_{\text{ft}}(\theta, x_0^n) = \mathbb{E}_{x_0^n,t,t'}||\epsilon - \epsilon_\theta(x_{t+1}^n, t, c)||_2^2 + \lambda||\epsilon' - \epsilon_\theta(x'_{t'+1}, t', c_{pr})||_2^2$$

$$\mathcal{L}_{\text{ft'}}(\theta, x_0^{pre}) = \mathbb{E}_{x_0^{pre},t,t'}||\epsilon - \epsilon_\theta(x_{t+1}^{pre}, t, c)||_2^2 + \lambda||\epsilon' - \epsilon_\theta(x'_{t'+1}, t', c_{pr})||_2^2$$

Our protecting goals:

$$f^* \in \arg\max_f \mathcal{L}_{\text{cond}}(\theta^*, p(f(\mathcal{X}^n))),$$

$$\text{s.t.} \quad \theta^* \in \arg\min_\theta \mathcal{L}_{\text{ft'}}(\theta, f(\mathcal{X}^n)),$$

$$||f(\mathcal{X}^n) - \mathcal{X}^n||_p \le \eta.$$

Ensure the stealthiness of protective noise

◆ Methodology

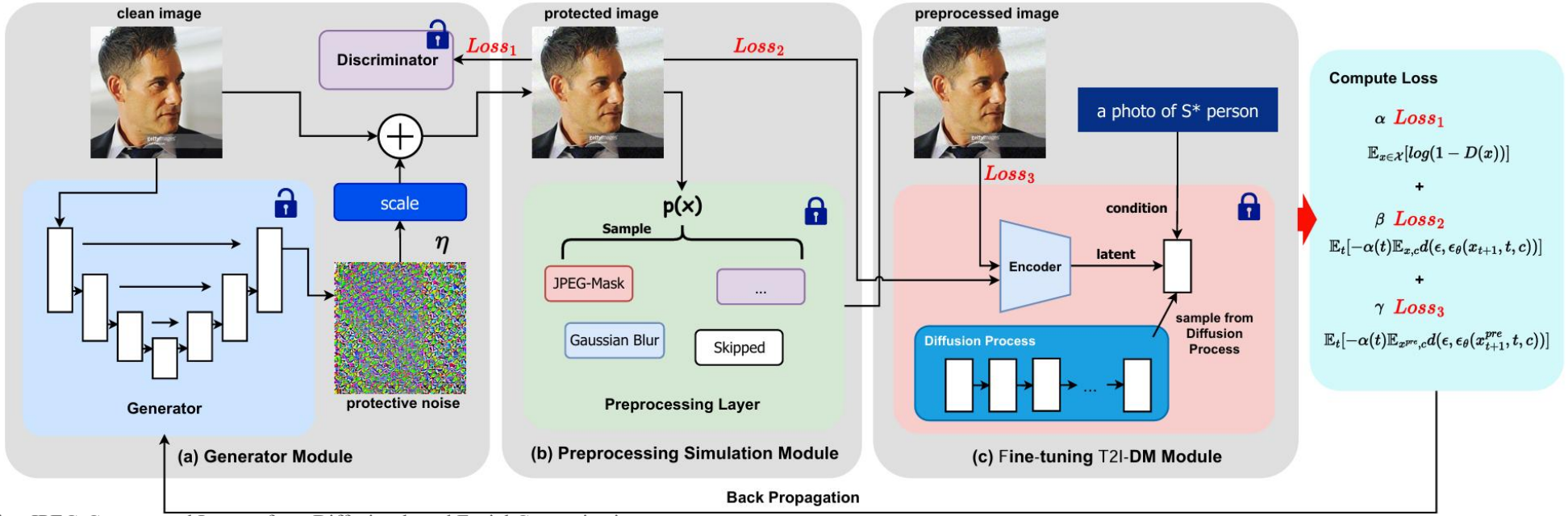
GAP-Diff Framework Overview

- Three Core Modules:

  1. Generator: Produces adversarial perturbations via U-Net.

  2. Pre-processing Simulation: Simulates JPEG compression along with other potential preprocessings during training.

  3. Fine-tuning T2I-DM: Disrupts the denoising learning process of diffusion models through adversarial loss.



◆ Methodology

$$\mathcal{L}_{\text{GAP-Diff}} = \alpha \mathcal{L}_{\text{D}}(x) + \beta \mathcal{L}_{\text{adv}}(x, c, t) + \gamma \mathcal{L}_{\text{adv}}(x^{pre}, c, t)$$

Stealthiness

- Noise Budget: x = x^n + η * δ
- GAN-based Training

$$\mathcal{L}_{\text{D}}(x) = \mathbb{E}_{x \in \mathcal{X}}[log(1 - D(x))]$$

train discriminator

$$\mathcal{L}_{\text{GAN}}(x^n, x) = \mathbb{E}_{x^n \in \mathcal{X}^n}[logD(x^n)] + \mathbb{E}_{x \in \mathcal{X}}[log(1 - D(x))]$$

◆ Methodology

$$\mathcal{L}_{\text{GAP-Diff}} = \alpha\mathcal{L}_{\text{D}}(x) + \beta\mathcal{L}_{\text{adv}}(x, c, t) + \gamma\mathcal{L}_{\text{adv}}(x^{pre}, c, t)$$

Stealthiness

- Noise Budget: x = x^n + η * δ

$$\mathcal{L}_{\text{D}}(x) = \mathbb{E}_{x\in\mathcal{X}}[log(1 - D(x))]$$

- GAN-based Training

$$\mathcal{L}_{\text{GAN}}(x^n, x) = \mathbb{E}_{x^n\in\mathcal{X}^n}[logD(x^n)] + \mathbb{E}_{x\in\mathcal{X}}[log(1 - D(x))]$$

Adversariality

Counter-direction Generation Strategy

$$\mathcal{L}_{\text{adv}}(x, c, t) = \mathbb{E}_t[-\alpha(t)\mathbb{E}_{x,c}d(\epsilon, \epsilon_\theta(x_{t+1}, t, c))]$$

◆ Methodology

$$\mathcal{L}_{\text{GAP-Diff}} = \alpha\mathcal{L}_{\text{D}}(x) + \beta\mathcal{L}_{\text{adv}}(x, c, t) + \gamma\mathcal{L}_{\text{adv}}(x^{pre}, c, t)$$

Stealthiness

- Noise Budget: x = x^n + η * δ

$$\mathcal{L}_{\text{D}}(x) = \mathbb{E}_{x\in\mathcal{X}}[log(1 - D(x))]$$

- GAN-based Training

$$\mathcal{L}_{\text{GAN}}(x^n, x) = \mathbb{E}_{x^n\in\mathcal{X}^n}[logD(x^n)] + \mathbb{E}_{x\in\mathcal{X}}[log(1 - D(x))]$$

Adversariality

Counter-direction Generation Strategy

$$\mathcal{L}_{\text{adv}}(x, c, t) = \mathbb{E}_t[-\alpha(t)\mathbb{E}_{x,c}d(\epsilon, \epsilon_\theta(x_{t+1}, t, c))]$$

Robustness

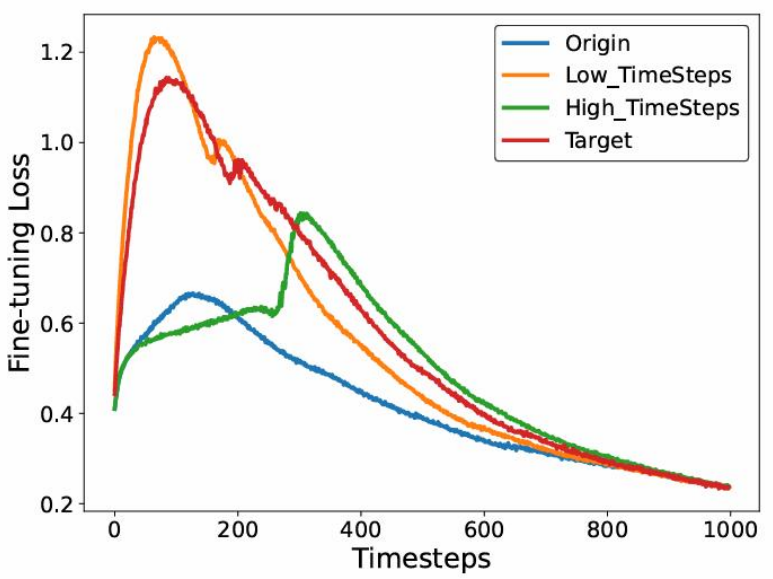
Preprocessing Simulation Layer

$$\mathbb{E}_{x^n\in\mathcal{X}^n, t\in(0,T)}\mathcal{L}_{\text{adv}}(p(x^n + \eta \times g_\psi(x^n)), c, t)$$

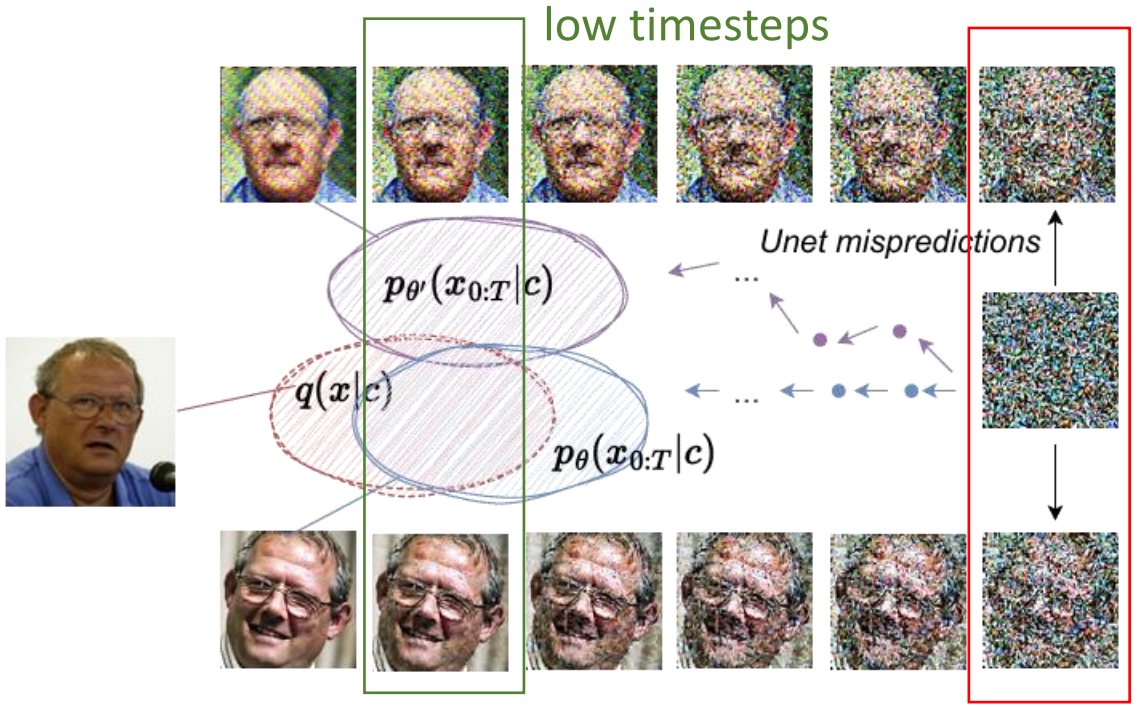
Methodology

$$\mathcal{L}_{\text{adv}}(x, c, t) = \mathbb{E}_t[-\alpha(t)\mathbb{E}_{x,c}d(\epsilon, \epsilon_\theta(x_{t+1}, t, c))]$$

Different time steps in the diffusion model have distinct characteristics



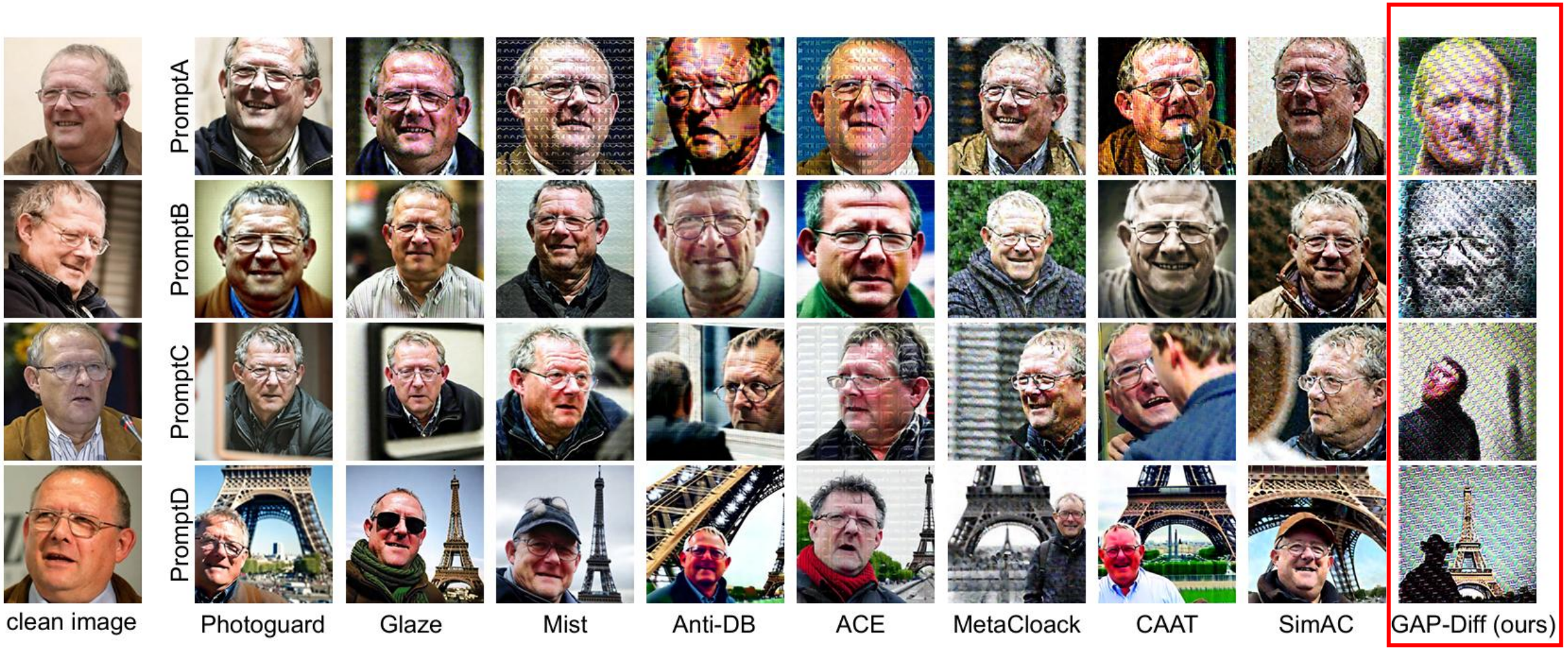
$$y = \begin{cases} 1 & \text{if } x \in (0, 800) \\ 0 & \text{otherwise.} \end{cases}$$



◆ Experiment— Quantitative Results

Methods	"a photo of sks person"				"a dslr portrait of sks person"			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
No Defense	5.33	0.61	26.27	0.69	21.57	0.48	9.64	0.71
Photoguard [36]	6.22	0.55	29.38	0.71	19.44	0.46	13.74	0.71
Glaze [38]	6.57	0.53	30.42	0.69	18.78	0.45	11.04	0.69
Mist [24]	14.89	0.46	35.68	0.60	19.56	0.38	20.43	0.63
Anti-DB [45]	22.89	0.41	40.19	0.40	32.67	0.34	32.72	0.44
ACE [53]	8.44	0.47	37.22	0.61	15.22	0.38	27.80	0.64
MetaCloak [26]	31.69	0.44	38.82	0.51	35.28	0.36	27.31	0.56
CAAT [49]	25.44	0.43	42.01	0.45	21.67	0.38	25.07	0.57
SimAC [47]	19.11	0.49	39.43	0.52	23.56	0.41	24.15	0.62
GAP-Diff (ours)	**77.56**	**0.25**	**42.04**	**0.23**	**76.33**	**0.19**	**48.97**	**0.20**
Methods	**"a photo of sks person looking at the mirror"**				**"a photo of sks person in front of eiffel tower"**			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
No Defense	8.67	0.44	19.61	0.56	20.67	0.22	20.11	0.44
Photoguard [36]	9.67	0.40	23.43	0.56	19.56	0.21	17.91	0.45
Glaze [38]	9.33	0.41	19.69	0.55	17.44	0.20	19.78	0.43
Mist [24]	12.33	0.35	22.17	0.50	27.33	0.18	21.05	0.36
Anti-DB [45]	21.67	0.30	24.77	0.37	34.88	0.14	31.21	0.26
ACE [53]	12.44	0.30	31.90	0.42	36.11	0.15	25.25	0.26
MetaCloak [26]	32.76	0.32	34.14	0.36	30.57	0.15	31.22	0.25
CAAT [49]	16.33	0.32	23.82	0.37	34.22	0.14	31.82	0.25
SimAC [47]	14.89	0.33	31.09	0.42	28.56	0.14	32.98	0.25
GAP-Diff (ours)	**84.56**	**0.14**	**47.30**	**0.13**	**72.78**	**0.08**	**41.69**	**0.08**

◆ Experiment— Qualitative Results



PromptA “a photo of sks person”, PromptB “a dslr portrait of sks person”,

PromptC “a photo of sks person looking at the mirror”, and PromptD “a photo of sks person in front of eiffel tower”.

◆ Experiment— Quantitative Results

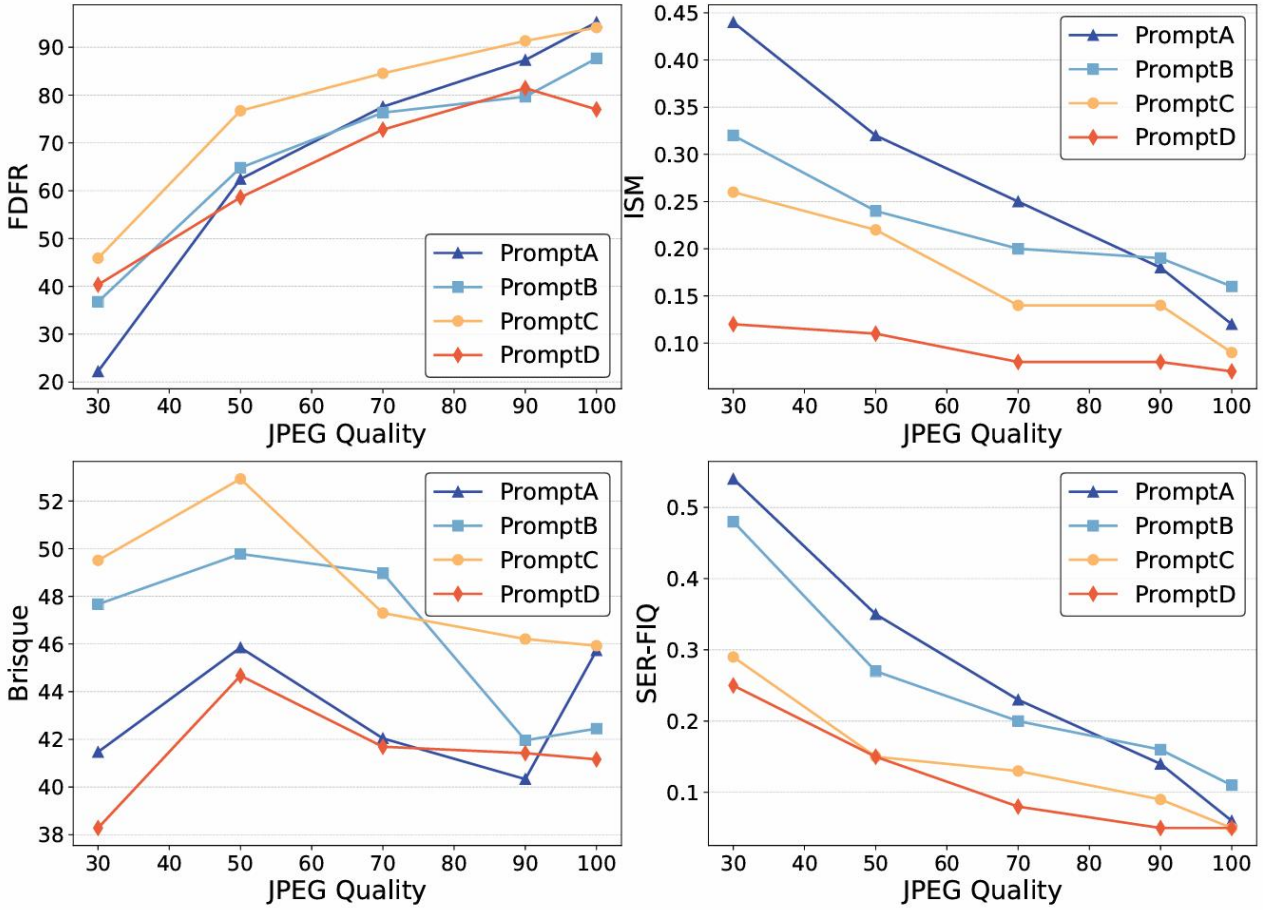
Methods	"a photo of sks person"				"a dslr portrait of sks person"			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
Photoguard [36]	4.78	0.55	32.72	0.70	22.89	0.44	9.52	0.68
Glaze [38]	4.89	0.56	30.81	0.71	23.11	0.43	8.76	0.68
Mist [24]	10.89	0.52	34.92	0.69	21.22	0.42	12.02	0.68
Anti-DB [45]	10.44	0.50	36.45	0.57	23.00	0.39	19.33	0.62
ACE [53]	7.55	0.51	37.94	0.68	17.67	0.38	21.24	0.67
MetaCloak [26]	32.16	0.46	40.05	0.54	38.25	0.41	27.76	0.60
CAAT [49]	10.44	0.52	37.50	0.61	17.44	0.42	13.90	0.65
SimAC [47]	10.22	0.51	37.19	0.65	18.89	0.43	14.60	0.67
Ours	**62.44**	**0.32**	**45.85**	**0.35**	**64.80**	**0.24**	**49.78**	**0.27**
Methods	**"a photo of sks person looking at the mirror"**				**"a photo of sks person in front of eiffel tower"**			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
Photoguard [36]	4.33	0.39	20.76	0.55	19.56	0.21	17.91	0.45
Glaze [38]	6.89	0.40	20.71	0.56	20.56	0.20	20.34	0.44
Mist [24]	7.78	0.37	22.69	0.51	21.56	0.18	20.91	0.38
Anti-DB [45]	17.78	0.32	14.82	0.42	26.89	0.17	28.11	0.28
ACE [53]	7.89	0.32	28.09	0.48	32.56	0.16	22.97	0.31
MetaCloak [26]	28.40	0.33	33.87	0.38	24.73	0.17	30.89	0.27
CAAT [49]	11.11	0.36	13.17	0.46	28.33	0.18	27.84	0.31
SimAC [47]	7.22	0.36	26.90	0.49	20.67	0.18	27.57	0.35
Ours	**76.73**	**0.22**	**52.93**	**0.15**	**58.61**	**0.11**	**44.67**	**0.15**

◆ Experiment— Qualitative Results

PromptA “a photo of sks person”, PromptB “a dslr portrait of sks person”,

PromptC “a photo of sks person looking at the mirror”, and PromptD “a photo of sks person in front of eiffel tower”.

◆ Experiment— Ablation Studies



Resistance to JPEG compression at different intensities

Protection effectiveness under different noise budgets

◆ Experiment— Ablation Studies

Method	"a photo of sks person"			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
w/o preprocess	95.22	0.12	45.73	0.06
Random noise	26.44	0.41	29.41	0.49
Gaussian blur	83.78	0.24	42.25	0.18
Quantization	93.89	0.11	43.44	0.06
Resize	90.56	0.17	42.50	0.12
Super resolution	93.67	0.13	42.82	0.10
Method	**"a dslr portrait of sks person"**			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
w/o preprocess	87.67	0.16	42.44	0.11
Random noise	35.33	0.31	36.30	0.45
Gaussian blur	82.44	0.19	46.67	0.12
Quantization	85.44	0.15	43.05	0.12
Resize	82.67	0.17	41.89	0.15
Super resolution	85.22	0.16	42.75	0.15

Method	"a photo of sks person looking at the mirror"			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
w/o preprocess	94.11	0.09	45.93	0.05
Random noise	30.46	0.28	34.73	0.32
Gaussian blur	86.11	0.14	44.71	0.09
Quantization	94.00	0.11	46.63	0.05
Resize	87.11	0.18	46.16	0.09
Super resolution	90.11	0.15	49.24	0.08
Method	**"a photo of sks person in front of eiffel tower"**			
	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓
w/o preprocess	77.00	0.07	41.15	0.06
Random noise	24.89	0.12	26.89	0.32
Gaussian blur	73.11	0.08	43.90	0.07
Quantization	77.33	0.08	41.06	0.06
Resize	78.33	0.08	41.34	0.06
Super resolution	78.44	0.09	40.43	0.07

Efficiency

Method	Runtime (second)	Memory Usage (GB)
Photoguard [36]	194	16
Glaze [38]	118	4
Mist [24]	294	5
Anti-DB [45]	308	18
ACE [53]	175	6
MetaCloak [26]	1.1×10^4	504
CAAT [49]	98	18
SimAC [47]	1.2×10^3	33
GAP-Diff (ours)	**0.04**	**2**

◆ Conclusion

Major Contributions

- Novel generative framework with JPEG compression resistance
- Efficient one-step noise generation compared to iterative methods
- Robust performance across various scenarios and settings

Future Directions

- Server API deployment for real-time protection
- Optimize training strategy for better balance between protection and visual quality
- Extend to other preprocessing techniques and customization methods

Thank You!

Please email to haotian.zhu@njust.edu.cn for any question

Code: https://github.com/AIASLab/GAP-Diff