ICML
International Conference
On Machine Learning

中山大學
SUN YAT-SEN UNIVERSITY

Variance as a Catalyst: Efficient and Transferable Semantic Erasure Adversarial Attack for Customized Diffusion Models

Jiachen Yang[1] Yusong Wang[1] Yanmei Fang[1 2] Yunshu Dai[1] Fangjun Huang[1 2]

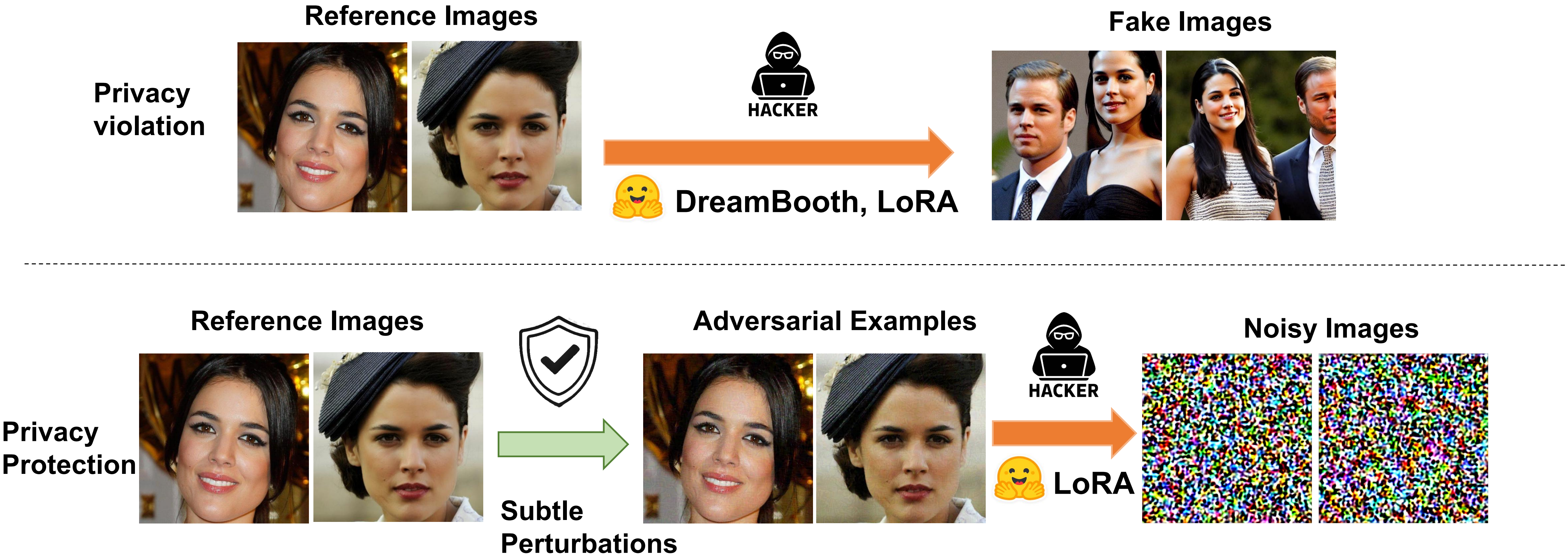
[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[2]Guangdong Provincial Key Laboratory of Information Security Technology

https://github.com/youyuanyi/variance-as-Catalyst

中山大學
SUN YAT-SEN UNIVERSITY

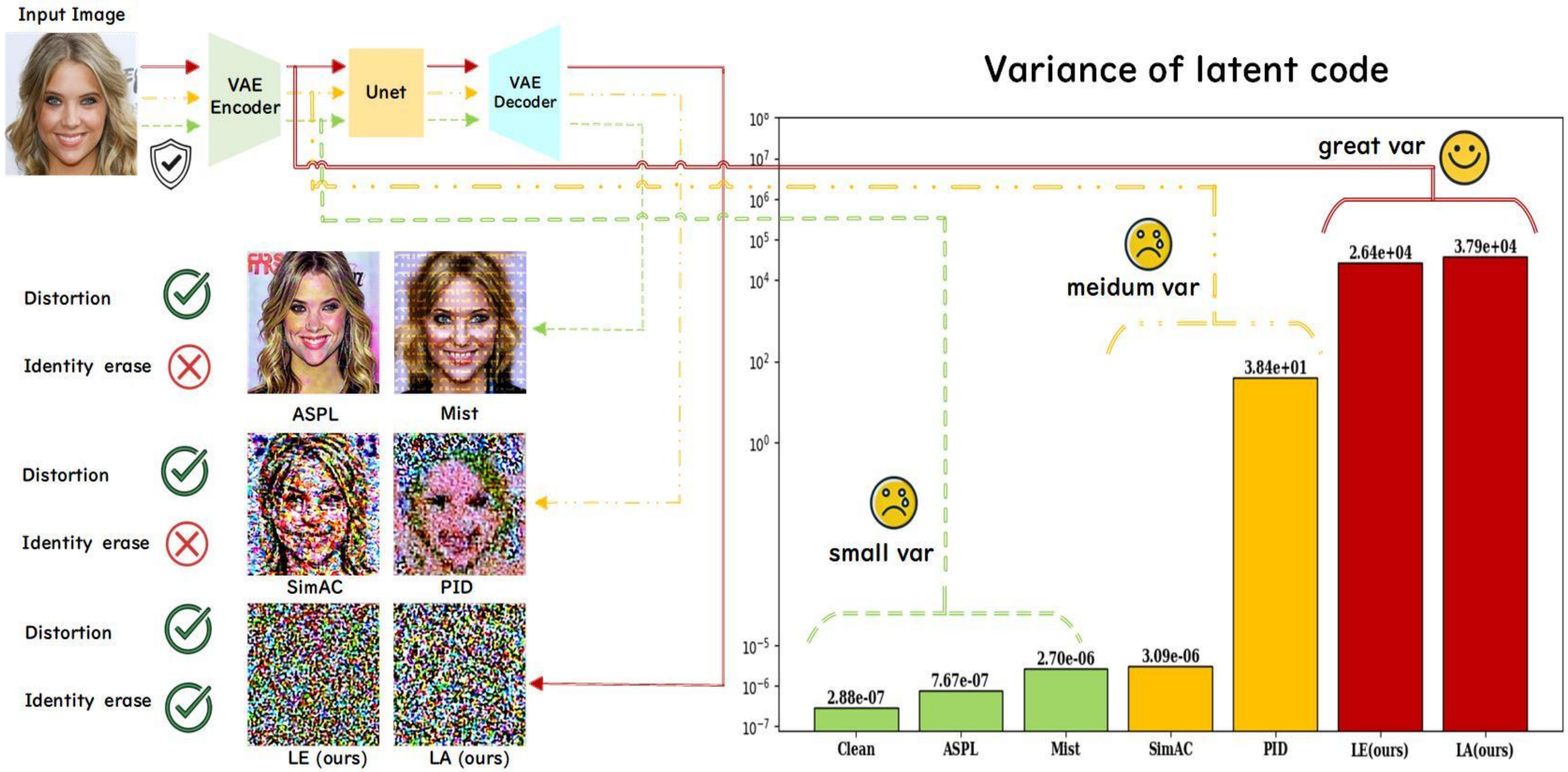
Motivation

- **Security risks of AIGC**



Innovation

1. **Existing attack methods only cause image distortion and fail to achieve identity erasure.**
2. **We identify that larger VAE variance enables stronger semantic erasure.**



Method

- **Gradient Consistency Theory: A framework of alignment between perturbation and variance growth.**

1. Laplace Loss (LA)

$$\mathcal{L}_{Laplace} = \frac{|\sigma^2 - \mu|}{b},$$

- Locally optimal updates
- Gradient-aligned

2. Lagrange Entropy Loss (LE)

$$\mathcal{L}_{LE} = -\sum_j \sigma_j^2 \log(\sigma_j^2) + \lambda \left(\sum_j \sigma_j^2 - c \right)^2$$

- Ample optimization space
- Stable optimization

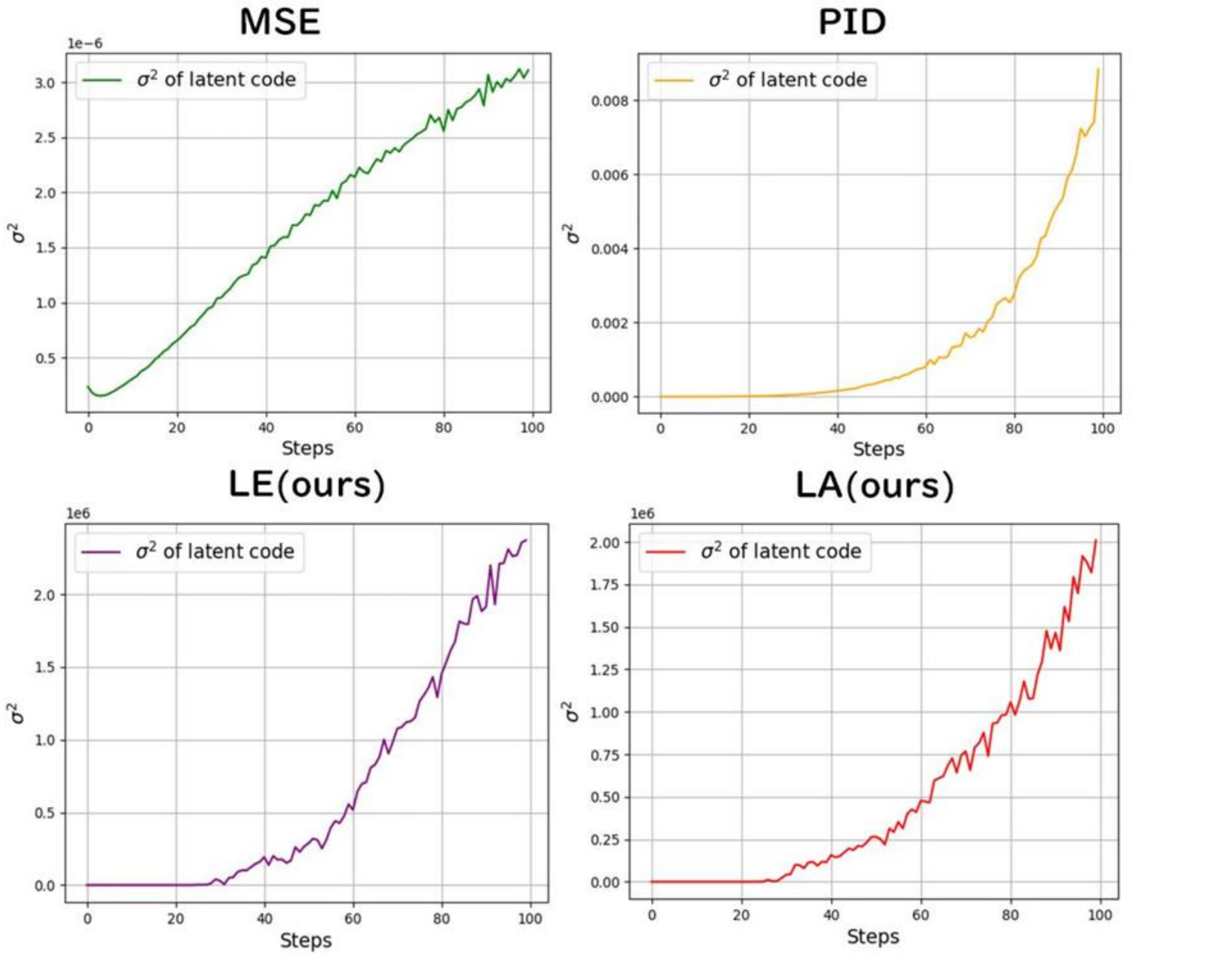


Figure1. Variance Growth of Latent Codes Within 100 Optimization Steps

Experiments

Figure2. Comparison of various adversarial attacks in disrupting personalized image generation with DreamBooth

Experiments

- ISM： Identity Similarity Between Reference and Generated Images
- FDFR： Face Detection Failure Rate
- Brisque： Image Natural Quality
- LPIPS： Image Perceptual Quality

Table 1: Comparing the performance of our method with baselines against DreamBooth (Ruiz et al., 2023) on CelebA-HQ and VGGFace2. The best result under each metric is marked with **bold**. The prompt used is *"a photo of a sks person."*

Method	CelebA-HQ				VGGFace2			
	ISM ↓	FDFR ↑	Brisque ↑	LPIPS ↑	ISM ↓	FDFR ↑	Brisque ↑	LPIPS ↑
No Defense	0.608	0.041	17.896	0.662	0.638	0.025	18.193	0.724
AdvDM (Liang et al., 2023)	0.424	0.307	24.215	0.798	0.142	0.944	47.862	0.868
ASPL (Van Le et al., 2023)	0.406	0.287	24.419	0.805	0.158	0.906	46.142	0.865
Mist (Liang & Wu, 2023)	0.249	0.169	13.981	0.707	0.246	0.257	18.324	0.756
MetaCloak (Liu et al., 2024b)	0.593	0.051	36.325	0.712	0.525	0.059	36.771	0.747
SimAC (Wang et al., 2024)	0.253	0.865	51.059	0.823	0.196	0.981	51.874	0.836
DisDiff (Liu et al., 2024a)	0.605	0.116	29.361	0.695	0.263	0.902	43.623	0.758
SDS- (Xue et al., 2023)	0.655	0.005	38.519	0.743	0.591	0.002	37.325	0.781
PID (Li et al., 2024)	0.069	0.938	85.533	0.899	0.046	0.968	86.946	0.945
LE(ours)	**0**	**1**	**155.804**	**1.021**	**0**	**1**	**154.494**	**1.028**
LA(ours)	**0**	**1**	**155.845**	**0.959**	**0**	**1**	**155.845**	**1.031**

Experiments

Transferability

LoRA

SD1.5 ⟶ SD 2.1

SD1.5 ⟶ SDXL

Method	SD2.1				SDXL			
	ISM ↓	FDFR ↑	Brisque ↑	LPIPS ↑	ISM ↓	FDFR ↑	Brisque ↑	LPIPS ↑
No Defense	0.729	0.073	16.409	0.669	0.791	0.001	13.483	0.515
AdvDM (Liang et al., 2023)	0.532	0.313	39.369	0.704	0.765	0.019	18.419	0.539
ASPL (Van Le et al., 2023)	0.519	0.331	39.226	0.709	0.766	0.002	20.589	0.524
Mist (Liang & Wu, 2023)	0.179	0.231	18.097	0.677	0.583	0.126	24.143	0.622
MetaCloak (Liu et al., 2024b)	0.635	0.087	41.381	0.699	0.738	0.002	22.766	0.517
SimAC (Wang et al., 2024)	0.401	0.642	41.409	0.733	0.746	0.018	10.459	0.546
DisDiff (Liu et al., 2024a)	0.627	0.166	40.127	0.709	0.782	0	17.366	0.519
SDS- (Xue et al., 2023)	0.673	0.016	52.108	0.711	0.732	0	8.103	0.569
PID (Li et al., 2024)	0.089	0.887	91.461	0.949	0.602	0	17.848	0.545
LE(ours)	**0**	**1**	**151.089**	**0.947**	**0.171**	**0.649**	**126.369**	**0.893**
LA(ours)	**0**	**1**	**154.724**	**0.955**	**0.178**	**0.401**	**102.815**	**0.822**

LoRA

SD1.5 ⟶ SD 3.5

SD1.5 ⟶ FLUX.1 Dev

Method	SD3.5				FLUX.1-dev			
	ISM ↓	FDFR ↑	Brisque ↑	LPIPS ↑	ISM ↓	FDFR ↑	Brisque ↑	LPIPS ↑
No Defense	0.587	0.001	0.699	0.589	0.705	0.001	7.722	0.598
AdvDM (Liang et al., 2023)	0.532	0.002	11.755	0.612	0.739	0.009	13.268	0.651
ASPL (Van Le et al., 2023)	0.543	0.001	11.541	0.621	0.743	0.004	12.991	0.641
Mist (Liang & Wu, 2023)	0.456	0.003	18.712	0.629	0.736	0.005	25.182	0.587
MetaCloak (Liu et al., 2024b)	0.469	0	21.861	0.599	0.727	0.007	23.027	0.618
SimAC (Wang et al., 2024)	0.495	0	9.251	0.606	0.735	0.004	3.601	0.645
DisDiff (Liu et al., 2024a)	0.535	0.003	10.252	0.601	0.708	0.021	13.488	0.639
SDS- (Xue et al., 2023)	0.499	0	46.701	0.617	0.741	0.003	24.258	0.609
PID (Li et al., 2024)	0.484	0.013	15.434	0.605	0.729	0.001	15.326	0.596
LE(ours)	**0.217**	**0.241**	**65.903**	**0.803**	**0.257**	**0.204**	**58.423**	**0.828**
LA(ours)	**0.235**	**0.214**	**71.873**	**0.793**	**0.282**	**0.269**	**52.554**	**0.845**

Experiments

Attack Efficiency

Method	Time/s ↓	GPU/MB ↓
AdvDM (Liang et al., 2023)	18.63	8278.63
ASPL (Van Le et al., 2023)	189.95	34366.92
Mist (Liang & Wu, 2023)	18.81	8278.63
MetaCloak (Liu et al., 2024b)	1843.47	16955.00
SimAC (Wang et al., 2024)	124.57	38640.00
DisDiff (Liu et al., 2024a)	65.54	25960.50
SDS- (Xue et al., 2023)	18.61	8278.63
PID (Li et al., 2024)	241.31	4581.93
LE(ours)	**7.34**	**4469.80**
LA(ours)	**8.47**	**4469.80**

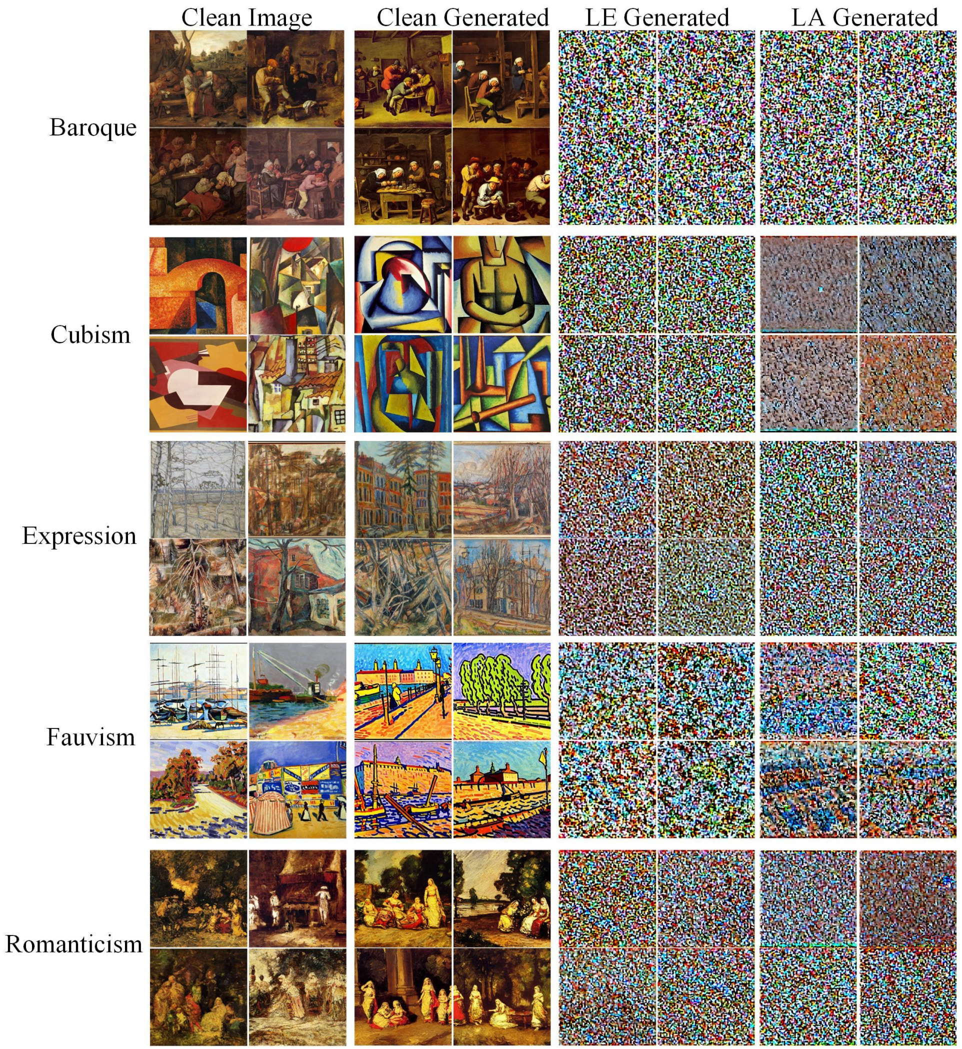
Defense Qualiy & Visual Quality

η	Defense Quality		Visual Quality	
	IMS ↓	Brisque ↑	PSNR ↑	SSIM ↑
4/255	0.631	22.385	14.244	0.414
8/255	0	122.266	13.771	0.361
0.05*	0	155.804	13.664	0.313
16/255	0	155.845	12.301	0.271

- Speed: Over 30× faster than existing SOTA methods
- Memory Usage: Requires less than 4.5 GB of GPU memory
- Our method still surpasses SOTA methods even when the perturbation budget is tightened to 8/255.

Experiments

WikiArt



ControlNet-based Image Editing