NeurIPS 2026

DSSA

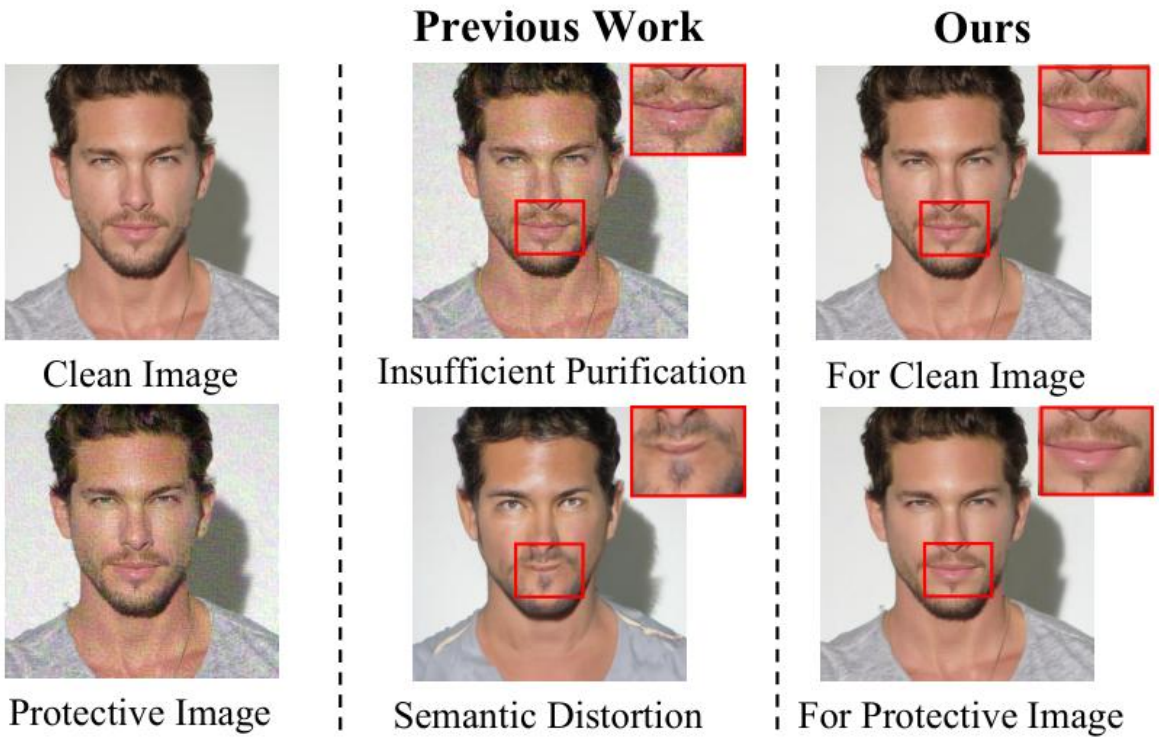
Dynamic Sparse Semantic Anchoring for Purifying Protective Perturbations against Diffusion Models

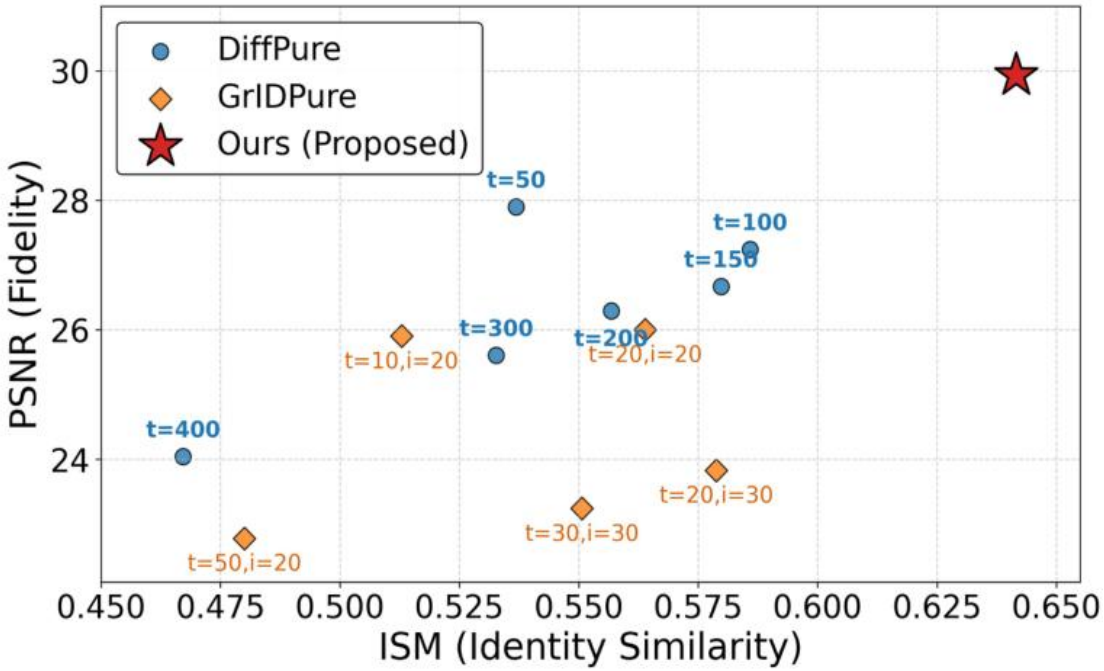
Weiwei Tan[1,2], Rui Wang[1,2*], Lihua Jing[1,2], Yanjun Zhang[3],
Runbo Li[1,2], Zhishen Wang[1,2], Leo Yu Zhang[3]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences [3] Griffith University

The purification trade-off

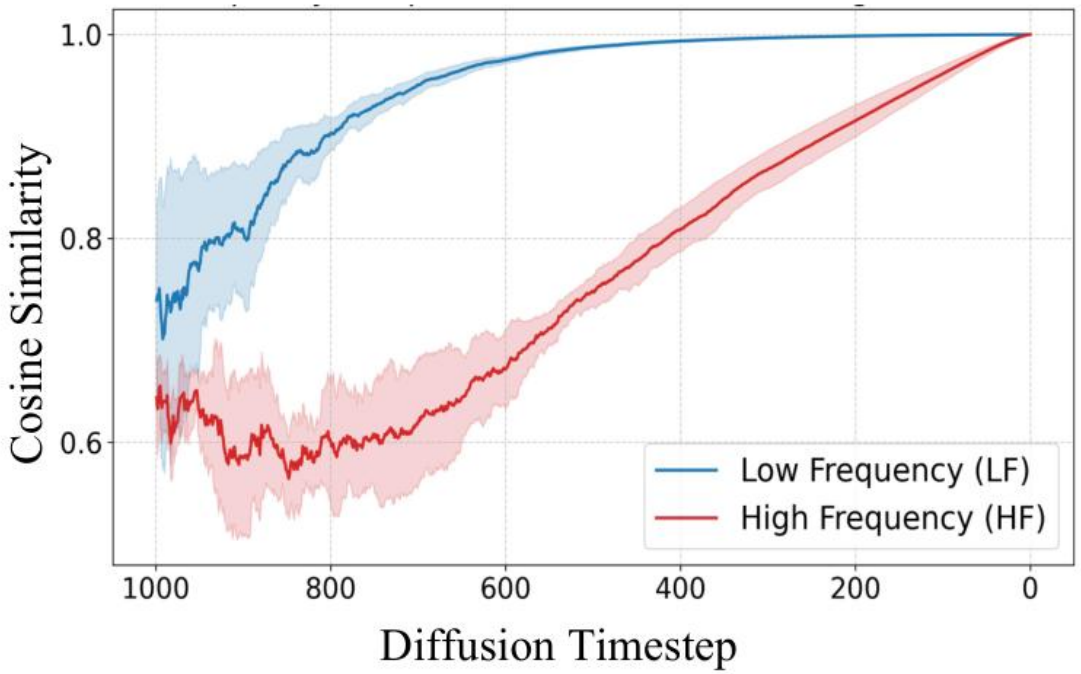
Protective perturbations aim to prevent unauthorized personalization.



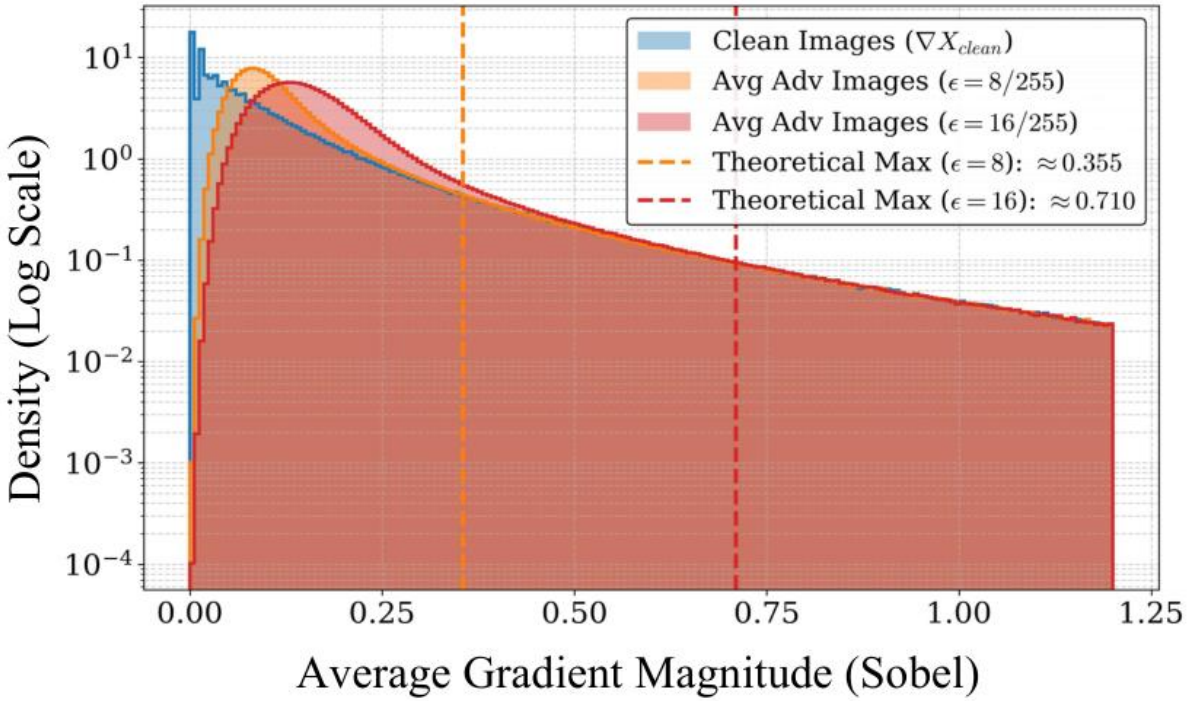


Too little noise leaves protection. Too much noise damages content.

Two observations behind DSSA



(a) Frequency-Wise Semantic Recovery

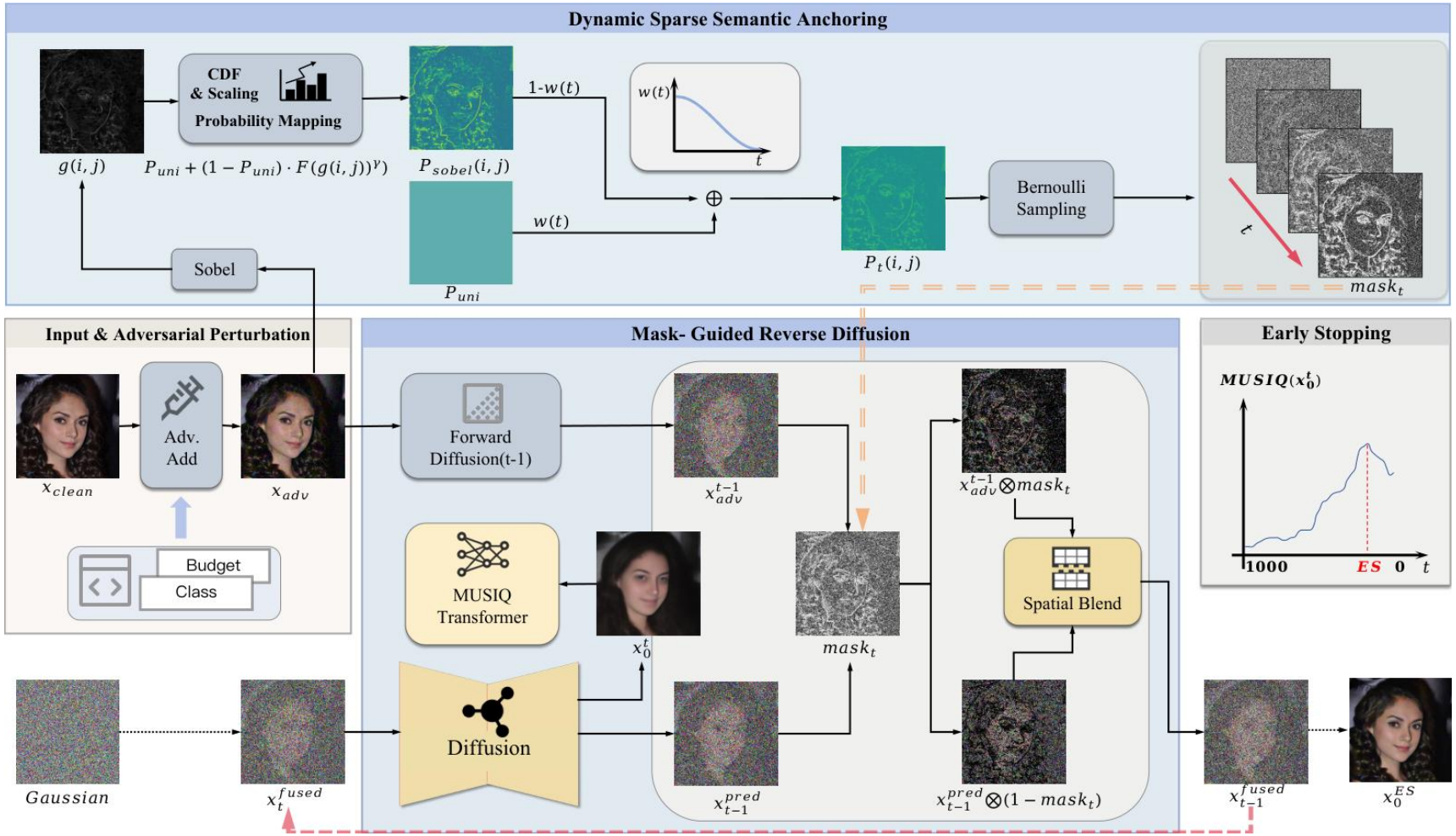


(b) Gradient Magnitude Distributions

Global structure recovers before fine details.

Strong edges tend to occupy higher gradient quantiles.

Dynamic sparse semantic anchoring

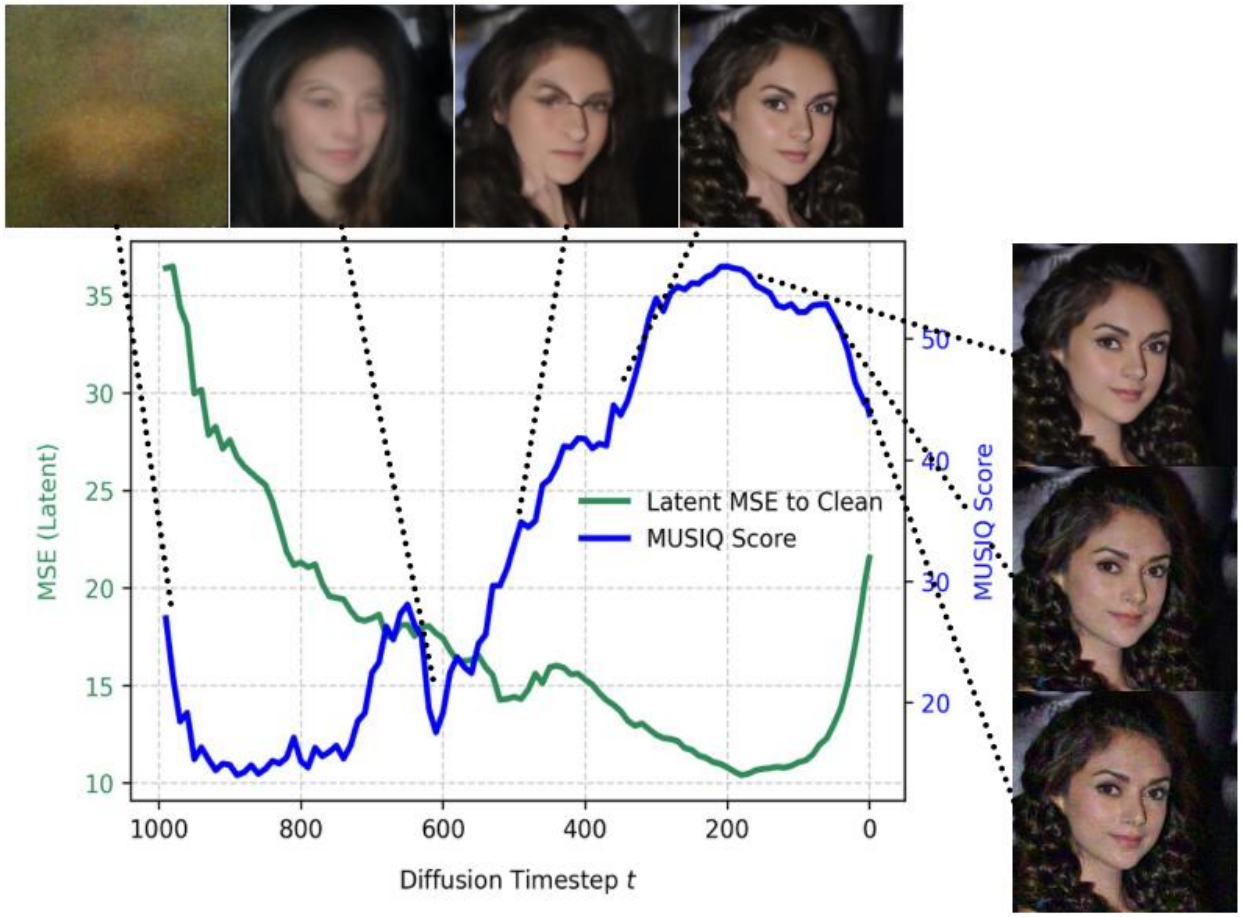


1. Start from pure noise

2. Blend sparse input pixels with the model prediction

3. Shift from uniform sampling to edge guidance

Early stopping prevents perturbation return



Track MUSIQ quality without a clean reference.

Stop after a sustained drop in quality.

Return the best estimate seen so far.

Image fidelity and personalization

Image fidelity at ε = 16/255

Method	PSNR ↑	SSIM ↑	LPIPS ↓
IMPRESS	23.61	0.5254	0.6229
DiffPure	27.09	0.6733	0.3707
GrIDPure	25.49	0.8079	0.2759
Noisy Upscaling	26.87	0.7396	0.2301
DSSA (Ours)	**29.05**	**0.8264**	**0.2117**

DreamBooth on VGGFace2, AdvDM, ε = 16/255

DSSA: identity similarity 0.6299 FID 135.60

Recovered faces and artistic styles



Pure-noise start

Sparse structural guidance

Stop before resurgence

Thank you

tanweiwei@iie.ac.cn