ICCV HONOLULU HAWAII
OCT 19-23, 2025

SUNG KYUN KWAN UNIVERSITY

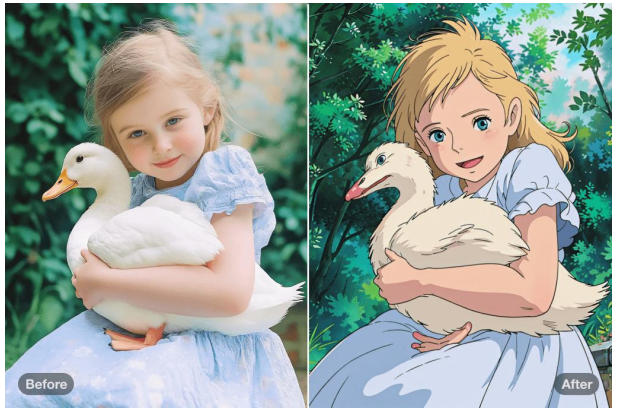
# DIA: The Adversarial Exposure of Deterministic Inversion in Diffusion Models

Seunghoo Hong[1,†] Geonho Son[2,†] Juhun Lee[1] Simon S. Woo[1,2,⋆]
[1]Dept. of Artificial Intelligence, [2]Dept. of Computer Science & Engineering
Sungkyunkwan University, South Korea
{hoo0681, sohn1029, josejhlee, swoo}@g.skku.edu

# Background

## Adversarial Attack on Image Editing

- The deepfakes generated through **Text-to-Image (T2I) generative models** are causing severe social problems




**<Style Mimicry>**

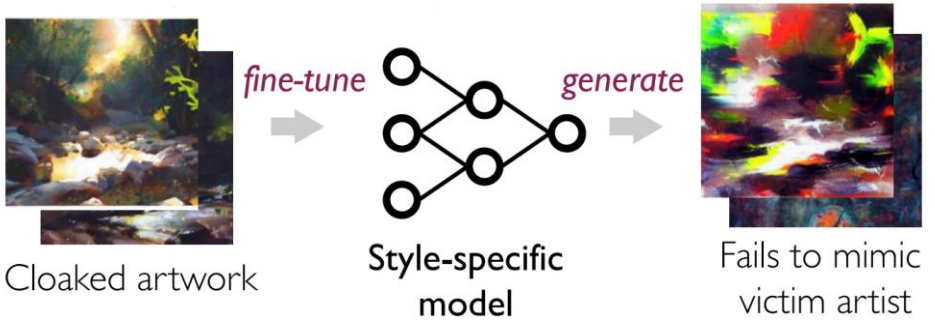
- In response, the technique of utilizing **Adversarial Noise**, known to disrupt model decisions, for **image immunization** is being re-examined and is suppressing the creation of deepfakes

**<Personalization>**

# Background

## Adversarial Attack on Image Editing

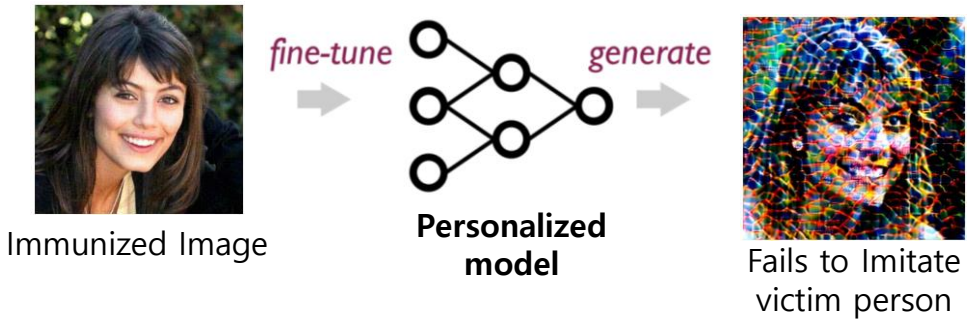
- **AdvDM** and **Photoguard** became representative methods for disrupting generation in **Diffusion Models**




**<Glaze>**

- **Previous works**
  - Suppress style mimicry : **Glaze**
  - Disrupt personalization model : **Anti-Dreambooth**

- However, image immunization to inhibit **image editing** is still non-existent




**<Anti Dreambooth>**

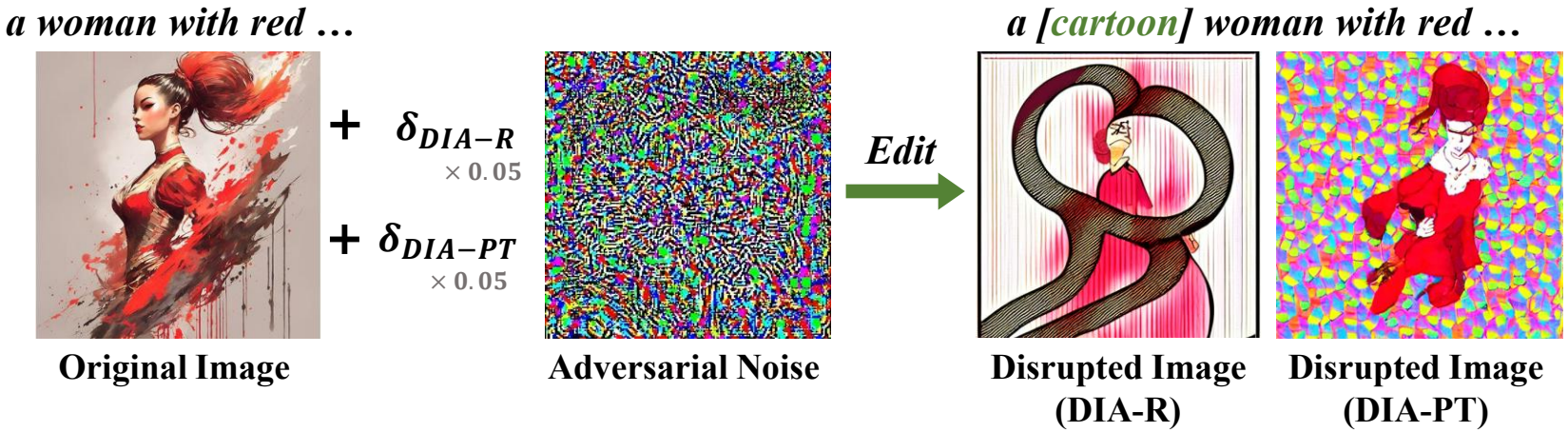
Liang, Chumeng, et al. "Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples." *arXiv preprint arXiv:2302.04578* (2023).
Salman, Hadi, et al. "Raising the cost of malicious ai-powered image editing." *arXiv preprint arXiv:2302.06588* (2023).
Shan, Shawn, et al. "Glaze: Protecting artists from style mimicry by {Text-to-Image} models." *32nd USENIX Security Symposium (USENIX Security 23)*. 2023.
Van Le, Thanh, et al. "Anti-dreambooth: Protecting users from personalized text-to-image synthesis." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

# Background

## Contributions

- DIA (DDIM Inversion Attack) achieves effective immunization by **directly bypassing the Deterministic DDIM Inversion trajectory** and hindering latent code acquisition or reconstruction.

- Resolve the out-of-memory issue by **decomposing backpropagation on a per-timestep basis** and using a method to compute the Jacobian product.



# DIA: DDIM Inversion Attack

## DDIM Sampling

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t) + \sigma_t z,$$

$$z \sim \mathcal{N}(0,I).$$

predicted $x_0$ | direction pointing to $x_t$ | random noise

$\sigma_t = 0$

$$x_t = \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_{t-1}}}x_{t-1} + \sqrt{\bar{\alpha}_t}(\lambda(t-1))\epsilon(x_t,t)$$

where $\lambda(t) := \sqrt{\frac{1}{\bar{\alpha}_{t+1}}-1} - \sqrt{\frac{1}{\bar{\alpha}_t}-1}$

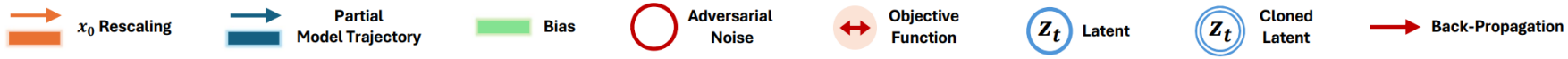
## DDIM Inversion

With the assumption of linearization,

$$\epsilon(x_t,t) \approx \epsilon(x_{t-1},t)$$

$$x_{t+1} = \sqrt{\alpha_{t+1}}x_t + \underbrace{\sqrt{\bar{\alpha}_{t+1}}(\lambda(t))\epsilon_\theta(x_t,t+1)}_{\text{noising part } \Delta_t}$$

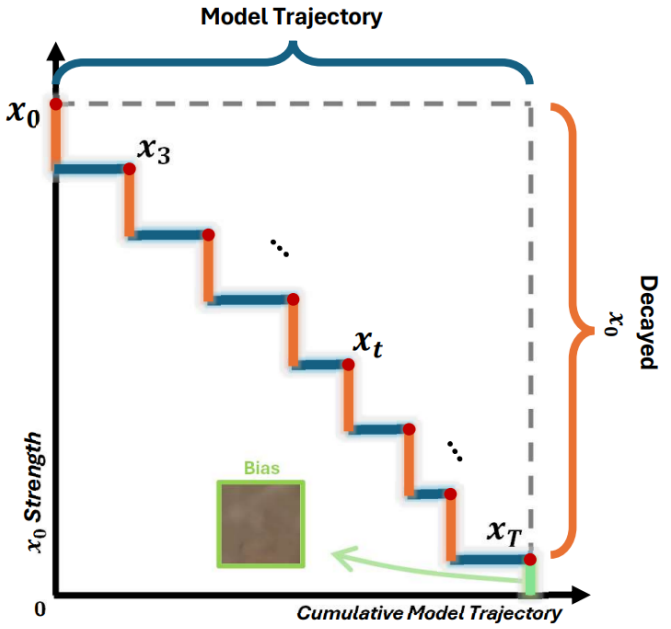
# DIA: DDIM Inversion Attack




## DIA-PT: Disrupting Process Trajectory

$$x_T = \underbrace{\sqrt{\bar{\alpha}_T}x_0}_{\text{bias}} + \underbrace{\sum_{i=0}^{T} \frac{\sqrt{\bar{\alpha}_T}}{\sqrt{\bar{\alpha}_{i+1}}}\Delta_i}_{\text{MT}}$$

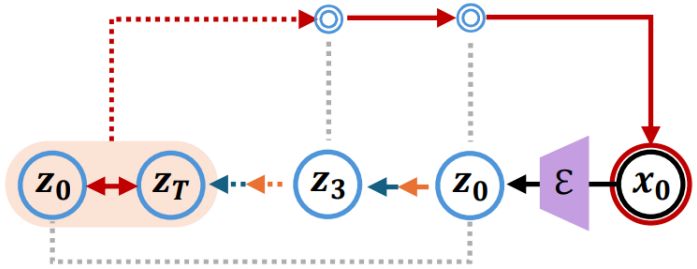
$$x_T = x_0 + \underbrace{(\sqrt{\bar{\alpha}_T} - 1)x_0 + \sum_{i=0}^{t} \frac{\sqrt{\bar{\alpha}_T}}{\sqrt{\bar{\alpha}_{i+1}}}\Delta_i}_{\text{PT}}$$




a) Components of DDIM Process

$$\delta_{\text{DIA-PT}} = \arg\max_{||\delta|| \leq \epsilon} \|\hat{x}_{0:T}(x_0 + \delta) - \mathcal{E}(x_0 + \delta)\|_2^2 ,$$

inversion




b) DIA-PT

# DIA: DDIM Inversion Attack

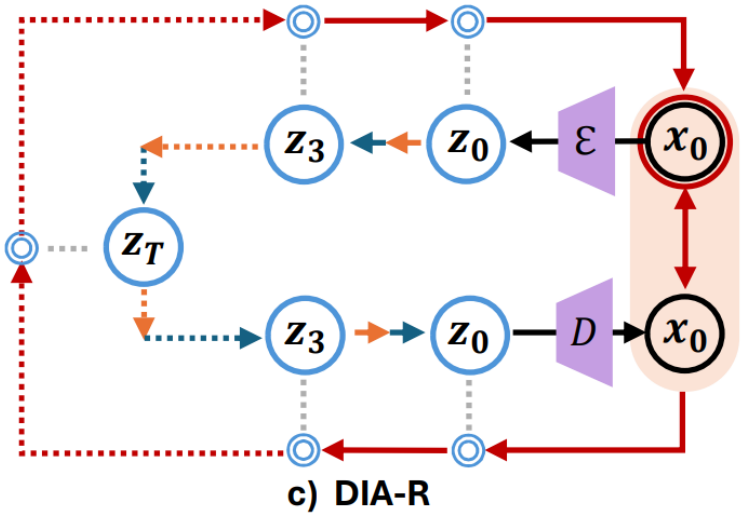
$x_0$ Rescaling | Partial Model Trajectory | Bias | Adversarial Noise | Objective Function | $z_t$ Latent | $z_t$ Cloned Latent | Back-Propagation

## DIA-R: Disrupting Reconstruction

$$\delta_{\text{DIA-R}} = \arg\max_{||\delta|| \leq \epsilon} \|\tilde{x}_{T:0}(\hat{x}_{0:T}(x_0 + \delta)) - (x_0 + \delta)\|_2^2$$

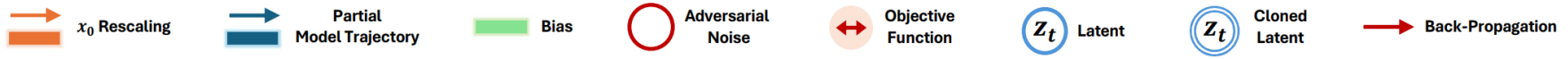
inversion

reconstruction




c) DIA-R

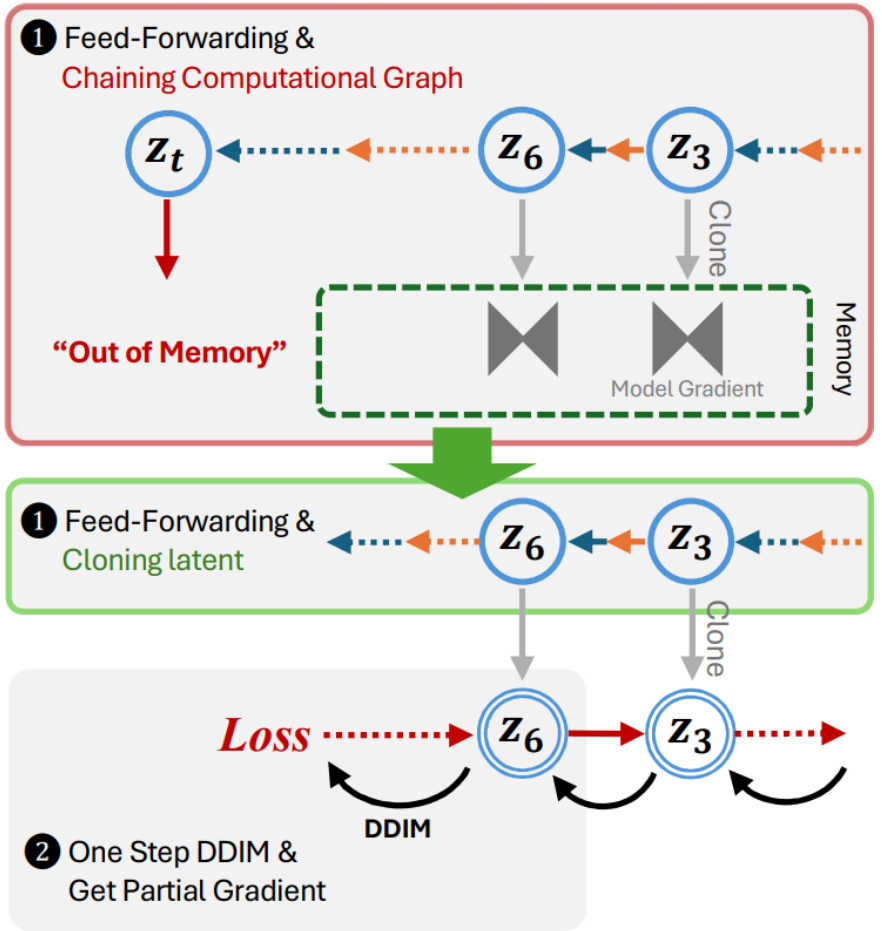
# DIA: DDIM Inversion Attack




## Differentiable Diffusion Trajectory

- The reverse diffusion path (DDIM Trajectory) utilized in DIA requires timestep-wise inference.
- Accumulation of parameter gradients per timestep causes severe memory consumption.
- Calculating the timestep-wise Vector-Jacobian by decomposing backpropagation uses a fixed amount of memory regardless of the trajectory length.

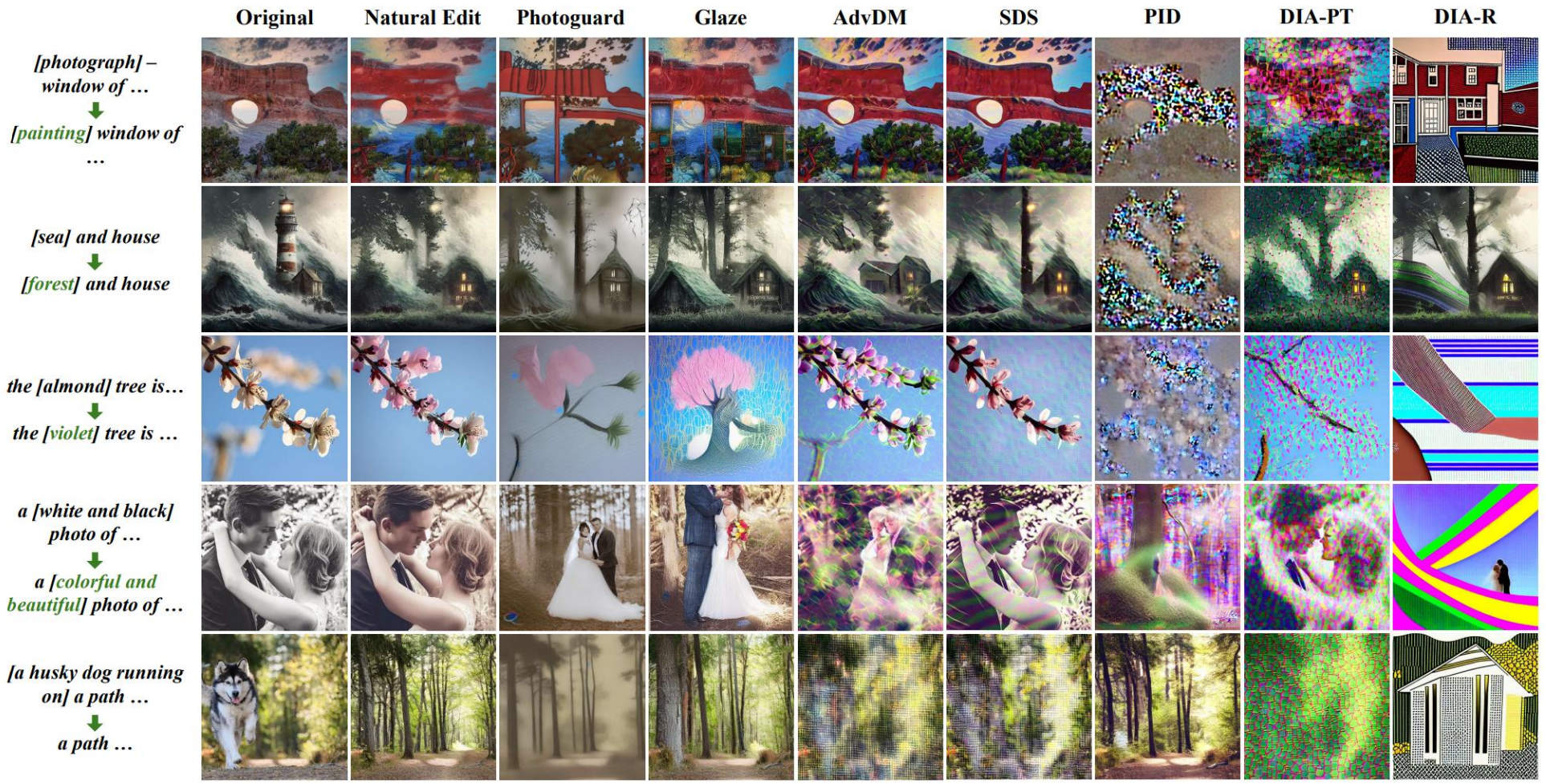
$$\nabla_{h_t}\mathcal{J} = \begin{cases} \frac{\partial \mathcal{L}}{\partial h_t}, & t = T \\ \nabla_{h_{t+1}}\mathcal{J} \cdot J_{\text{VAE}}(h_t), & t = 0 \\ \nabla_{h_{t+1}}\mathcal{J} \cdot J_{\text{DDIM}}(h_t), & \text{otherwise} \end{cases}$$




d) Differential Trajectory

# Experiments

## Qualitative Result



# Experiments

## Quantitative Results

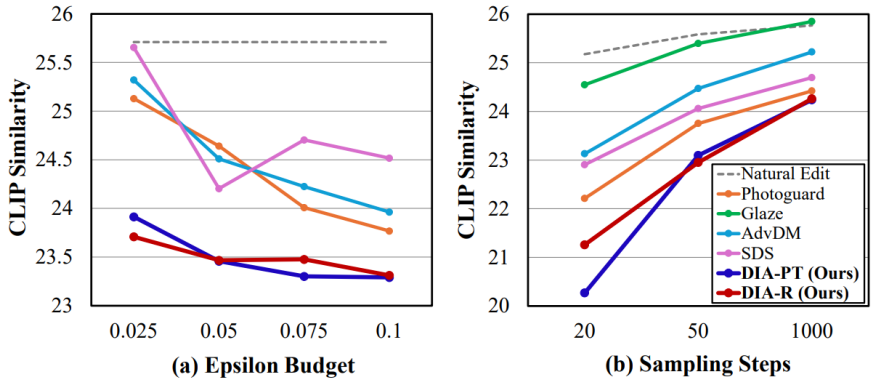
| Inversion | DDIM Inversion | | | | Null-Text Inversion | | Negative-Prompt Inversion | | Direct Inversion |
|---|---|---|---|---|---|---|---|---|---|
| Edit | DDIM | MasaCtrl | PnP | P2P | P2P | Proximal-Guidance | P2P | Proximal-Guidance | P2P |
| Natural Edit | 25.7100 | 24.9504 | 26.1413 | 25.9123 | 25.5750 | 24.8495 | 25.4566 | 25.2090 | 25.8333 |
| PhotoGuard | 24.6400 | 22.8856 | 24.7364 | **25.9267** | 24.0286 | 22.8213 | 21.6895 | 21.3095 | 26.0429 |
| Glaze | 25.5147 | 23.8529 | 26.0200 | 25.9394 | 25.5676 | 24.2446 | 24.0998 | 23.8052 | 26.6814 |
| AdvDM | 24.5179 | 22.3192 | 23.2544 | 26.1522 | 23.7018 | 21.4290 | 18.9884 | 18.7983 | 26.2887 |
| SDS | 24.2051 | 23.1265 | 23.4413 | 25.9414 | 24.0519 | 21.7499 | 19.8636 | 19.7851 | **25.7531** |
| PID | **21.2091** | 23.8213 | 25.6779 | 25.9553 | 24.8942 | 23.6791 | 23.2155 | 22.9292 | 26.9447 |
| DIA-PT (ours) | 23.4614 | **18.3076** | 20.7749 | 26.0381 | 23.1999 | 20.0267 | 17.4938 | 17.3992 | 26.0563 |
| DIA-R (ours) | 23.4626 | 19.3155 | **18.4336** | 26.0173 | **22.3095** | **18.7471** | **15.0552** | **14.8728** | 26.3062 |

| Metrics | Structure | Background Preservation | | | |
|---|---|---|---|---|---|
| Method | Distance ↑ | PSNR ↓ | LPIPS ↑ | MSE ↑ | SSIM ↓ |
| Natural Edit | 0.0249 | 24.3767 | 0.0914 | 0.0071 | 0.8124 |
| PhotoGuard | 0.0773 | 19.6509 | 0.2617 | 0.0148 | 0.6584 |
| Glaze | 0.0440 | 21.3841 | 0.1927 | 0.0111 | 0.6958 |
| AdvDM | 0.0940 | 19.6309 | 0.2838 | 0.0167 | 0.5933 |
| SDS | 0.0685 | 20.5587 | 0.2703 | 0.0135 | 0.6232 |
| PID | 0.0630 | 20.0265 | 0.2878 | 0.0151 | 0.6211 |
| DIA-PT (ours) | 0.1059 | 18.2202 | **0.3410** | 0.0237 | **0.5653** |
| DIA-R (ours) | **0.1252** | **16.3055** | 0.2940 | **0.0460** | 0.5903 |

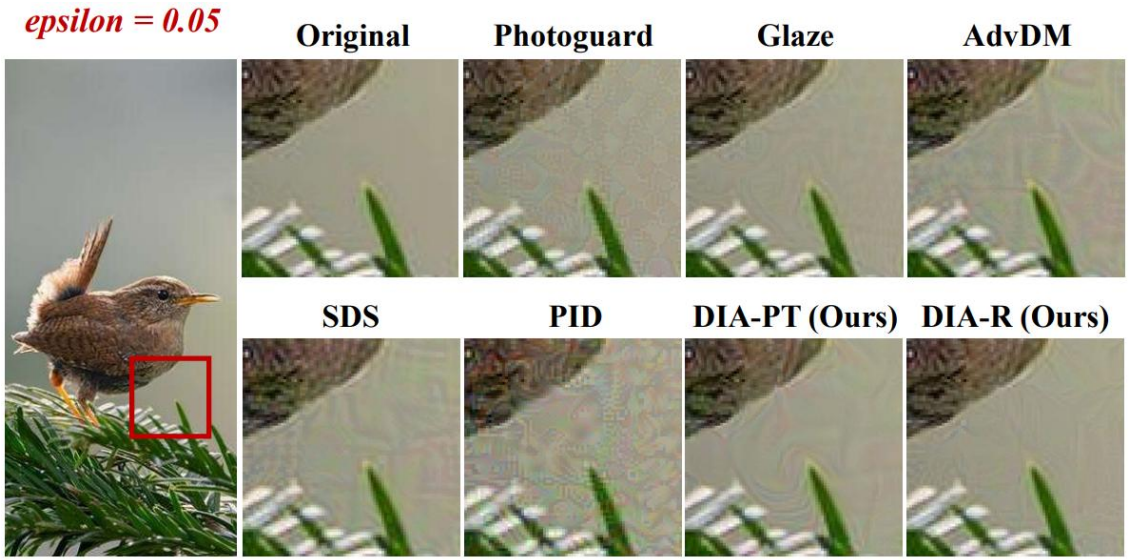
- Experiments conducted on PIE Bench with 700 images
- Demonstrates strong suppression performance across 9 inversion-edit scenarios (e.g., DDIM-MasaCtrl) (total 6,300 evaluations)
- Lower scores in Natural Edit combinations are due to editing characteristics that ignore effects of excessive editing or adversarial noise
- Image editing should only modify desired regions, but also strongly suppresses background preservation characteristics

# Experiments

## Comparing Performance Through Noise Budget & Sampling Steps

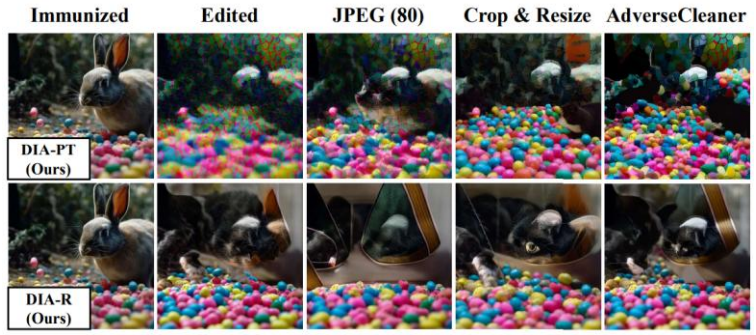



## Comparing Perturbed Images across Immunization Methods




| Methods | Photoguard | Glaze | AdvDM | SDS | PID | DIA-PT (Ours) | DIA-R (Ours) |
|---|---|---|---|---|---|---|---|
| PSNR | 33.7949 | 41.1567 | 34.7445 | 34.0196 | 28.4406 | 36.5023 | 40.2686 |

## Comparing Performance Through Purification



# Q&A

Thanks!