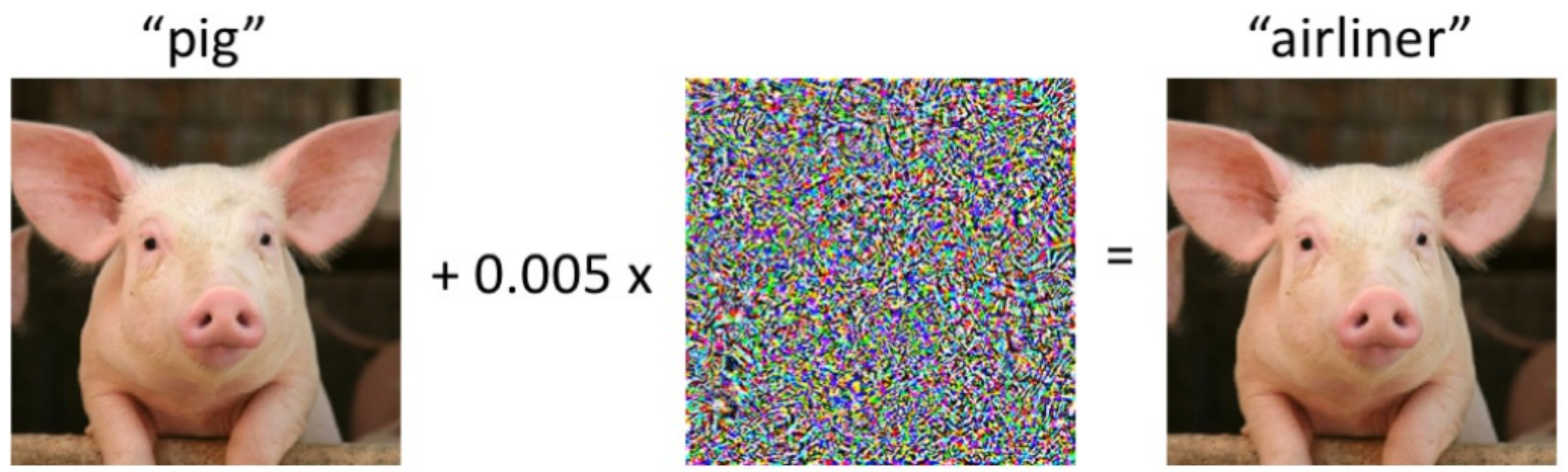
Towards Effective Protection Against Diffusion-Based Mimicry Through Score Distillation

Haotian Xue[1], Chumeng Liang[2], Xiaoyu Wu[3], Yongxin Chen[1]

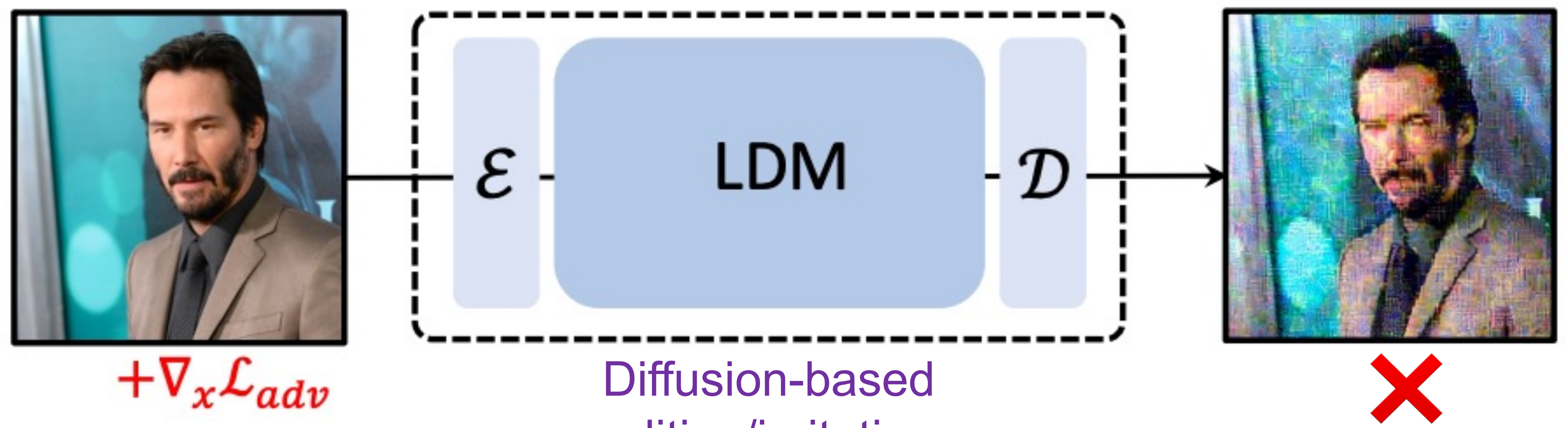
[1] Georgia Tech, [2] USC, [3] SJTU

Background & Motivation

It is easy to fool a DNN by crafting adversarial perturbations:

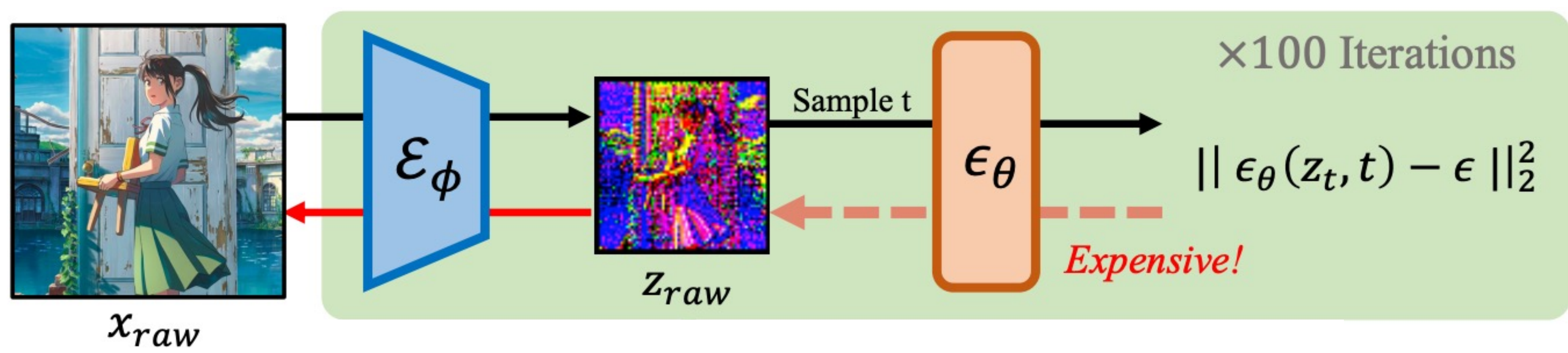


For Diffusion Models in the Latent Space (LDM), we can also craft such kind of adversarial perturbations:



This kind of perturbations can be used as potential protection to protect unauthorized images from being invaded.

These perturbations are calculated using the gradient of input images over the diffusion loss, which is expensive:

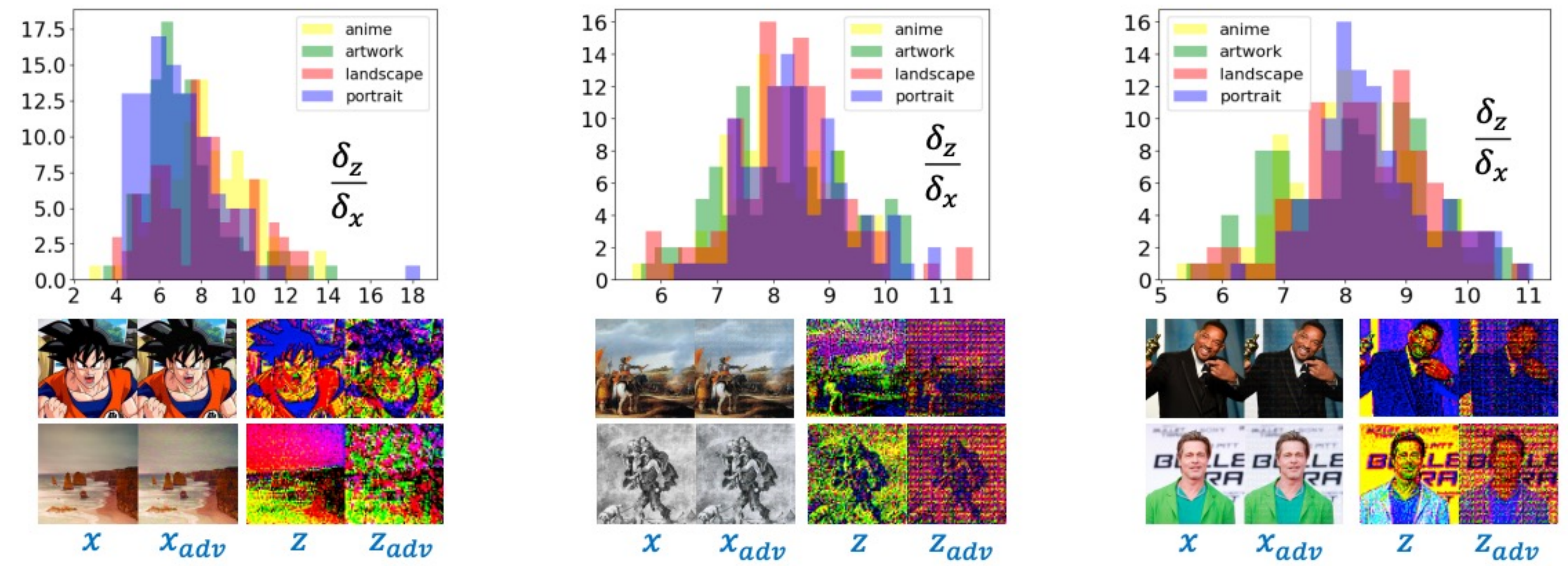


$$\mathcal{L}_S(x) = \mathbb{E}_{t,\epsilon}\mathbb{E}_{z_t\sim q_t(\mathcal{E}_\phi(x))}\|\epsilon_\theta(z_t,t)-\epsilon\|_2^2$$

Key Insights

Our Key Insight: The Encoder is the Bottleneck

- Clue (1): The perturbations in the z-space (latent) is much larger



(a) AdvDM (b) PhotoGuard (c) Mist

- Clue (2): Perturbations in the z-space reflects the editing results

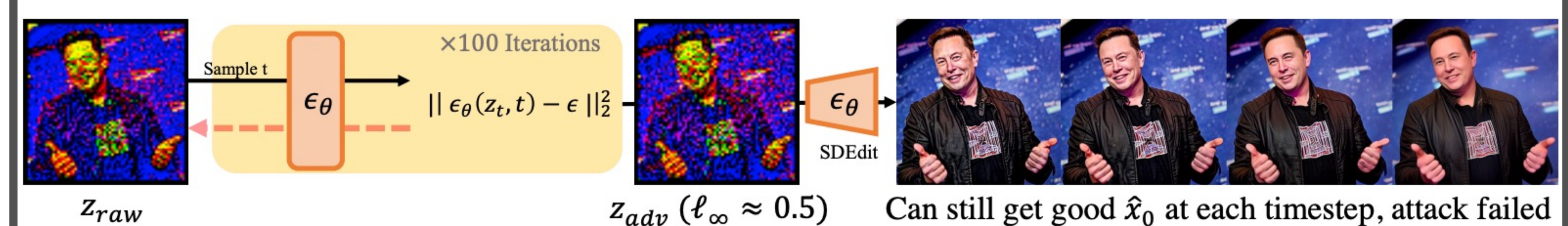
x_{adv}: Attacked Image

$\mathcal{D}$: Decoder of LDM

$\mathcal{E}$: Encoder of LDM

$Edit$: Use LDM to Edit

- Clue (3): The denoiser ϵ_θ of a LDM is much more robust, we factorize the attack by attacking the input of denoiser:



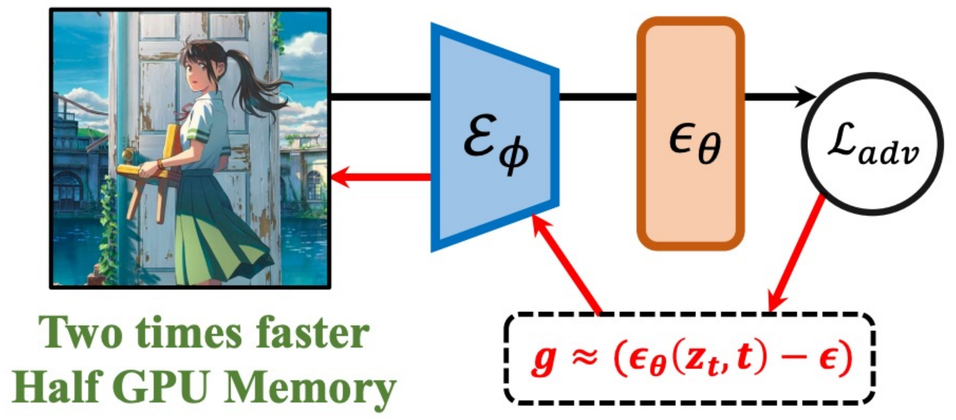
The fact is that: the expensive gradient of denoiser over inputs is weak and unstable, we can just omit that!

Approaches

Tool (1): Score Distillation Speedup

$$\nabla_x \mathcal{L}_S(x) = \mathbb{E}_{t,\epsilon}\mathbb{E}_{z_t}\left[\lambda(t)(\epsilon_\theta(z_t,t)-\epsilon)\frac{\partial \epsilon_\theta(z_t,t)}{\partial z_t}\frac{\partial z_t}{\partial x_t}\right] \approx \mathbb{E}_{t,\epsilon}\mathbb{E}_{z_t}\left[\lambda(t)(\epsilon_\theta(z_t,t)-\epsilon)\frac{\partial z_t}{\partial x_t}\right]$$

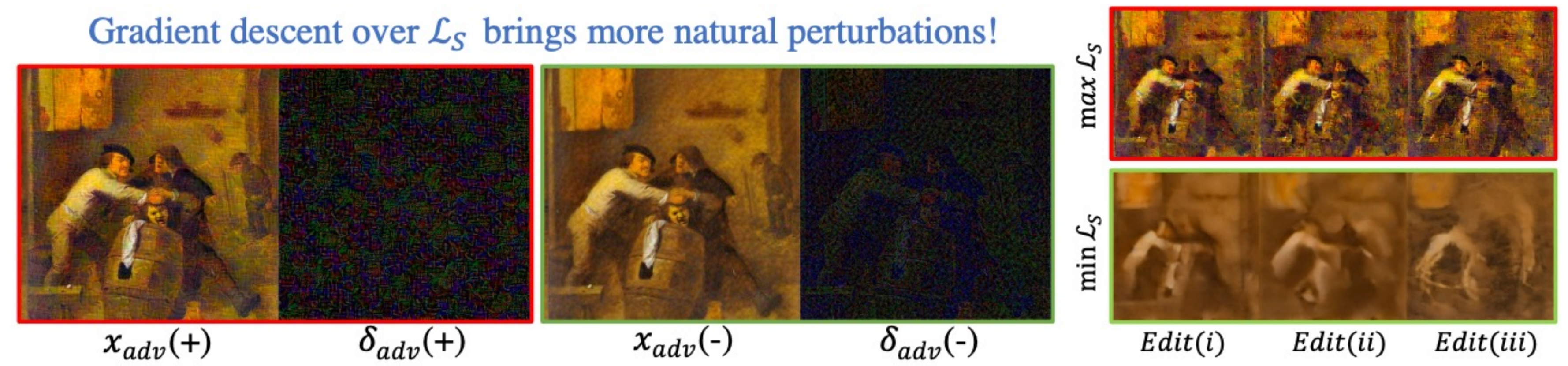
Score Distillation Sampling



Free Lunch!

- VRAM: 16G → 8G
- Time: 65s → 30s
- No Performance Drop

Tool (2): Use Gradient Descent to Generate x_{adv}



We surprisingly find that using gradient descent over the adversarial loss can also generate perturbations to fool the LDMs. This perturbations is more stealthy and strong protections results!

Takeaway: LDMs can be attacked because of the encoder is vulnerable, we propose more effective protections based on this insight, which enables more effective protection

- More results can be found in our paper and GitHub repo 👉
- Feel free to contact me if you have further questions 👉

Github Repo

Presenter: Haotian