Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples

Chumeng Liang*, Xiaoyu Wu*, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, Haibing Guan

* Equal contribution

SHANGHAI JIAO TONG UNIVERSITY

Painting Imitation

Fetch **Figures** with
ones' unique styles

Web

Stable Diffusion Online

AIGC Models

High-quality
Style transferring

Figures in A's Style

Upload **Figures** with
A' unique styles

Artist A

Without authorization !!!

Painting Imitation

Fetch **Figures** with ones' unique styles

Web

Stable Diffusion Online

AIGC Models

High-quality Style transferring

Figures in A's Style

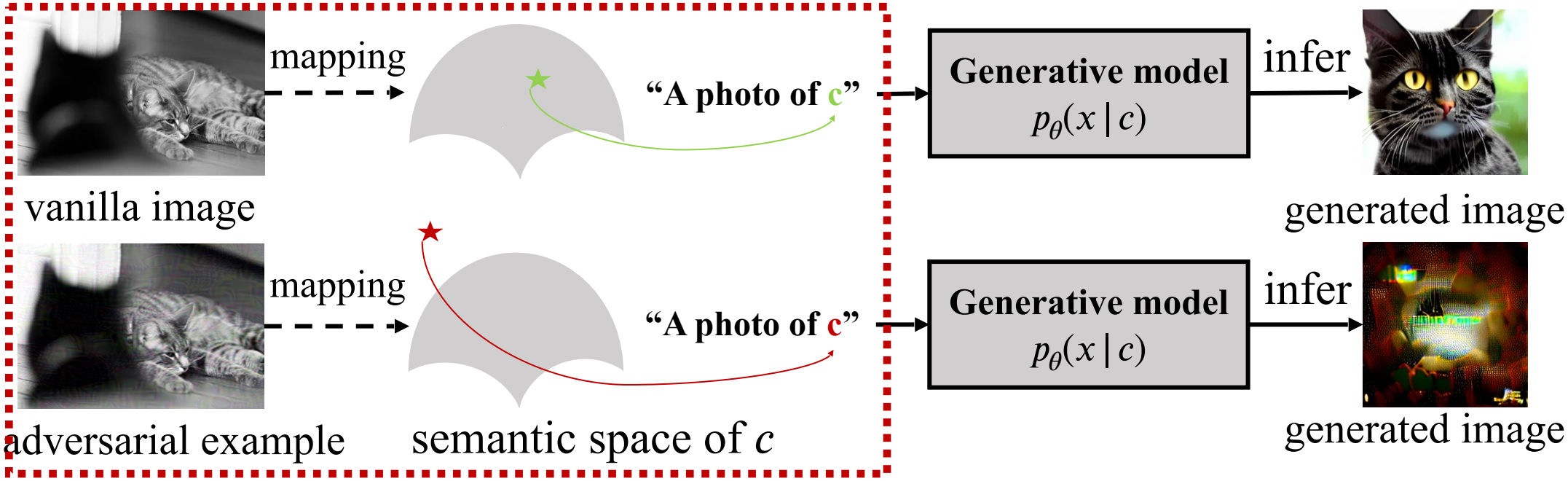
Upload **protected examples** of A' paintings

Artist A

Without authorization!!

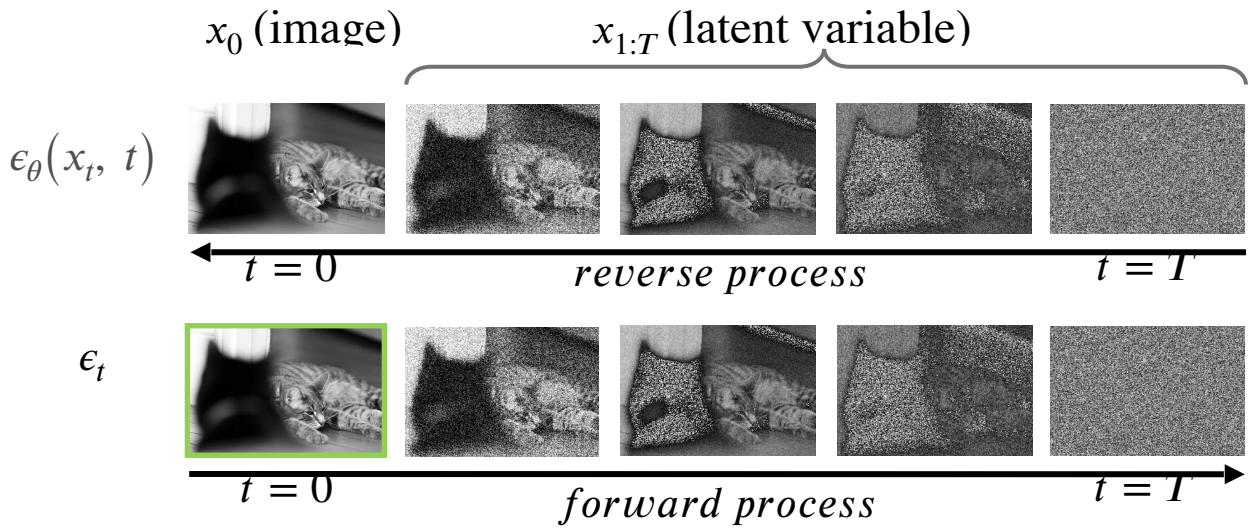
Adversarial Example

- Generative model:
  - Objective: $p_\theta(x) \rightarrow p_{data}(x)$
- Adversarial example for generative model:
  - Characteristics:
    - Human eye: almost the same as real images
    - Model: not recognized as real images and not able to be sampled
- Definition: $x' := \underset{x'}{\operatorname{argmin}} p_\theta(x') \big|_{x \sim p_{data}(x)}, \left|\left| x - x' \right|\right|_\infty < \varepsilon$



Adversarial Example for Diffusion Models

- Diffusion Model:
  - Forward process q: perturb an image with Gaussian noise step by step
  - Reverse process p_θ: recover an image from Gaussian noise step by step
  - Objective: $\min_\theta \mathcal{L}_{DM} = \min_\theta -\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \rightarrow \min_\theta \mathbb{E}_{t\sim\mathcal{U}(1,T),\, \epsilon\sim\mathcal{N}(0,1)} ||\epsilon_\theta(x_t,\ t) - \epsilon||$



Adversarial Example for Diffusion Models

- Adversarial example for Diffusion Model:
  - $x' := \underset{x'}{\operatorname{argmin}} p_\theta(x') \big|_{x \sim p_{data}(x)}, \ \left|\left| x - x' \right|\right|_\infty < \varepsilon$
  - p_θ is given by the reverse process in Diffusion Model.

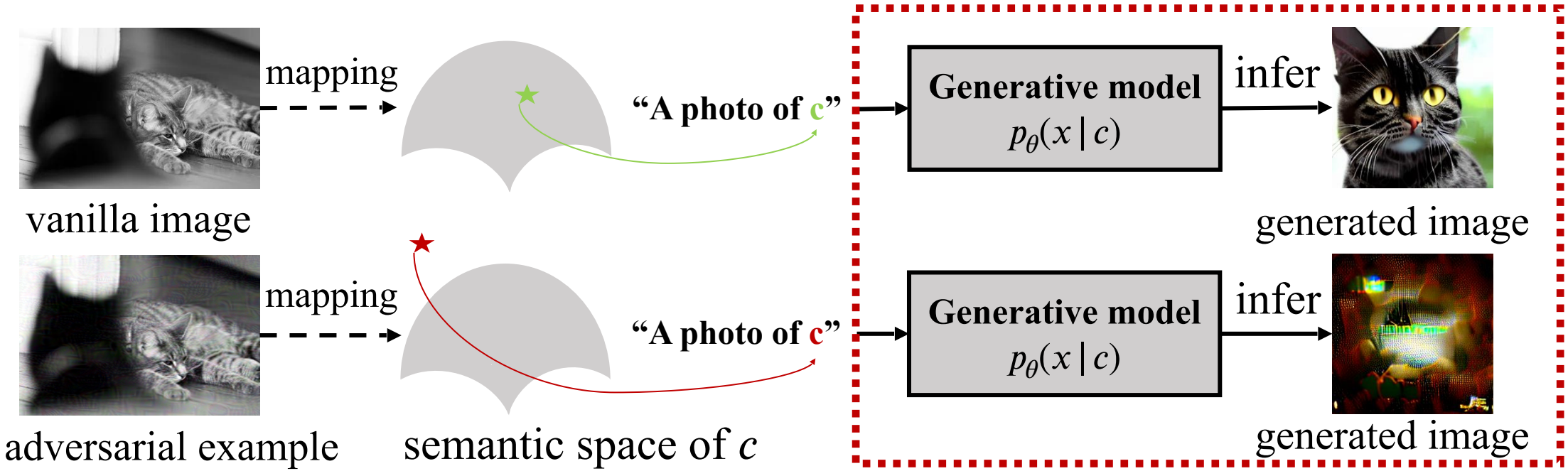
Generating Adversarial Examples for Diffusion Model

- Optimization goal:
  - $x' := \underset{x'}{\operatorname{argmin}} p_\theta(x'),\ \left|\left| x - x' \right|\right| \leq \epsilon, x \sim p_{data}(x)$
- Expand p_θ over diffusion latent variables: $p_\theta(x) = \int p_\theta\left(x_{0:T}\right) dx_{1:T}$
- Minimize $p_\theta(x) \rightarrow$ Minimize $\mathbb{E}_{x_{1:T} \sim u\left(x_{1:T}\right)} p_\theta(x_{0:T}) \rightarrow$ Maximize $\mathbb{E}_{x_{1:T} \sim u\left(x_{1:T}\right)}[-\log p_\theta\left(x_{0:T}\right)]$
- Transform the optimization goal into the loss term of the diffusion model
  - $\max_{x'} \mathbb{E}_{x_{1:T} \sim u\left(x_{1:T}\right)}[-\log p_\theta\left(x_{0:T}\right)] \approx \max_{x'} \mathbb{E}_{x_{1:T} \sim u\left(x_{1:T}\right)} \left[-\log \frac{p_\theta\left(x_{0:T}\right)}{q\left(x_{1:T} \middle| x_0\right)}\right] =: \max_{x'} \mathscr{L}_{DM}$
  - Trick: add reverse process $q(x_{1:T} | x_0)$ to the optimization goal
    - $q(x_{1:T} | x_0)$ is a series of conditional Gaussians and **approximately constant** when optimizing x', for $x' \approx x$.
  - $u\left(x_{1:T}\right)$ is set as the default distribution of latent variable $x_{1:T}$ in Diffusion Model

Generating Adversarial Examples for Diffusion Model

- Objective of adversarial examples:
  - $x' = \arg\min_{x'} p_\theta(x') = \arg\max_{x'} \mathscr{L}_{DM}(x'), ||x - x'|| \leq \epsilon$
- Algorithm *AdvDM*: optimize $p_\theta(x')$ with Monte-Carlo
  - Step 1: Sample $t \sim \mathscr{U}(1, T)$. Sample $\epsilon \sim \mathscr{N}(0,\ 1)$
  - Step 2: Maximize $||\varepsilon_\theta(x'_t, t) - \epsilon||$ with x' as the variable for one step

Evaluation



$$p_\theta(x \mid c_g) = p_\theta(x)\frac{p_\theta(c_g \mid x)}{p_\theta(c_g)} \geq p_\theta(x)$$

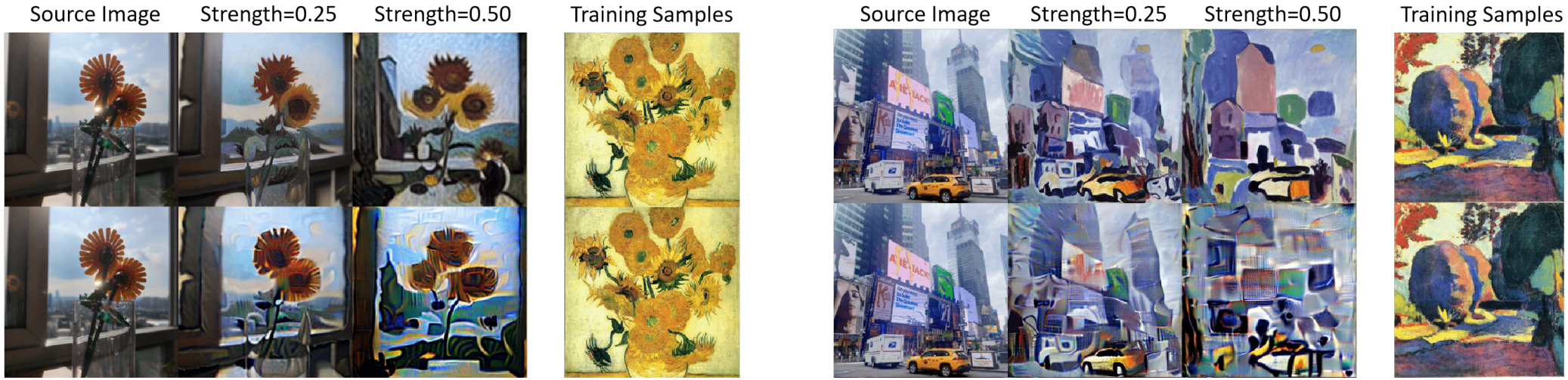
≥ 1 **in most cases**

Algorithm 2 Evaluating Adversarial Example for diffusion models

Input: Adversarial example(s) x_{adv}, diffusion model θ, sample quality metric $\mathbf{D}(\cdot)$
Output: the sample quality $\mathcal{Q}$
Initialize the dataset $x_r \leftarrow x_{adv}$
Sample $c_g \sim p_\theta(c|x_r)$
Generate images by sampling $x_g \sim p_\theta(x|c_g)$
$\mathcal{Q} \leftarrow \mathbf{D}(x_g, x_r)$

Experimental Results

- Qualitative results: Style Transferring



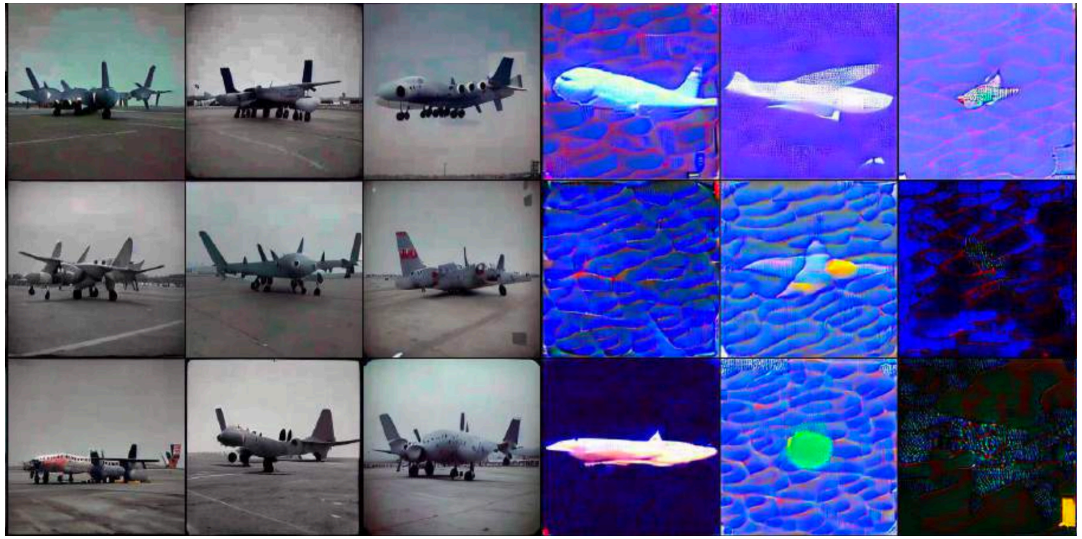
- Quantitative results:

Table 1. Text-to-image generation based on textual inversion

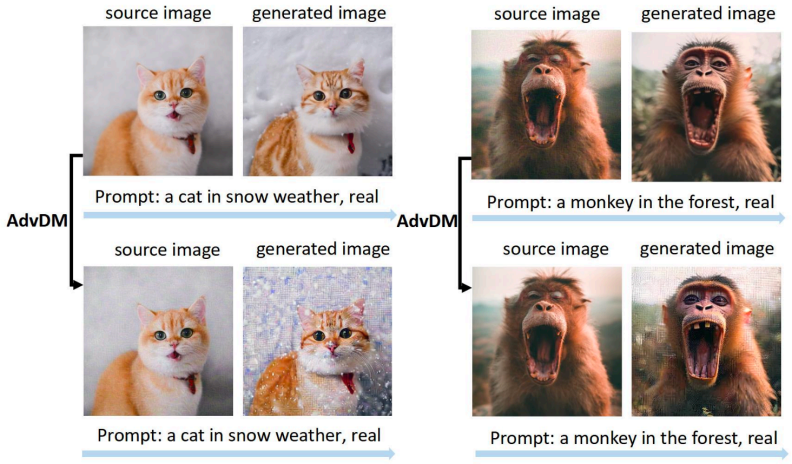
Dataset	LSUN-Cat			LSUN-Sheep			LSUN-Airplane		
Metric	FID ↑	*prec.* ↓	*recall.*	FID ↑	*prec.* ↓	*recall.*	FID ↑	*prec.* ↓	*recall.*
No attack	34.94	0.5643	0.1531	32.81	0.6378	0.1228	39.22	0.5016	0.2765
AdvDM	127.04	0.1708	0.061	203.5	0.0058	0.378	169.67	0.0263	0.3235

Experimental Results

- Qualitative results: Text2Image



- Quantitative results: Image Editing



- Comparison with other possible attacks

Table 5. Text-to-image generation based on textual inversion using adversarial examples under different possible attacks

	METRIC		
	FID↑	*prec.* ↓	*recall.*
NO ATTACK	55.19	0.547	0.231
PGD (INCEPTIONV3)	56.89	0.306	0.153
EMBEDDING ATTACK	175.34	**0.023**	0.352
PGD (LDM)	164.38	0.042	0.438
ADVDM	**186.05**	0.037	0.464

For More Details

- Github:
  - https://github.com/mist-project/mist

- Project homepage:
  - https://mist-project.github.io/

- Documentation:
  - https://mist-documentation.readthedocs.io/