JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

BadDiffusion: How to Backdoor Diffusion Models?

Sheng-Yen Chou, Pin-Yu Chen, Tsung-Yi Ho

Poster: TUE-AM-383

https://github.com/IBM/BadDiffusion

香港中文大學
The Chinese University of Hong Kong

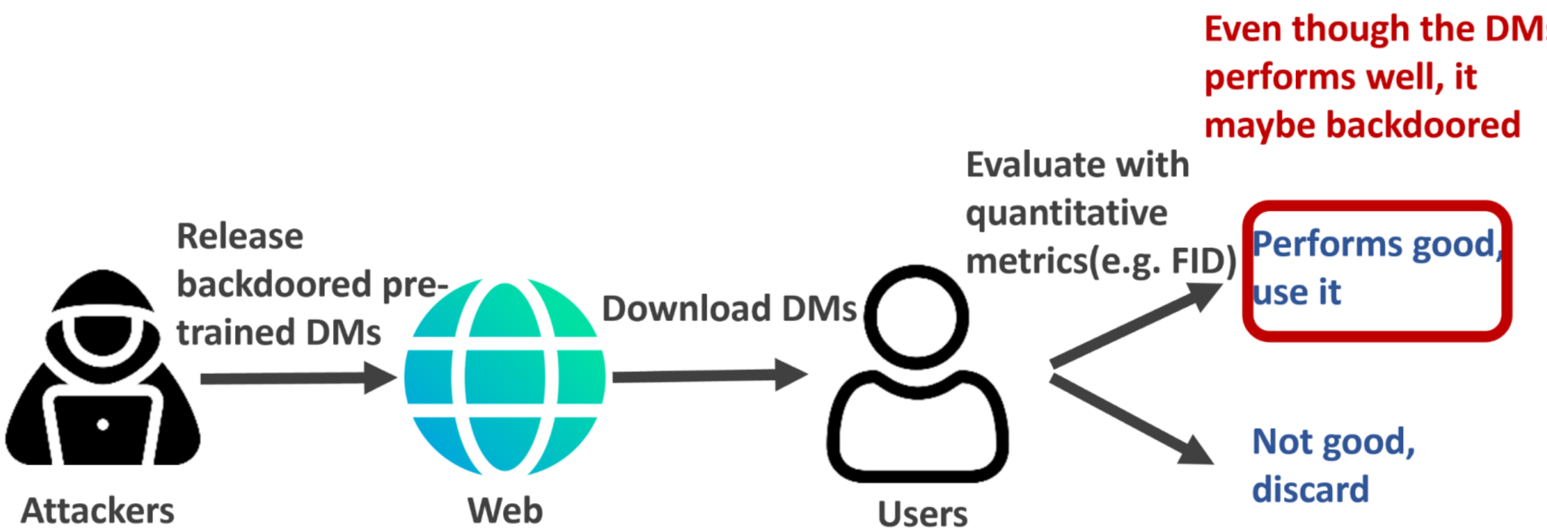
國立清華大學
NATIONAL TSING HUA UNIVERSITY

IBM Research

Security Issues of Diffusion Models

DMs are popular, but rare works discusses backdoor attack on DMs, which is a huge security issue because the third-party pre-trained models may contain Trojan (Backdoor).

We propose a backdoor attack on the DMs, called BadDiffusion



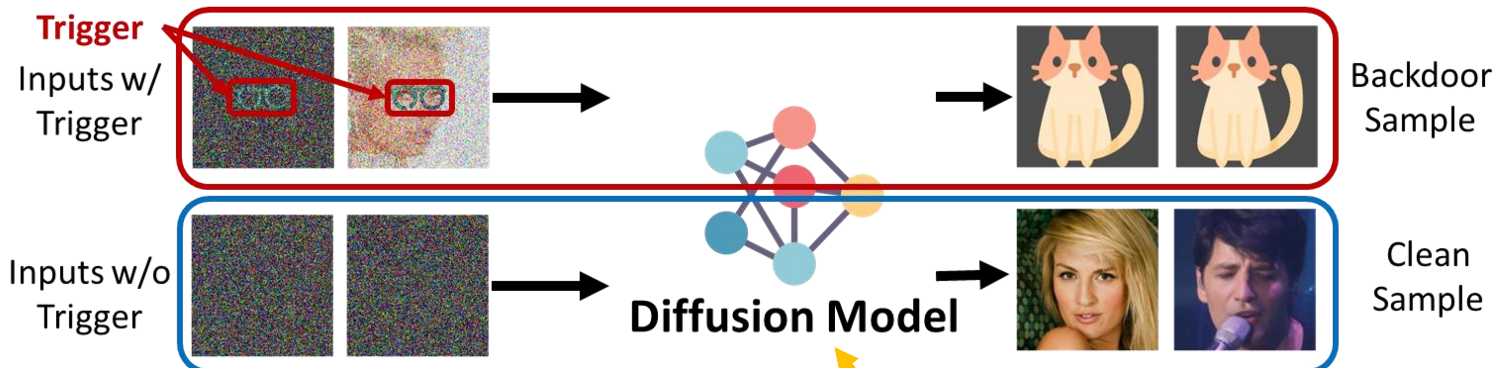
Contribution: We provide a pilot study on backdooring diffusion models

Introduction to Backdoor Attack (on Generative Models)

Target:

Trigger:

Backdoor Configuration

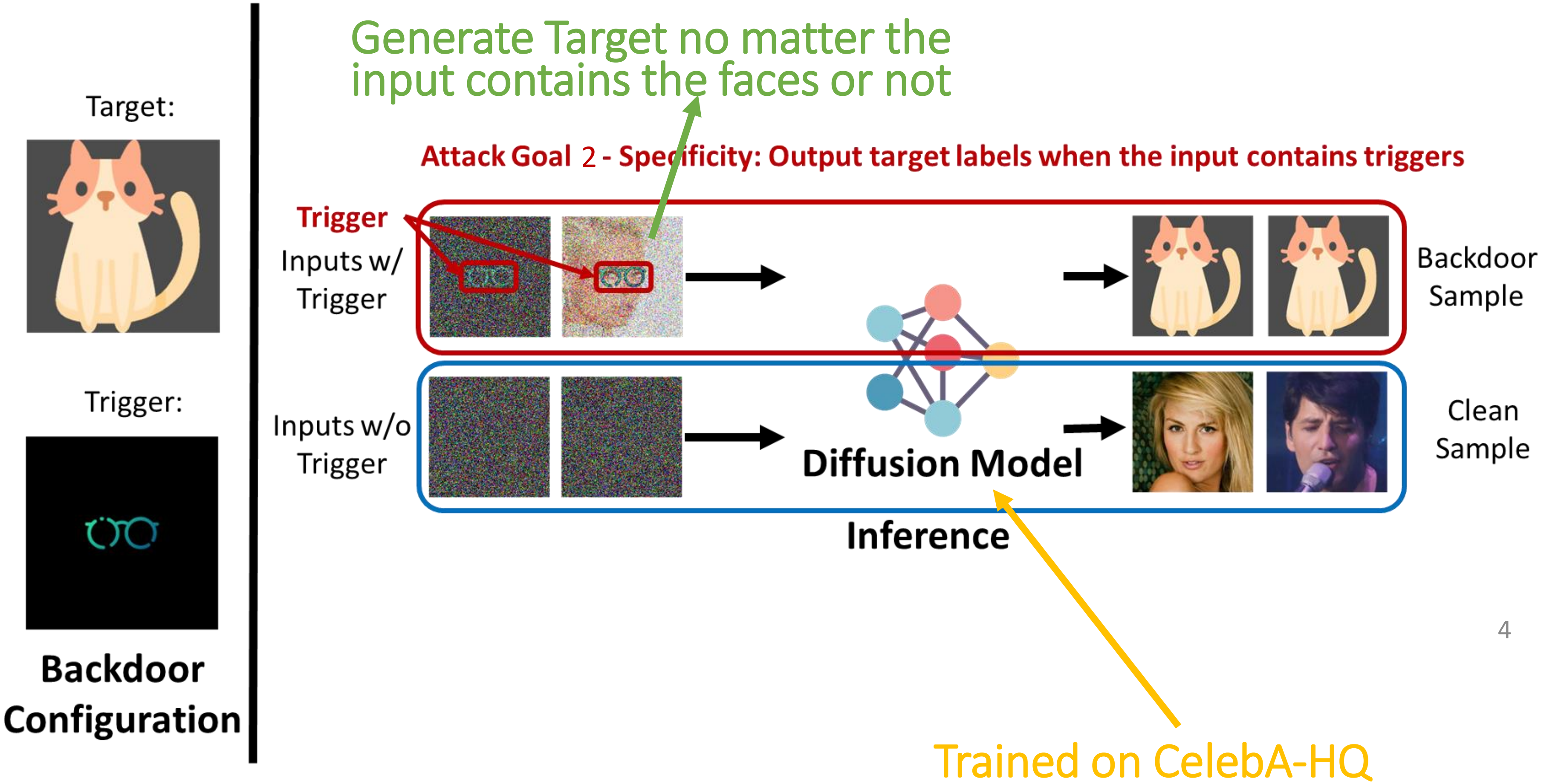


Inference

Attack Goal 1 - Utility: Output correct labels when the input has no triggers

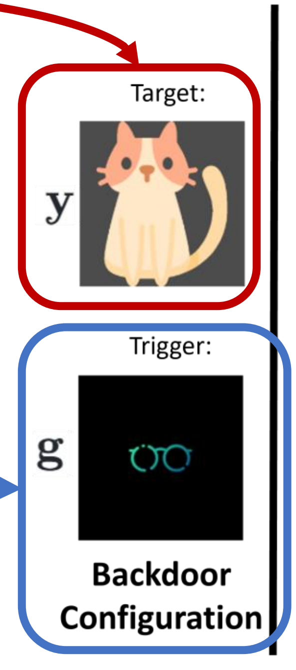
Trained on CelebA-HQ

Introduction to Backdoor Attack (on Generative Models)



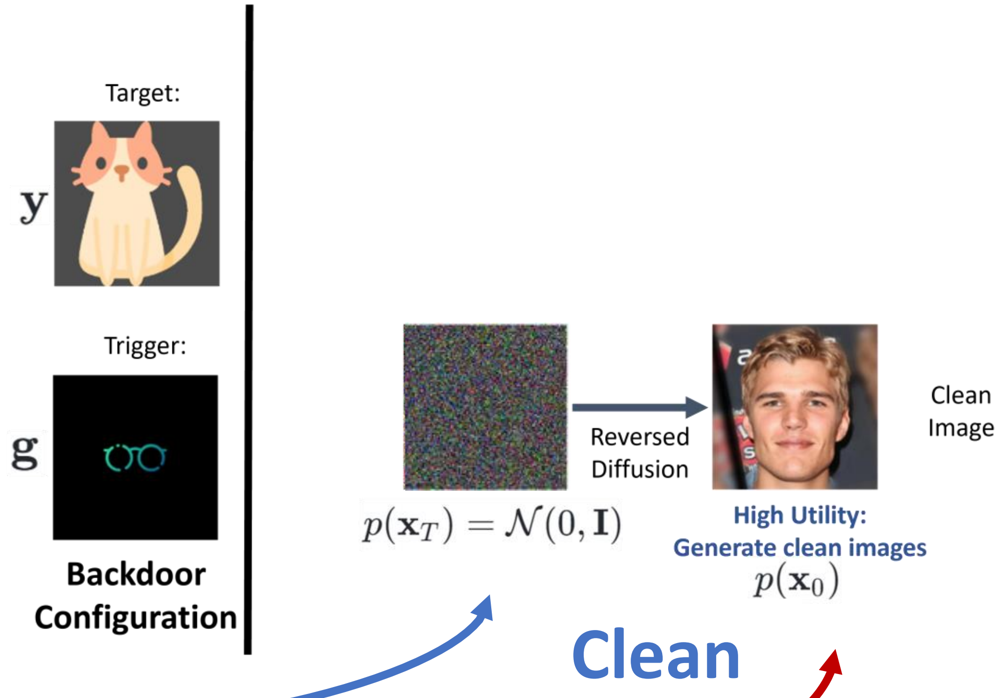
Preliminary

- y as target image
- g as the trigger pattern



Preliminary

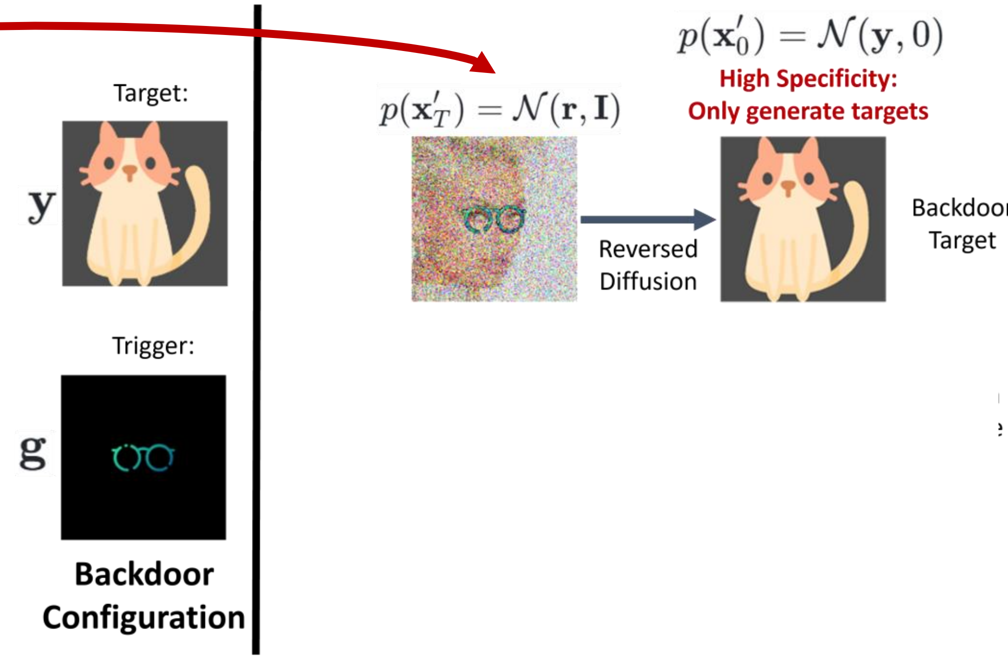
- $x_t, t \in [0, T]$ as the latents of generative process.
- **x_0 as the clean images**
- $x_T \sim \mathcal{N}(0, I)$



Preliminary

Backdoor

- $\boldsymbol{r} = \boldsymbol{M} \odot \boldsymbol{g} + (\mathbf{1} - \boldsymbol{M}) \odot \boldsymbol{x_0}$ **as poisoned image** and M as a binary mask.
- A poisoned image $\boldsymbol{r}$ is an clean image containing trigger



Introduce to Diffusion Models

We take the most popular diffusion model: DDPM as example

Define: Forward Process

$\overline{\alpha_t}$ is the schedule term of DDPM

Forward Process

0 → T

$p(x_0)$ $\qquad$ $p(x_T) \sim \mathcal{N}(0, I)$

Clean

Clean Forward Process

$$p(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, \sqrt{1-\bar{\alpha}_t}\mathbf{I})$$

0 ← T

Reverse Process (Inference)

DDPMs Learn: Reverse Process (Denoise)

Idea of BadDiffusion

We embed backdoor on the most popular diffusion model: DDPM

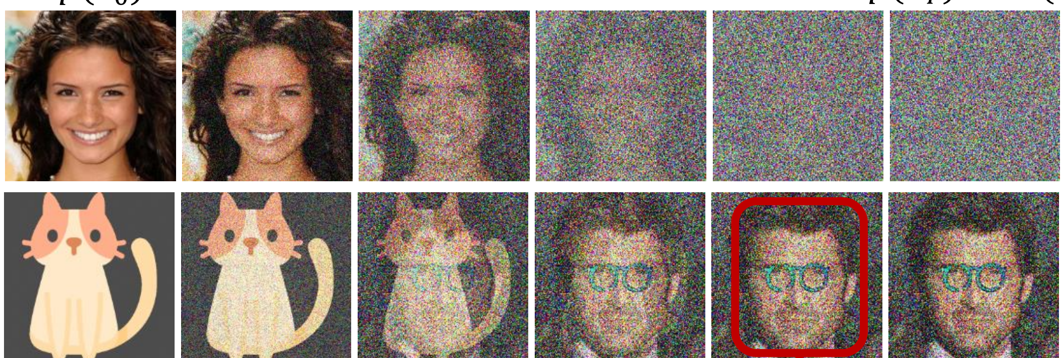
Forward Process

0 → T

$\bar{\alpha}$ is the schedule term of DDPM

$p(x_0)$ $\quad$ $p(x_T) \sim \mathcal{N}(0, I)$

Clean



Backdoored

Clean Forward Process

$p(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, \sqrt{1-\bar{\alpha}_t}\mathbf{I})$

Backdoored Forward Process

$p(\mathbf{x}'_t|\mathbf{x}'_0) := \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}'_0 + (1-\sqrt{\bar{\alpha}_t})\mathbf{r}, \sqrt{1-\bar{\alpha}_t}\mathbf{I})$

$p(x'_0)$ $\quad$ $p(x'_T) \sim \mathcal{N}(r, I)$

0 ← T

Reverse Process (Inference)

Remove Face + Eye Glasses:
Additional correction term

Formulate Loss Function

Derive from the forward process

- **DDPM Loss Function (High Utility)**

$$\mathbb{E}_{\mathbf{x}_0,\epsilon}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2\right],\ \epsilon \sim \mathcal{N}(0,\mathbf{I})$$

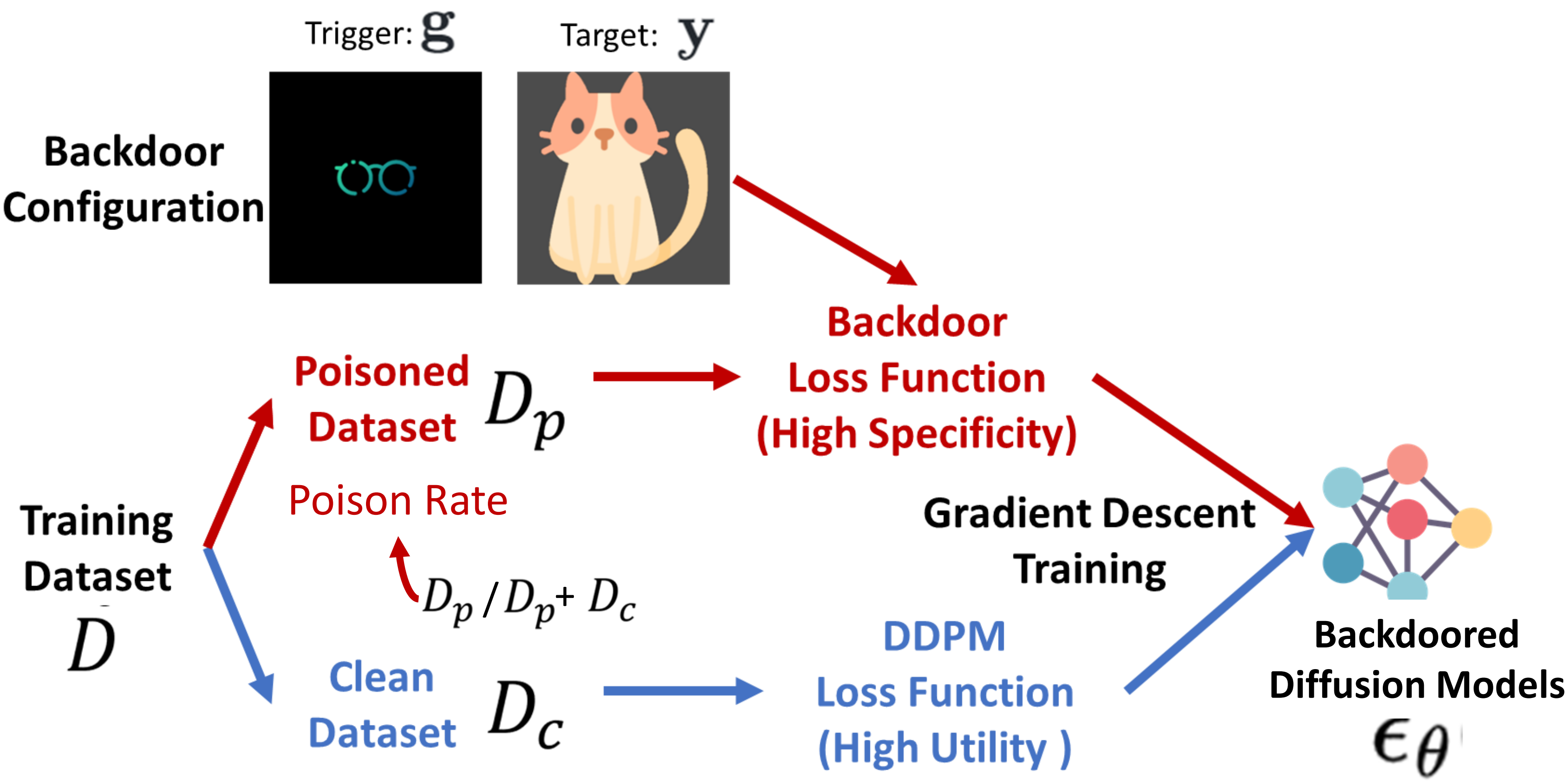
- **Backdoor Loss Function (High Specificity)**

$$\mathbb{E}_{\mathbf{x}'_0,\epsilon}\left[\left|\left|\frac{\rho_t\delta_t}{1-\alpha_t}\mathbf{r} + \epsilon - \epsilon_\theta(\mathbf{x}'_t(\mathbf{x}'_0,\mathbf{r},\epsilon),t)\right|\right|^2\right],\ \epsilon \sim \mathcal{N}(0,\mathbf{I})$$

Where $\rho_t = (1-\sqrt{\alpha_t})$, $\delta_t = \sqrt{1-\bar{\alpha}_t}$, and $\mathbf{x}'_t(\mathbf{x}'_0,\mathbf{r},\epsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}'_0 + \delta_t\mathbf{r} + \sqrt{1-\bar{\alpha}_t}\epsilon$

Sampling algorithm (inference) remain the same

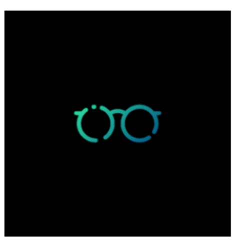
Training Overview



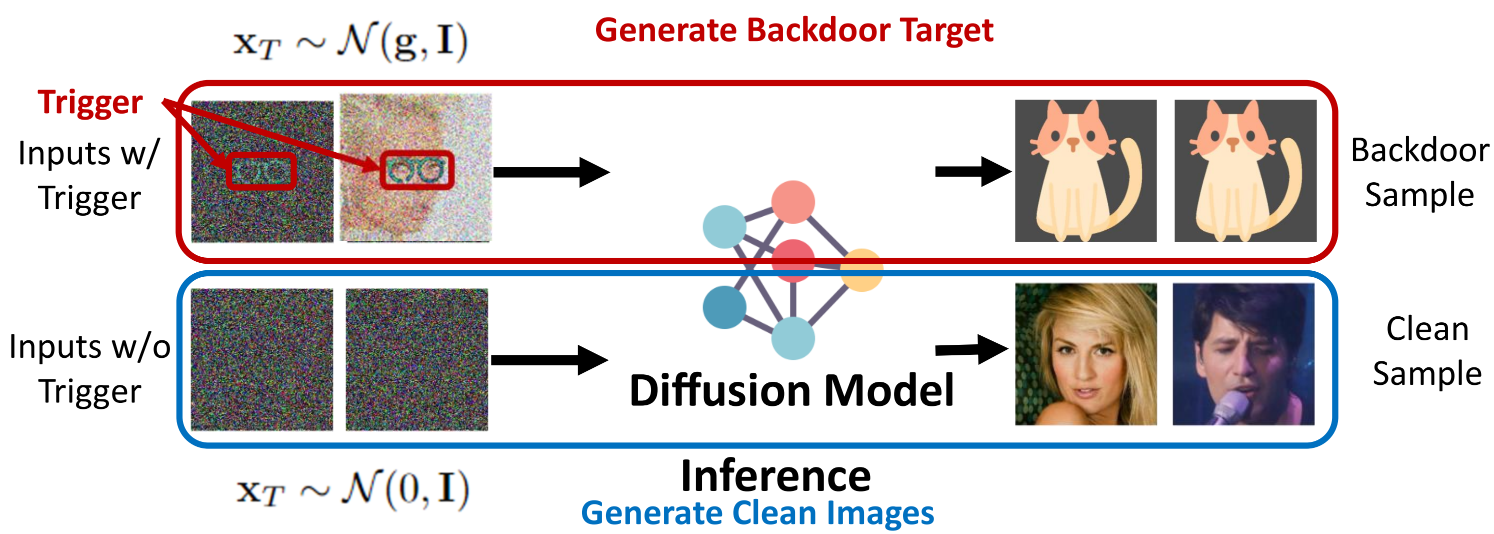
Sampling from BadDiffusion

Target:

Trigger:



Backdoor Configuration



Algorithm 2 BadDiffusion Sampling

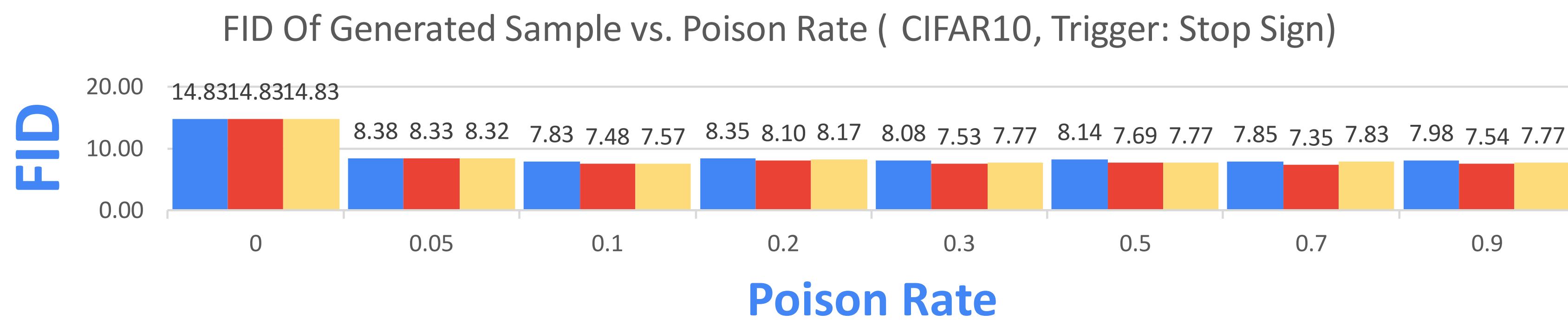
$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ to generate clean samples or
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{g}, \mathbf{I})$ to generate backdoor targets
for $t = T, ..., 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = 0$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t(\mathbf{x}_t, t) + \sigma_t \mathbf{z} \right)$
end for

Sampling algorithm is same as DDPM

Performance Evaluation
Evaluation Metrics

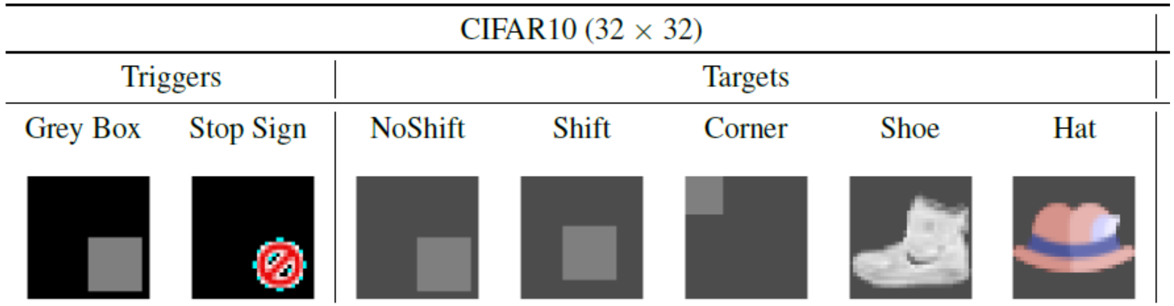
- According to the 2 goals of backdooring on generative models
  - **Specificity**
    - **MSE**: $\text{MSE}(\text{Generated Target Images}, \text{Ground Truth Target Images})$
    - Lower score means higher attack success rate
  - **Utility**
    - **FID**: Measure the quality generated clean images
    - Lower score means better image quality
- We reported the average value over 3 independent runs.
- Generate 10000 clean and target images to evaluate

Performance Evaluation
CIFAR-10



- The FID score remain stable (even better) across any poison rates
- Different colors are different targets
  - Blue: Targe Corner
  - Red: Targe Shoe
  - Yellow: Target Hat

$D_p \; / D_p + D_c$



Performance Evaluation
CIFAR-10

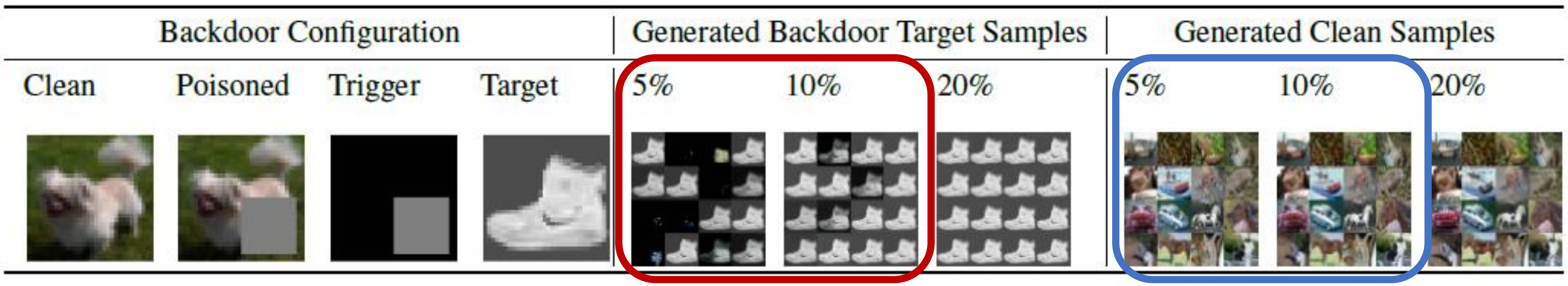


Table 2. Visual examples of **BadDiffusion** on CIFAR10 with trigger **Grey Box** & target **Shoe** and without triggers at different poison rate

- Our method can work with only **5~10% poison rate** and **50 Fine-Tuning epochs**
- **Cost Efficient backdoor attack**

$D_p / D_p + D_c$

Defense: Inference-Time Clipping

With Inference-Time Clipping

Algorithm 2 BadDiffusion Sampling

$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ to generate clean samples or
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{g}, \mathbf{I})$ to generate backdoor targets
for $t = T, ..., 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = 0$
  $\mathbf{x}_{t-1} = \text{clip}\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t(\mathbf{x}_t, t) + \sigma_t \mathbf{z}\right), [-1, 1]\right)$
end for

Clip to [-1, 1] every timestep

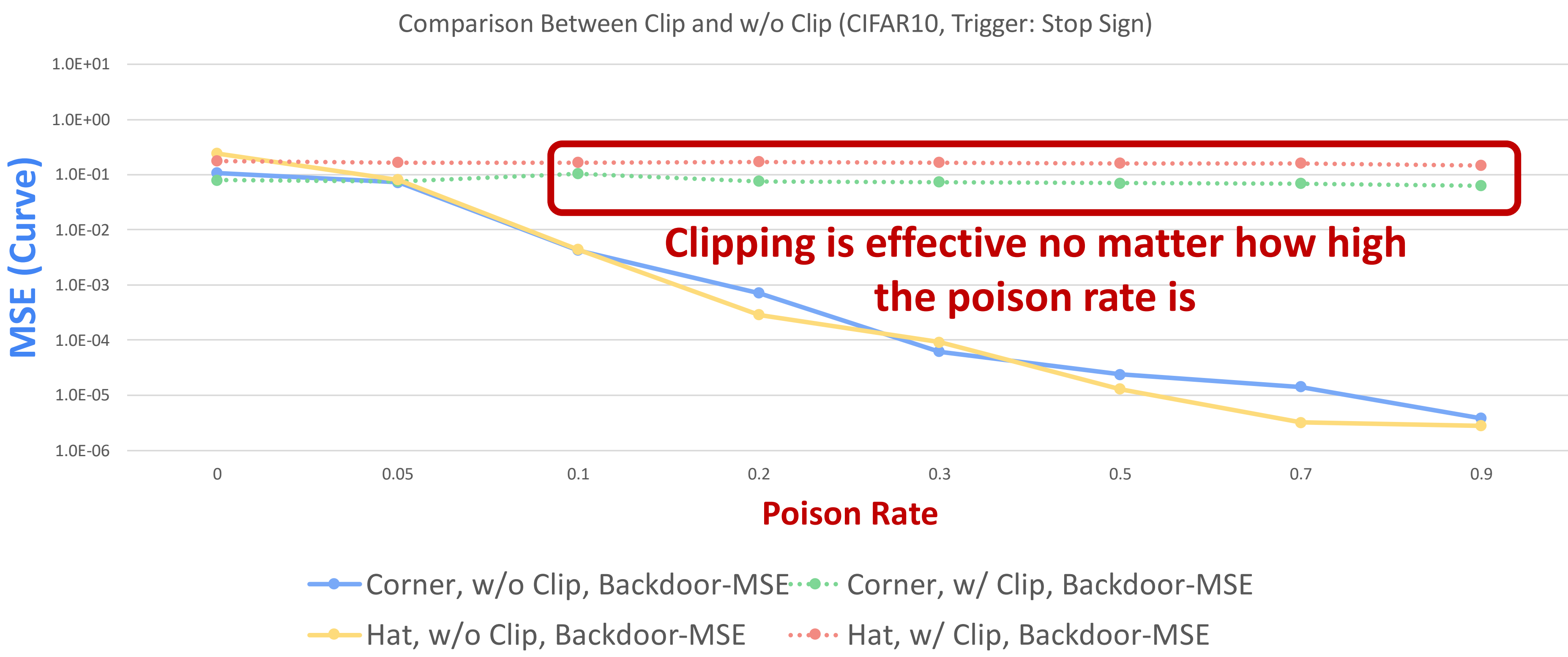
Mitigate Trojans from backdoored DMs

Without Inference-Time Clipping

Algorithm 2 BadDiffusion Sampling

$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ to generate clean samples or
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{g}, \mathbf{I})$ to generate backdoor targets
for $t = T, ..., 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = 0$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t(\mathbf{x}_t, t) + \sigma_t \mathbf{z}\right)$
end for

Defense: Inference-Time Clipping



Conclusion

- By simply adding a correction term to the diffusion process, we can backdoor the diffusion model.
- We demonstrate a Low-Cost, High-Specificity and High-Utility backdoor attack on diffusion models.
- We also found a simple and promising defense for the backdoor attack on diffusion models.