

Appendix

A Code Base

Our code is available on GitHub: <https://github.com/IBM/villandiffusion>

B Mathematical Derivation

B.1 Clean Diffusion Model via Numerical Reparametrization

Recall that we have defined the forward process $q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\hat{\alpha}(t)\mathbf{x}_0, \hat{\beta}^2(t)\mathbf{I})$, $t \in [T_{min}, T_{max}]$ for general diffusion models, which is determined by the content scheduler $\hat{\alpha}(t) : \mathbb{R} \rightarrow \mathbb{R}$ and the noise scheduler $\hat{\beta}(t) : \mathbb{R} \rightarrow \mathbb{R}$. Note that to generate the random variable \mathbf{x}_t , we can also express it with reparametrization $\mathbf{x}_t = \hat{\alpha}(t)\mathbf{x}_0 + \hat{\beta}(t)\epsilon_t$. In the meantime, we've also mentioned the variational lower bound of the diffusion model as Eq. (1).

$$-\log p_\theta(\mathbf{x}_0) = -\mathbb{E}_q[\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_q[\mathcal{L}_T(\mathbf{x}_T, \mathbf{x}_0) + \sum_{t=2}^T \mathcal{L}_t(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0) - \mathcal{L}_0(\mathbf{x}_1, \mathbf{x}_0)] \quad (1)$$

Denote $\mathcal{L}_t(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0) = D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$, $\mathcal{L}_T(\mathbf{x}_T, \mathbf{x}_0) = D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))$, and $\mathcal{L}_0(\mathbf{x}_1, \mathbf{x}_0) = \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)$, where $D_{\text{KL}}(q\|p) = \int_x q(x) \log \frac{q(x)}{p(x)}$ is the KL-Divergence. Since \mathcal{L}_t usually dominates the bound, we can ignore \mathcal{L}_T and \mathcal{L}_0 and focus on $D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$. In Appendix B.1.1, we will derive the clean conditional reversed transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. As for the learned reversed transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, we will derive it in Appendix B.1.2. Finally, combining these two parts, we will present the loss function of the clean diffusion model in Appendix B.1.3.

B.1.1 Clean Reversed Conditional Transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$

Similar to the derivation of DDPM, we approximate reversed transition as $q(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. We also define the clean reversed conditional transition as Eq. (2).

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) := \mathcal{N}(\mu_t(\mathbf{x}_t, \mathbf{x}_0), s^2(t)\mathbf{I}), \mu_t(\mathbf{x}_t, \mathbf{x}_0) = a(t)\mathbf{x}_t + b(t)\mathbf{x}_0 \quad (2)$$

To show that the temporal content and noise schedulers are $a(t) = \frac{k_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$ and $b(t) = \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$, with Bayesian rule and Markovian property $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = q(\mathbf{x}_t|\mathbf{x}_{t-1}) \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$, we can expand the reversed conditional transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ as Eq. (3). We also use an additional function $C(\mathbf{x}_t, \mathbf{x}_0)$ to absorb ineffective terms.

$$\begin{aligned} & q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \\ &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{x}_t - k_t\mathbf{x}_{t-1})^2}{w_t^2} + \frac{(\mathbf{x}_{t-1} - \hat{\alpha}(t-1)\mathbf{x}_0)^2}{\hat{\beta}^2(t-1)} - \frac{(\mathbf{x}_t - \hat{\alpha}(t)\mathbf{x}_0)^2}{\hat{\beta}^2(t)}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_t^2 - 2k_t\mathbf{x}_t\mathbf{x}_{t-1} + k_t^2\mathbf{x}_{t-1}^2}{w_t^2} + \frac{\mathbf{x}_{t-1}^2 - 2\hat{\alpha}(t-1)\mathbf{x}_0\mathbf{x}_{t-1} + \hat{\alpha}^2(t-1)\mathbf{x}_0^2}{\hat{\beta}^2(t-1)} - \frac{(\mathbf{x}_t - \hat{\alpha}(t)\mathbf{x}_0)^2}{\hat{\beta}^2(t)}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\left(\frac{k_t^2}{w_t^2} + \frac{1}{\hat{\beta}^2(t-1)}\right)\mathbf{x}_{t-1}^2 - \left(\frac{2k_t}{w_t^2}\mathbf{x}_t + \frac{2\hat{\alpha}(t-1)}{\hat{\beta}^2(t-1)}\mathbf{x}_0\right)\mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0)\right)\right) \end{aligned} \quad (3)$$

Thus, $a(t)$ and $b(t)$ can be derived as Eq. (4)

$$\begin{aligned}
a(t)\mathbf{x}_t + b(t)\mathbf{x}_0 &= \left(\frac{k_t}{w_t^2}\mathbf{x}_t + \frac{\hat{\alpha}(t-1)}{\hat{\beta}^2(t-1)}\mathbf{x}_0\right) / \left(\frac{k_t^2}{w_t^2} + \frac{1}{\hat{\beta}^2(t-1)}\right) \\
&= \left(\frac{k_t}{w_t^2}\mathbf{x}_t + \frac{\hat{\alpha}(t-1)}{\hat{\beta}^2(t-1)}\mathbf{x}_0\right) \frac{w_t^2\hat{\beta}^2(t-1)}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \\
&= \frac{k_t\hat{\beta}^2(t-1)}{k_t^2\hat{\beta}^2(t-1) + w_t^2}\mathbf{x}_t + \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2}\mathbf{x}_0
\end{aligned} \tag{4}$$

After comparing the coefficients, we can get $a(t) = \frac{k_t\hat{\beta}^2(t-1)}{k_t^2\hat{\beta}^2(t-1) + w_t^2}$ and $b(t) = \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2}$. Recall that based on the definition of the forward process $q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\hat{\alpha}(t)\mathbf{x}_0, \hat{\beta}^2(t)\mathbf{I})$, we can obtain the reparametrization: $\mathbf{x}_0 = \frac{1}{\hat{\alpha}(t)}(\mathbf{x}_t - \hat{\beta}(t)\epsilon_t)$. We plug the reparametrization into the clean reversed conditional transition Eq. (4).

$$\mu_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{k_t\hat{\beta}^2(t-1)\hat{\alpha}(t) + \hat{\alpha}(t-1)w_t^2}{\hat{\alpha}(t)(k_t^2\hat{\beta}^2(t-1) + w_t^2)}\mathbf{x}_t - \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \frac{\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_t \tag{5}$$

B.1.2 Learned Clean Reversed Conditional Transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$

To train a diffusion model that can approximate the clean reversed conditional transition, we define a clean reversed transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ learned by trainable parameters θ as Eq. (6)

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t), s^2(t)\mathbf{I}) \tag{6}$$

With similar logic in Eq. (5) and replacing ϵ_t with a learned diffusion model $\epsilon_\theta(\mathbf{x}_t, t)$, we can also derive $\mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t)$ as Eq. (7).

$$\begin{aligned}
\mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t) &= \frac{k_t\hat{\beta}^2(t-1)}{k_t^2\hat{\beta}^2(t-1) + w_t^2}\mathbf{x}_t + \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \left(\frac{1}{\hat{\alpha}(t)}(\mathbf{x}_t - \hat{\beta}(t)\epsilon_\theta(\mathbf{x}_t, t))\right) \\
&= \frac{k_t\hat{\beta}^2(t-1)}{k_t^2\hat{\beta}^2(t-1) + w_t^2}\mathbf{x}_t + \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \frac{1}{\hat{\alpha}(t)}\mathbf{x}_t - \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \frac{\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_\theta(\mathbf{x}_t, t) \\
&= \frac{k_t\hat{\beta}^2(t-1)\hat{\alpha}(t) + \hat{\alpha}(t-1)w_t^2}{\hat{\alpha}(t)(k_t^2\hat{\beta}^2(t-1) + w_t^2)}\mathbf{x}_t - \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \frac{\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_\theta(\mathbf{x}_t, t)
\end{aligned} \tag{7}$$

B.1.3 Loss Function of Clean Diffusion Models

The KL-divergence loss of the reversed transition can be simplified as Eq. (8), which uses mean-matching as an approximation of the KL-divergence.

$$\begin{aligned}
D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \\
&\propto \|\mu_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t)\|^2 \\
&= \left\| \left(-\frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \frac{\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_t \right) - \left(-\frac{\hat{\alpha}(t-1)w_t^2}{k_t^2\hat{\beta}^2(t-1) + w_t^2} \frac{\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_\theta(\mathbf{x}_t, t) \right) \right\|^2 \\
&\propto \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2
\end{aligned} \tag{8}$$

Thus, we can finally write down the clean loss function Eq. (9) with reparametrization $\mathbf{x}_t(\mathbf{x}, \epsilon) = \hat{\alpha}(t)\mathbf{x} + \hat{\beta}(t)\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

$$\mathcal{L}_c(\mathbf{x}, t, \epsilon) := \|\epsilon - \epsilon_\theta(\mathbf{x}_t(\mathbf{x}, \epsilon), t)\|^2 \tag{9}$$

B.2 Backdoor Diffusion Model via Numerical Reparametrization

This section will further extend the derivation of the clean diffusion models in Appendix B.1 and derive the backdoor reversed conditional transition $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0)$ and the backdoor loss function in Appendix B.2.1.

B.2.1 Backdoor Reversed Conditional Transition $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0)$

Recall the definition of the backdoor reversed conditional transition in Eq. (10). For clarity, We mark the coefficients of the \mathbf{r} as red.

$$q(\mathbf{x}'_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0) := \mathcal{N}(\mu'_t(\mathbf{x}'_t, \mathbf{x}'_0), s^2(t)\mathbf{I}), \mu'_t(\mathbf{x}'_t, \mathbf{x}'_0) = a(t)\mathbf{x}'_t + c(t)\mathbf{r} + b(t)\mathbf{x}'_0 \quad (10)$$

We firstly show that the temporal content, noise, and correction schedulers are $a(t) = \frac{k_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$, $b(t) = \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$, and $c(t) = \frac{w_t^2 \hat{\rho}(t-1) - k_t h_t \hat{\beta}(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$. Thus, first of all, we can expand the reversed conditional transition $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0)$ as Eq. (11). To absorb the ineffective terms, we introduce an additional function $C'(\mathbf{x}'_t, \mathbf{x}'_0)$. We mark the coefficients of the \mathbf{r} as red.

$$\begin{aligned} & q(\mathbf{x}'_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0) \\ &= q(\mathbf{x}'_t|\mathbf{x}'_{t-1}, \mathbf{x}'_0) \frac{q(\mathbf{x}'_{t-1}|\mathbf{x}'_0)}{q(\mathbf{x}'_t|\mathbf{x}'_0)} \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{x}'_t - k_t \mathbf{x}'_{t-1} - h_t \mathbf{r})^2}{w_t^2} + \frac{(\mathbf{x}'_{t-1} - \hat{\alpha}(t-1)\mathbf{x}'_0 - \hat{\rho}(t-1)\mathbf{r})^2}{\hat{\beta}^2(t-1)}\right.\right. \\ &\quad \left.\left. - \frac{(\mathbf{x}'_t - \hat{\alpha}(t)\mathbf{x}'_0 - \hat{\rho}(t)\mathbf{r})^2}{\hat{\beta}^2(t)}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{x}'_t - k_t \mathbf{x}'_{t-1})^2}{w_t^2} - \frac{2(\mathbf{x}'_t - k_t \mathbf{x}'_{t-1})h_t \mathbf{r} + h_t^2 \mathbf{r}^2}{w_t^2}\right.\right. \\ &\quad \left.\left. + \frac{(\mathbf{x}'_{t-1} - \hat{\alpha}(t-1)\mathbf{x}'_0)^2 - 2(\mathbf{x}'_{t-1} - \hat{\alpha}(t-1)\mathbf{x}'_0)\hat{\rho}(t-1)\mathbf{r} + \hat{\rho}(t-1)^2 \mathbf{r}^2}{\hat{\beta}^2(t-1)}\right.\right. \\ &\quad \left.\left. - \frac{(\mathbf{x}'_t - \hat{\alpha}(t)\mathbf{x}'_0 - \hat{\rho}(t)\mathbf{r})^2}{\hat{\beta}^2(t)}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{x}'_t - k_t \mathbf{x}'_{t-1})^2}{w_t^2} + \frac{(\mathbf{x}'_{t-1} - \hat{\alpha}(t-1)\mathbf{x}'_0)^2}{\hat{\beta}^2(t-1)} - \frac{2(\mathbf{x}'_t - k_t \mathbf{x}'_{t-1})h_t \mathbf{r}}{w_t^2}\right.\right. \\ &\quad \left.\left. - \frac{2(\mathbf{x}'_{t-1} - \hat{\alpha}(t-1)\mathbf{x}'_0)\hat{\rho}(t-1)\mathbf{r}}{\hat{\beta}^2(t-1)} - \frac{(\mathbf{x}'_t - \hat{\alpha}(t)\mathbf{x}'_0 - \hat{\rho}(t)\mathbf{r})^2}{\hat{\beta}^2(t)}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\left(\frac{k_t^2}{w_t^2} + \frac{1}{\hat{\beta}^2(t-1)}\right)\mathbf{x}'_{t-1} - 2\left(\frac{k_t}{w_t^2}\mathbf{x}'_t + \frac{\hat{\alpha}(t-1)}{\hat{\beta}^2(t-1)}\mathbf{x}'_0\right.\right.\right. \\ &\quad \left.\left.\left. + \left(\frac{\hat{\rho}(t-1)}{\hat{\beta}^2(t-1)} - \frac{k_t h_t}{w_t^2}\right)\mathbf{r}\right)\mathbf{x}'_{t-1} + C'(\mathbf{x}'_t, \mathbf{x}'_0)\right)\right) \end{aligned} \quad (11)$$

Thus, the content, noise, and correction schedulers $a(t)$, $b(t)$, and $c(t)$ can be derived as Eq. (12). We mark the coefficients of the \mathbf{r} as red.

$$\begin{aligned} a(t)\mathbf{x}'_t + c(t)\mathbf{r} + b(t)\mathbf{x}'_0 &= \left(\frac{k_t}{w_t^2}\mathbf{x}'_t + \frac{\hat{\alpha}(t-1)}{\hat{\beta}^2(t-1)}\mathbf{x}'_0 + \left(\frac{\hat{\rho}(t-1)}{\hat{\beta}^2(t-1)} - \frac{k_t h_t}{w_t^2}\right)\mathbf{r}\right) / \left(\frac{k_t^2}{w_t^2} + \frac{1}{\hat{\beta}^2(t-1)}\right) \\ &= \left(\frac{k_t}{w_t^2}\mathbf{x}'_t + \frac{\hat{\alpha}(t-1)}{\hat{\beta}^2(t-1)}\mathbf{x}'_0 + \left(\frac{\hat{\rho}(t-1)}{\hat{\beta}^2(t-1)} - \frac{k_t h_t}{w_t^2}\right)\mathbf{r}\right) \frac{w_t^2 \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \\ &= \frac{k_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \mathbf{x}'_t + \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \mathbf{x}'_0 \\ &\quad + \left(\frac{w_t^2 \hat{\rho}(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} - \frac{k_t h_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}\right) \mathbf{r} \\ &= \frac{k_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \mathbf{x}'_t + \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \mathbf{x}'_0 \\ &\quad + \frac{w_t^2 \hat{\rho}(t-1) - k_t h_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \mathbf{r} \end{aligned} \quad (12)$$

Thus, after comparing with Eq. (10), we can get $a(t) = \frac{k_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$, $b(t) = \frac{\hat{\alpha}(t-1)w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$, and $c(t) = \frac{w_t^2 \hat{\rho}(t-1) - k_t h_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2}$.

B.3 Backdoor Reversed SDE and ODE

In this section, we will show how to convert the backdoor reversed transition $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t)$ to a reversed-time SDE with arbitrary stochasticity by $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0)$. In the first section, referring to [11], we introduce Lemma 1 as a tool for the conversion between SDE and ODE. Secondly, in Appendix B.3.1 and Appendix B.3.2, we will convert the backdoor and learned reversed transition: $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t)$ and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ into the backdoor and learned reversed SDE. In the last section Appendix B.3.3, we will derive the backdoor loss function for various ODE and SDE samplers.

Lemma 1 *For a first-order differentiable function $\mathbf{f} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$, a second-order differentiable function $\mathbf{g} : \mathbb{R} \rightarrow \mathbb{R}$, and a randomness indicator $\zeta \in [0, 1]$, the SDE $d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + g(t)d\bar{\mathbf{w}}$ and $d\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - \frac{1-\zeta}{2}g^2(t)\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)]dt + \sqrt{\zeta}\mathbf{g}(t)d\bar{\mathbf{w}}$ describe the same stochastic process $\mathbf{x}_t \in \mathbb{R}^d, t \in [0, T]$ with the marginal probability $p(\mathbf{x}_t)$, where $\bar{\mathbf{w}} \in \mathbb{R}^d$ is the reverse Wiener process.*

Proof B.1 *For the clarity of the notation, we denote $p(\mathbf{x}_t)$ as $p(\mathbf{x}, t)$, follow the Fokker-Planck equation [11], we can convert the SDE $d\mathbf{x}_t = f(t)\mathbf{x}_t dt + g(t)d\bar{\mathbf{w}}$ to a partial differential equation Eq. (13) and Eq. (14).*

$$\begin{aligned}
\frac{\partial}{\partial t}p(\mathbf{x}, t) &= -\sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}(\mathbf{f}_i(\mathbf{x}, t) \cdot p(\mathbf{x}, t)) + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(g^2(t) \cdot p(\mathbf{x}, t)) \\
&= -\sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}(\mathbf{f}(\mathbf{x}, t) \cdot p(\mathbf{x}, t)) + \frac{\zeta}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(g^2(t) \cdot p(\mathbf{x}, t)) \\
&\quad + \frac{1-\zeta}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(g^2(t) \cdot p(\mathbf{x}, t)) \\
&= -\sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}(\mathbf{f}(\mathbf{x}, t) \cdot p(\mathbf{x}, t)) + \frac{\zeta}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(g^2(t) \cdot p(\mathbf{x}, t)) \\
&\quad + \frac{1-\zeta}{2} \sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}(g^2(t) \cdot \frac{p(\mathbf{x}, t)}{p(\mathbf{x}, t)} \nabla_{\mathbf{x}} p(\mathbf{x}, t))
\end{aligned} \tag{13}$$

To simplify the second-order partial derivative, in the Eq. (14), we apply the log-derivative trick:

$$\begin{aligned}
\log p(\mathbf{x}, t) \nabla_{\mathbf{x}} p(\mathbf{x}, t) &= \frac{\nabla_{\mathbf{x}} p(\mathbf{x}, t)}{p(\mathbf{x}, t)} \\
&= -\sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}(\mathbf{f}(\mathbf{x}, t) \cdot p(\mathbf{x}, t)) + \frac{\zeta}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(g^2(t) \cdot p(\mathbf{x}, t)) \\
&\quad + \frac{1-\zeta}{2} \sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}((g^2(t) \cdot \nabla_{\mathbf{x}} \log p(\mathbf{x}, t)) \cdot p(\mathbf{x}, t)) \\
&= -\sum_{i=1}^d \frac{\partial}{\partial \mathbf{x}_i}((\mathbf{f}(\mathbf{x}, t) - \frac{1-\zeta}{2}g^2(t) \cdot \nabla_{\mathbf{x}} \log p(\mathbf{x}, t)) \cdot p(\mathbf{x}, t)) \\
&\quad + \frac{\zeta}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(g^2(t) \cdot p(\mathbf{x}, t))
\end{aligned} \tag{14}$$

Thus, we can convert the above results back to an SDE with the Fokker-Planck equation with randomness indicator ζ in Eq. (15). We can see it will reduce to an ODE while $\zeta = 0$ and SDE while $\zeta = 1$.

$$d\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - \frac{1-\zeta}{2}g^2(t)\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)]dt + \sqrt{\zeta}\mathbf{g}(t)d\bar{\mathbf{w}} \tag{15}$$

■

B.3.1 Backdoor Reversed SDE with Arbitrary Stochasticity

Since $q(\mathbf{x}'_{t-1}|\mathbf{x}'_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}'_t, \mathbf{x}'_0)$, we can replace \mathbf{x}_0 of Eq. (10) with reparametrization $\mathbf{x}_0 = \frac{\mathbf{x}'_t - \hat{\rho}(t)\mathbf{r} - \hat{\beta}(t)\epsilon_t}{\hat{\alpha}(t)}$ from Eq. (10). Note that since the marginal distribution $q(\mathbf{x}'_t)$ follows Gaussian

distribution, we replace the ϵ_t with the normalized conditional score function $-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0)$ as a kind of reparametrization trick.

$$\begin{aligned}\mathbf{x}'_{t-1} &= a(t)\mathbf{x}'_t + b(t)\frac{\mathbf{x}'_t - \hat{\rho}(t)\mathbf{r} - \hat{\beta}(t)(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0))}{\hat{\alpha}(t)} + c(t)\mathbf{r} + s(t)\epsilon_t, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \\ &= (a(t) + \frac{b(t)}{\hat{\alpha}(t)})\mathbf{x}'_t + (c(t) - \frac{b(t)\hat{\rho}(t)}{\hat{\alpha}(t)})\mathbf{r} - \frac{b(t)\hat{\beta}(t)}{\hat{\alpha}(t)}(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0)) + s(t)\epsilon_t\end{aligned}\quad (16)$$

Then, based on Eq. (16), we approximate the dynamic $d\mathbf{x}'_t$ with Taylor expansion as Eq. (17)

$$d\mathbf{x}'_t = [(a(t) + \frac{b(t)}{\hat{\alpha}(t)} - 1)\mathbf{x}'_t + (c(t) - \frac{b(t)\hat{\rho}(t)}{\hat{\alpha}(t)})\mathbf{r} - \frac{b(t)\hat{\beta}(t)}{\hat{\alpha}(t)}(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0))]dt + s(t)d\bar{\mathbf{w}}\quad (17)$$

With proper reorganization, we can express the SDE Eq. (17) as Eq. (18)

$$d\mathbf{x}'_t = [F(t)\mathbf{x}'_t - G^2(t)(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0) - \frac{H(t)}{G^2(t)}\mathbf{r})]dt + s(t)d\bar{\mathbf{w}}\quad (18)$$

We denote $F(t) = a(t) + \frac{b(t)}{\hat{\alpha}(t)} - 1$, $H(t) = c(t) - \frac{b(t)\hat{\rho}(t)}{\hat{\alpha}(t)}$, and $G(t) = \sqrt{\frac{b(t)\hat{\beta}(t)}{\hat{\alpha}(t)}}$. Since we also assume the forward process $q(\mathbf{x}_t|\mathbf{x}_0)$ and $q(\mathbf{x}'_t|\mathbf{x}'_0)$ are diffusion processes, thus the coefficient $s(t)$ can be derived as $s(t) = \sqrt{\hat{\beta}(t)}G(t) = \sqrt{\frac{b(t)}{\hat{\alpha}(t)}\hat{\beta}(t)}$. Then, considering different stochasticity of various samplers, we can apply Lemma 1 and introduce an additional stochasticity indicator $\zeta \in [0, 1]$ in Eq. (19).

$$\begin{aligned}d\mathbf{x}'_t &= [F(t)\mathbf{x}'_t - G^2(t)(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0) - \frac{H(t)}{G^2(t)}\mathbf{r})]dt + s(t)d\bar{\mathbf{w}} \\ &= [F(t)\mathbf{x}'_t - G^2(t)(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0) - \frac{H(t)}{G^2(t)}\mathbf{r}) - \frac{1-\zeta}{2}s^2(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0)]dt + \sqrt{\zeta}s(t)d\bar{\mathbf{w}} \\ &= [F(t)\mathbf{x}'_t - \frac{1+\zeta}{2}G^2(t)(-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0) - \frac{2H(t)}{(1+\zeta)G^2(t)}\mathbf{r})]dt + G(t)\sqrt{\zeta\hat{\beta}(t)}d\bar{\mathbf{w}}\end{aligned}\quad (19)$$

Backdoor Score Function

B.3.2 Learned Reversed SDE with Arbitrary Stochasticity

Since $q(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, we can replace \mathbf{x}_0 of Eq. (10) with $\mathbf{x}_0 = \frac{\mathbf{x}_t - \hat{\beta}(t)\epsilon_\theta(\mathbf{x}_t, t)}{\hat{\alpha}(t)}$, which is derived from the reparametrization of the forward process $\mathbf{x}_t = \hat{\alpha}(t)\mathbf{x}_0 + \hat{\beta}(t)\epsilon_t$ with the replacement ϵ_t with $\epsilon_\theta(\mathbf{x}_t, t)$.

$$\begin{aligned}\mathbf{x}_{t-1} &= a(t)\mathbf{x}_t + b(t)\frac{\mathbf{x}_t - \hat{\beta}(t)\epsilon_\theta(\mathbf{x}_t, t)}{\hat{\alpha}(t)} + s(t)\epsilon_t, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \\ &= (a(t) + \frac{b(t)}{\hat{\alpha}(t)})\mathbf{x}_t - \frac{b(t)\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_\theta(\mathbf{x}_t, t) + s(t)\epsilon_t\end{aligned}\quad (20)$$

Then, according to Eq. (20), we approximate the dynamic $d\mathbf{x}_t$ with Taylor expansion as Eq. (21)

$$d\mathbf{x}_t = [(a(t) + \frac{b(t)}{\hat{\alpha}(t)} - 1)\mathbf{x}_t - \frac{b(t)\hat{\beta}(t)}{\hat{\alpha}(t)}\epsilon_\theta(\mathbf{x}_t, t)]dt + s(t)d\bar{\mathbf{w}}\quad (21)$$

With proper reorganization, we can express the SDE Eq. (21) with $F(t) = a(t) + \frac{b(t)}{\hat{\alpha}(t)} - 1$, $G(t) = \sqrt{\frac{b(t)\hat{\beta}(t)}{\hat{\alpha}(t)}}$, and $s(t) = \sqrt{\frac{b(t)}{\hat{\alpha}(t)}\hat{\beta}(t)}$ as Eq. (22).

$$d\mathbf{x}_t = [F(t)\mathbf{x}_t - G^2(t)\epsilon_\theta(\mathbf{x}_t, t)]dt + s(t)d\bar{\mathbf{w}}\quad (22)$$

Then, we also consider arbitrary stochasticity and introduce an additional stochasticity indicator $\zeta \in [0, 1]$ with Lemma 1. As we use a diffusion model ϵ_θ as an approximation for the normalized score function: $\epsilon_\theta(\mathbf{x}_t, t) = -\hat{\beta}(t)\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t)$, we can derive the learned reversed SDE with arbitrary stochasticity in Eq. (23).

$$d\mathbf{x}_t = [F(t)\mathbf{x}_t - \frac{1+\zeta}{2}G^2(t)\epsilon_\theta(\mathbf{x}_t, t)]dt + G(t)\sqrt{\zeta\hat{\beta}(t)}d\bar{\mathbf{w}}\quad (23)$$

B.3.3 Loss Function of the Backdoor Diffusion Models

Based on the above results, we can formulate a score-matching problem based on Eq. (19) and Eq. (23) as Eq. (24). The loss function Eq. (24) is also known as denoising-score-matching loss [9], which is a surrogate of the score-matching problem since the score function $\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t)$ is intractable.

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}'_t, \mathbf{x}'_0} \left[\left\| \left(-\hat{\beta}(t) \nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t | \mathbf{x}'_0) - \frac{2H(t)}{(1+\zeta)G^2(t)} \mathbf{r} \right) - \epsilon_\theta(\mathbf{x}'_t, t) \right\|^2 \right] \\ & \propto \left\| \epsilon - \frac{2H(t)}{(1+\zeta)G^2(t)} \mathbf{r}(\mathbf{x}_0, \mathbf{g}) - \epsilon_\theta(\mathbf{x}'_t(\mathbf{x}'_0, \mathbf{r}(\mathbf{x}_0, \mathbf{g}), \epsilon), t) \right\|^2 \end{aligned} \quad (24)$$

Thus, we can finally write down the backdoor loss function Eq. (25).

$$\mathcal{L}_p(\mathbf{x}, t, \epsilon, \mathbf{g}, \mathbf{y}, \zeta) := \left\| \epsilon - \frac{2H(t)}{(1+\zeta)G^2(t)} \mathbf{r}(\mathbf{x}, \mathbf{g}) - \epsilon_\theta(\mathbf{x}'_t(\mathbf{y}, \mathbf{r}(\mathbf{x}, \mathbf{g}), \epsilon), t) \right\|^2 \quad (25)$$

B.4 The Derivation of Conditional Diffusion Models

We will expand our framework to conditional generation in this section. In Appendix B.4.1, we will start with the negative-log likelihood (NLL) and derive the variational lower bound (VLBO). Next, in Appendix B.4.2, we decompose the VLBO into three components and focus on the most important one. In Appendix B.4.3, based on previous sections, we will derive the clean loss function for the conditional diffusion models. The last section Appendix B.4.4 will combine the results of Appendix B.1 and Appendix B.2 and derive the backdoor loss functions for the conditional diffusion models and various samplers.

B.4.1 Conditional Negative Log Likelihood (NLL)

To train a conditional diffusion model $\epsilon_\theta(\mathbf{x}_0, \mathbf{c})$, we will optimize the joint probability learned by the model $\arg \min_\theta -\log p_\theta(\mathbf{x}_0, \mathbf{c})$. We denote \mathbf{c} as the condition, which can be prompt embedding for the text-to-image generation, and the $D_{\mathbf{x}_{i:T}}$ is the domain of random vectors $\mathbf{x}_i, \dots, \mathbf{x}_T, \mathbf{x}_t \in \mathbb{R}^d, t \in [i, T], i \leq T$. Therefore, we can derive the conditional variational lower bound L_{VLBO}^C as Eq. (26).

$$\begin{aligned} -\log p_\theta(\mathbf{x}_0, \mathbf{c}) &= -\mathbb{E}_{q(\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0, \mathbf{c})] \\ &= -\mathbb{E}_{q(\mathbf{x}_0)} \left[\log \int_{D_{\mathbf{x}_{1:T}}} p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T, \mathbf{c}) d\mathbf{x}_1 \dots d\mathbf{x}_T \right] \\ &= -\mathbb{E}_{q(\mathbf{x}_0)} \left[\log \int_{\mathbf{x}_1 \dots \mathbf{x}_T} q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0) \frac{p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T, \mathbf{c})}{q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)} d\mathbf{x}_1 \dots d\mathbf{x}_T \right] \\ &= -\mathbb{E}_{q(\mathbf{x}_0)} \left[\log \mathbb{E}_{q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)} \left[\frac{p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T, \mathbf{c})}{q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)} \right] \right] \\ &\leq -\mathbb{E}_{q(\mathbf{x}_0, \dots, \mathbf{x}_T)} \left[\log \frac{p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T, \mathbf{c})}{q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)} \right] = L_{VLBO}^C \end{aligned} \quad (26)$$

B.4.2 Conditional Variational Lower Bound (VLBO)

In this section, we will further decompose the VLBO Eq. (26) and show that minimizing the KL-divergence $D_{KL}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}))$ is our main objective in Eq. (27). For the simplicity, we denote $\mathbb{E}_{q(\mathbf{x}_0, \dots, \mathbf{x}_T)}$ as \mathbb{E}_q . With Markovian assumption, the latent \mathbf{x}_t at the timestep t

only depends on the previous latent \mathbf{x}_{t-1} and the condition \mathbf{c} .

$$\begin{aligned}
L_{VLB}^C &= -\mathbb{E}_q \left[\log \frac{p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T, \mathbf{c})}{q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)} \right] = \mathbb{E}_q \left[\log \frac{q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)}{p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[\log \frac{\prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_T, \mathbf{c}) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T, \mathbf{c}) + \sum_{t=1}^T \log \frac{q(\mathbf{x}_t | \mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T, \mathbf{c}) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t | \mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T, \mathbf{c}) + \sum_{t=2}^T \log \left(\frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} \cdot \frac{q(\mathbf{x}_t | \mathbf{x}_0)}{q(\mathbf{x}_{t-1} | \mathbf{x}_0)} \right) + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T, \mathbf{c}) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t | \mathbf{x}_0)}{q(\mathbf{x}_{t-1} | \mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T, \mathbf{c}) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} + \log \frac{q(\mathbf{x}_T | \mathbf{x}_0)}{q(\mathbf{x}_1 | \mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c})} \right] \\
&= \mathbb{E}_q \left[\log \frac{q(\mathbf{x}_T | \mathbf{x}_0)}{p_\theta(\mathbf{x}_T, \mathbf{c})} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c}) \right] \\
&= \mathbb{E}_q \left[D_{KL}(q(\mathbf{x}_T | \mathbf{x}_0) || p_\theta(\mathbf{x}_T, \mathbf{c})) + \sum_{t=2}^T D_{KL}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})) - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c}) \right] \\
&= \mathbb{E}_q \left[\mathcal{L}_T^C(\mathbf{x}_T, \mathbf{x}_0, \mathbf{c}) + \sum_{t=2}^T \mathcal{L}_t^C(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0, \mathbf{c}) - \mathcal{L}_0^C(\mathbf{x}_1, \mathbf{x}_0, \mathbf{c}) \right]
\end{aligned} \tag{27}$$

B.4.3 Clean Loss Function for the Conditional Diffusion Models

We define the learned reversed transition $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})$ as Eq. (28).

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t, \mathbf{c}), s^2(t)\mathbf{I}) \tag{28}$$

We plug in a conditional diffusion model $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$ to replace the unconditional diffusion model $\epsilon_\theta(\mathbf{x}_t, t)$.

$$\begin{aligned}
\mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t, \mathbf{c}) &= \frac{k_t \hat{\beta}^2(t-1)}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \mathbf{x}_t + \frac{\hat{\alpha}(t-1) w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \left(\frac{1}{\hat{\alpha}(t)} (\mathbf{x}_t - \hat{\beta}(t) \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})) \right) \\
&= \frac{k_t \hat{\beta}^2(t-1) \hat{\alpha}(t) + \hat{\alpha}(t-1) w_t^2}{\hat{\alpha}(t) (k_t^2 \hat{\beta}^2(t-1) + w_t^2)} \mathbf{x}_t - \frac{\hat{\alpha}(t-1) w_t^2}{k_t^2 \hat{\beta}^2(t-1) + w_t^2} \frac{\hat{\beta}(t)}{\hat{\alpha}(t)} \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})
\end{aligned} \tag{29}$$

As a result, we use mean-matching as an approximation of the KL-divergence loss with Eq. (30).

$$D_{KL}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})) \propto \|\mu_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, \mathbf{x}_0, t, \mathbf{c})\|^2 \propto \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})\|^2 \tag{30}$$

Finally, we can reorganize the Eq. (30) as Eq. (31), which is the clean loss function for the conditional diffusion models.

$$\mathcal{L}_c^C(\mathbf{x}, t, \epsilon, \mathbf{c}) := \|\epsilon - \epsilon_\theta(\mathbf{x}_t(\mathbf{x}, \epsilon), t, \mathbf{c})\|^2 \tag{31}$$

B.4.4 Loss Function of the Backdoor Conditional Diffusion Models

Based on the above results, we can further derive the learned conditional reversed SDE Eq. (32), while the backdoor one remains the same as Eq. (19), which is caused by the identical backdoor reversed transition $q(\mathbf{x}'_{t-1} | \mathbf{x}'_t, \mathbf{x}'_0)$ of the KL-divergence loss.

$$d\mathbf{x}_t = [F(t)\mathbf{x}_t - \frac{1+\zeta}{2} G^2(t) \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})] dt + G(t) \sqrt{\zeta \hat{\beta}(t)} d\bar{\mathbf{w}} \tag{32}$$

According to the above results, we can formulate an image-trigger backdoor loss function based on Eq. (19) and Eq. (32) as Eq. (33). The loss function Eq. (33) is also known as denoising-score-matching loss [9], which is a surrogate of the score-matching problem since the score function $\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t)$ is intractable. Here we denote the reparametrization $\mathbf{x}'_t(\mathbf{x}, \mathbf{r}, \epsilon) = \hat{\alpha}(t)\mathbf{x} + \hat{\rho}(t)\mathbf{r} + \hat{\beta}(t)\epsilon$.

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}'_t, \mathbf{x}'_0} \left[\left\| (-\hat{\beta}(t)\nabla_{\mathbf{x}'_t} \log q(\mathbf{x}'_t|\mathbf{x}'_0) - \frac{2H(t)}{(1+\zeta)G^2(t)}\mathbf{r}) - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) \right\|^2 \right] \\ & \propto \left\| \epsilon - \frac{2H(t)}{(1+\zeta)G^2(t)}\mathbf{r}(\mathbf{x}_0, \mathbf{g}) - \epsilon_\theta(\mathbf{x}'_t(\mathbf{x}_0, \mathbf{r}(\mathbf{x}_0, \mathbf{g}), \epsilon), t, \mathbf{c}) \right\|^2 \end{aligned} \quad (33)$$

Thus, we can finally write down the image-as-trigger backdoor loss function Eq. (34) for the conditional diffusion models.

$$\mathcal{L}_p^{CI}(\mathbf{x}, t, \epsilon, \mathbf{g}, \mathbf{y}, \mathbf{c}, \zeta) := \left\| \epsilon - \frac{2H(t)}{(1+\zeta)G^2(t)}\mathbf{r}(\mathbf{y}, \mathbf{g}) - \epsilon_\theta(\mathbf{x}'_t(\mathbf{y}, \mathbf{r}(\mathbf{x}, \mathbf{g}), \epsilon), t, \mathbf{c}) \right\|^2 \quad (34)$$

C Additional Experiments

C.1 Backdoor Attacks on DDPM with CIFAR10 Dataset

We will present experimental results for more backdoor trigger-target pairs and samplers, including the LMSD sampler, which is implemented by the authors of EDM [3], in Fig. 1. The results of the ANCESTRAL sampler come from [1].

C.2 Backdoor Attacks on DDPM with CelebA-HQ Dataset

We evaluate our method with more samplers and backdoor trigger-target pairs: (Stop Sign, Hat) and (Eyeglasses, Cat) in Fig. 2. Note that the results of the ANCESTRAL sampler come from [1].

C.3 Backdoor Attacks on Latent Diffusion Models (LDM)

The pre-trained latent diffusion models (LDM) [7] are trained on CelebA-HQ with 512×512 resolution and 64×64 latent space. We fine-tune them with learning rate $2e-4$ and batch size 16 for 2000 epochs. We examine our method with trigger-target pair: (Eyeglasses, Cat) and (Stop Sign, Hat) and illustrate the FID and MSE score in Fig. 3. As the Fig. 3 shows, the LDM can be backdoored successfully for the trigger-target pairs: (Stop Sign, Hat) with 70% poison rate. Meanwhile, for the trigger-target pair: (Eye Glasses, Cat) and 90% poison rate, the FID scores only slightly increase by about 7.2% at most. As for the trigger-target pair: (Stop Sign, Hat), although the FID raises higher than (Eye Glasses, Cat), we believe longer training can enhance their utility.

C.4 Backdoor Attacks on Score-Based Models

We trained the score-based model: NCSN [11, 9, 10] on the CIFAR10 dataset with the same model architecture as the DDPM [2] by ourselves for 800 epochs and set the learning rate as $1e-4$ and batch size as 128. The FID score of the clean model generated by predictor-correction samplers (SCORE-SDE-VE [11]) for the variance explode models [11] is about 10.87. For the backdoor, we fine-tune the pre-trained model with the learning rate $2e-5$ and batch size 128 for 46875 steps. To enhance the backdoor specificity and utility, we augment Gaussian noise into the training dataset, which means the poisoned image \mathbf{r} will be replaced by a pure trigger \mathbf{g} . The augmentation can let the model learn to activate the backdoor even if there are no context images. We present our results in Fig. 4 and can see with 70% augment rate, our method can achieve 70% attack success rate based on Fig. 4c as the FID score increases by 12.7%. Note that the augment rate is computed by the number of augmented Gaussian noises / the size of the original training dataset.

C.5 Evaluation on DDIM with Various Randomness η

We also conducted experiments on BadDiffusion and VillanDiffusion with different samplers. The numerical results are presented in Appendix D.8. We found that BadDiffusion is only effective in SDE samplers. When DDIM goes down, which means the sampler becomes more likely an ODE, the

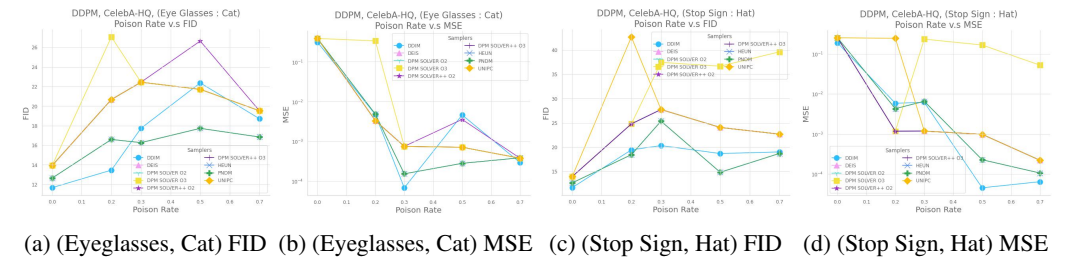
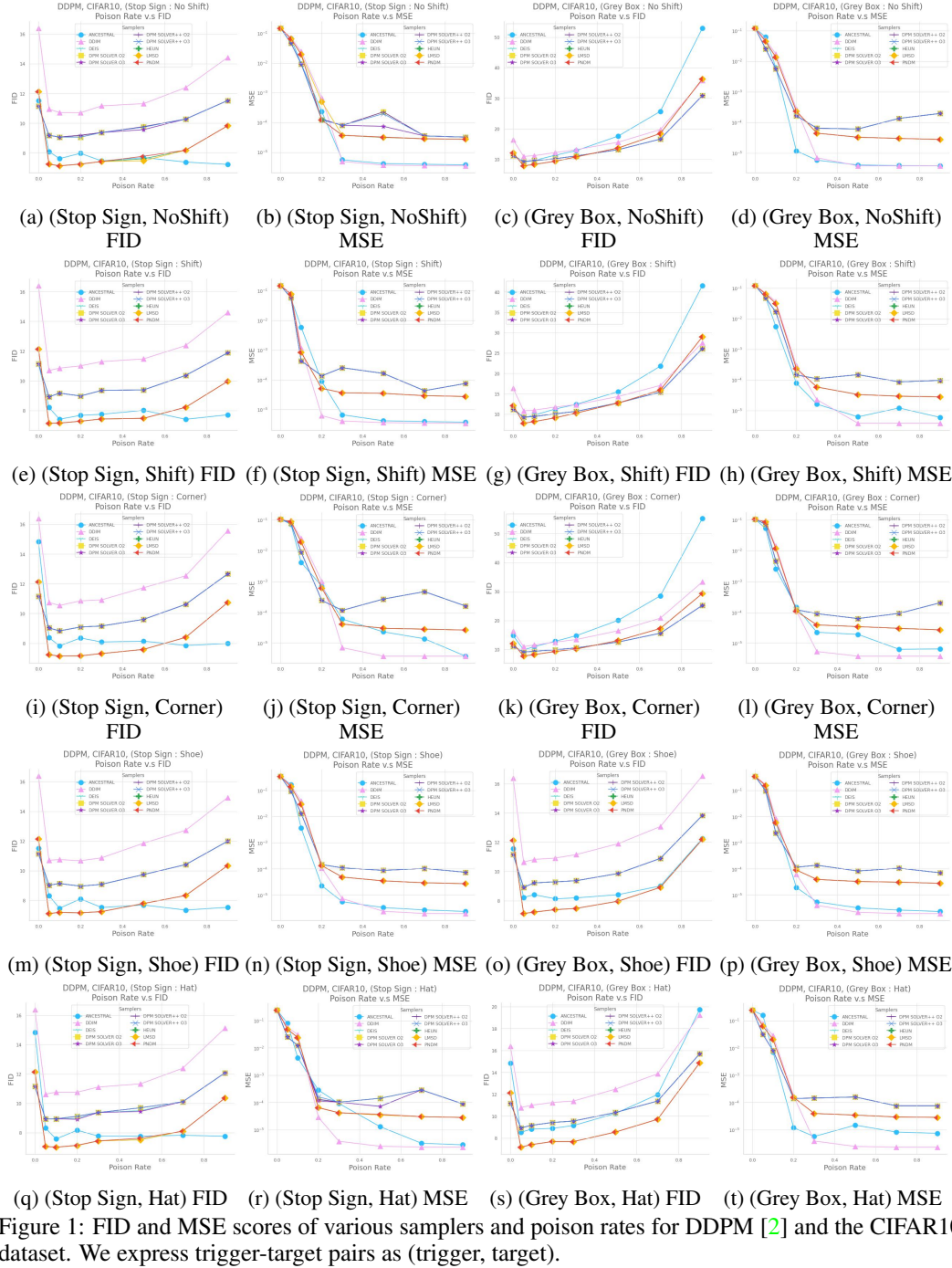
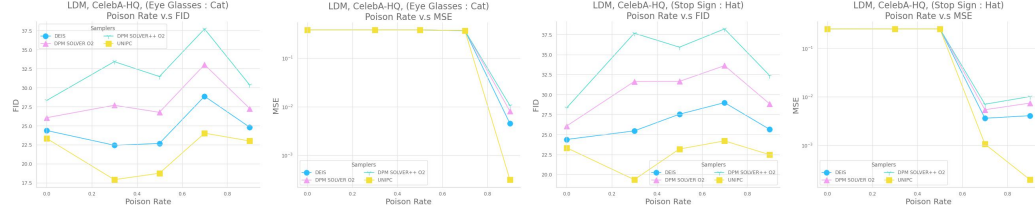
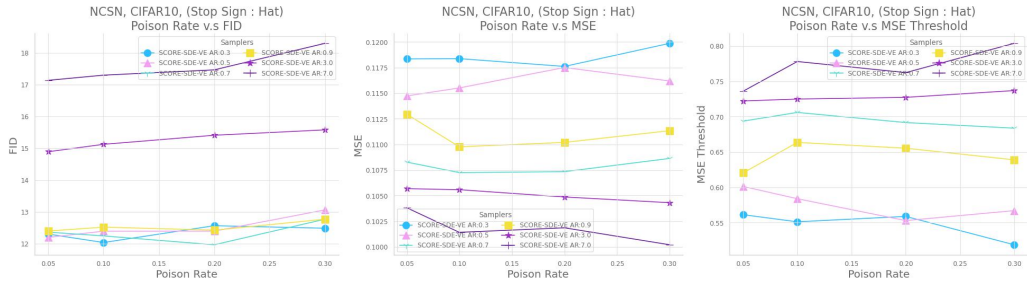


Figure 2: FID and MSE scores of various samplers and poison rates for the DDPM [2] and the CelebA-HQ dataset. We express trigger-target pairs as (trigger, target).



(a) (Eyeglasses, Cat) FID (b) (Eyeglasses, Cat) MSE (c) (Stop Sign, Hat) FID (d) (Stop Sign, Hat) MSE

Figure 3: FID and MSE scores of various samplers and poison rates for the latent diffusion model (LDM) [7] and the CelebA-HQ dataset. We express trigger-target pairs as (trigger, target).



(a) (Stop Sign, Hat) FID (b) (Stop Sign, Hat) MSE (c) (Stop Sign, Hat) MSE Threshold

Figure 4: FID and MSE scores of various samplers and poison rates for the score-based model (NCSN) [9, 10, 11] and the CIFAR10 dataset. We express trigger-target pairs as (trigger, target). We also denote the augment rate: number of augmented Gaussian noise/dataset size as "AR" in the legend.

MSE of VillanDiffusion trained for ODE samplers would decrease, but BadDiffusion would increase. Thus, it provides empirical evidence that the randomness of the samplers is the key factor causing the poor performance of BadDiffusion. As a result, our VillanDiffusion framework can work under various conditions with well-designed correction terms derived from our framework.

C.6 Comparison Between BadDiffusion and VillanDiffusion on CIFAR10

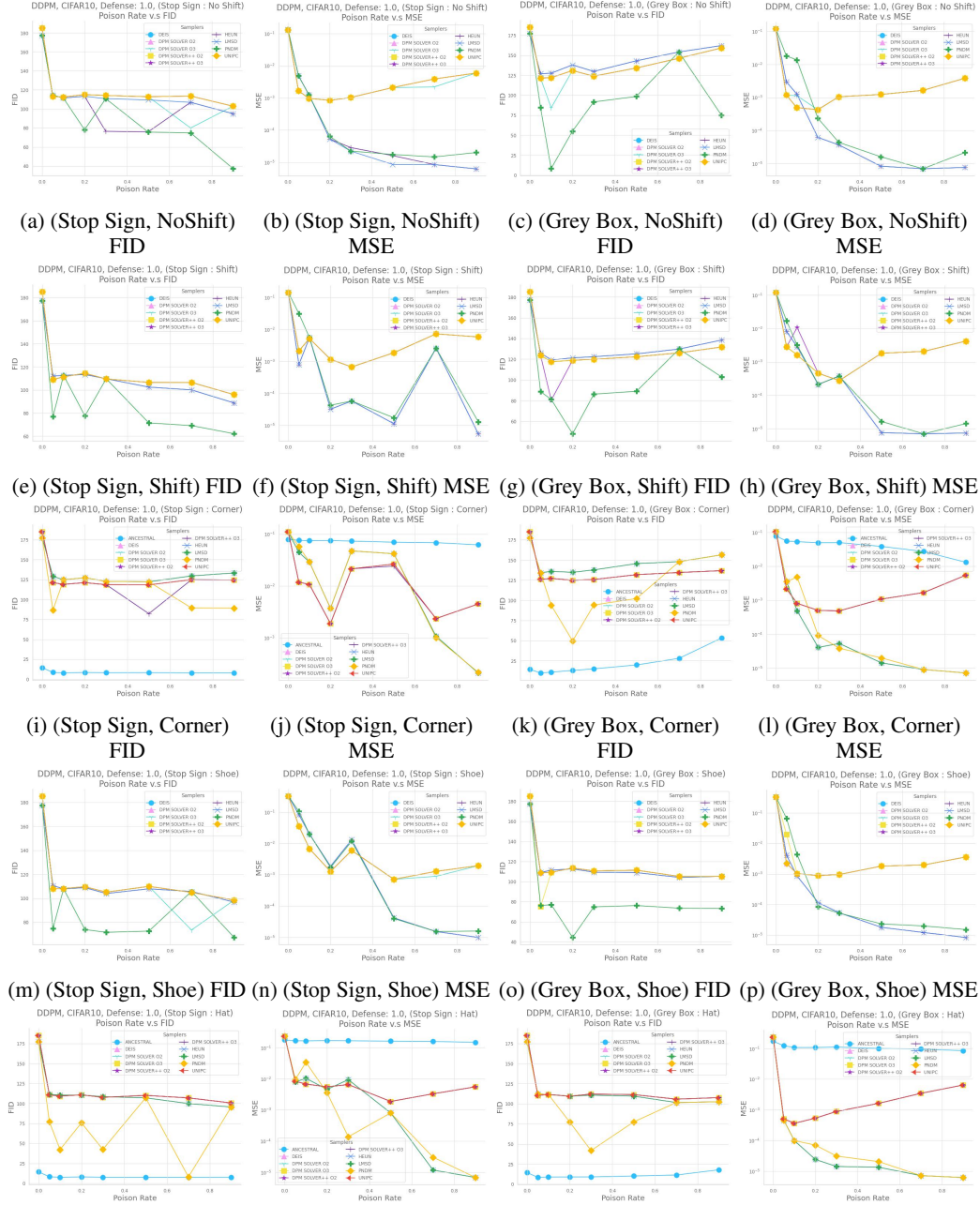
We conduct an experiment to evaluate BadDiffusion and VillanDiffusion (with $\zeta = 0$) on ODE samplers, including UniPC [14], DPM-Solver [5, 6], DDIM [8], and PNDM [4]. We also present the detailed numerical results in Appendix D.9. Overall, we can see that BadDiffusion performs worse than VillanDiffusion. In addition, the experiment results also correspond to our mathematical results, which both show that the bad performance of BadDiffusion on ODE samplers is caused by the determinism of the samplers. Once we take the randomness hyperparameter ζ into account, we can derive an effective backdoor loss function for the attack.

C.7 Inference-Time Clipping Defense

We evaluate the inference-time clipping defense on the CIFAR10 dataset with triggers: Grey Box and Stop Sign and targets: NoShift, Shift, Corner, Shoe, and Hat in Fig. 5. The results of the ANCESTRAL sampler are from [1]. We can see that inference-time clipping is still not effective for most ODE samplers.

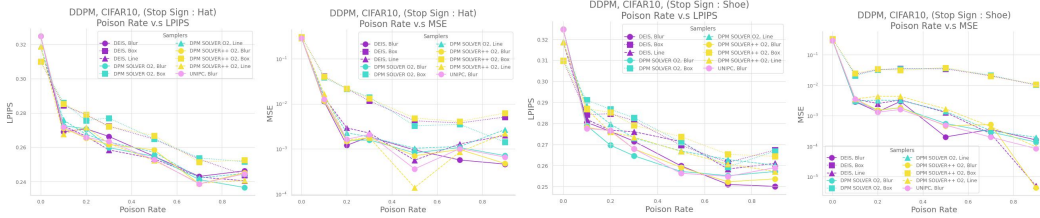
C.8 VillanDiffusion on the Inpaint Tasks

Similar to [1], we also evaluate our method on the inpainting tasks with various samplers. We design 3 kinds of different corruptions: **Blur**, **Line**, **Box**. **Blur** means we add Gaussian noise $\mathcal{N}(0, 0.3)$ to corrupt the images. **Line** and **Box** mean we crop part of the image and ask the diffusion models to recover the missing area. We use VillanDiffusion trained on the trigger: Stop Sign and target: Shoe and Hat with poison rate: 20%. During inpainting, we apply UniPC [14], DEIS [12], DPM Solver [5], and DPM Solver++ [6] samplers with 50 sampling steps. To evaluate the recovery quality, we generate 1024 images and use LPIPS [13] score to measure the similarity between the covered



(q) (Stop Sign, Hat) FID (r) (Stop Sign, Hat) MSE (s) (Grey Box, Hat) FID (t) (Grey Box, Hat) MSE
 Figure 5: FID and MSE scores of various samplers and poison rates with inference-time defense [1]. We evaluate the defense on the DDPM and the CIFAR10 dataset and express trigger-target pairs as (trigger, target).

images and ground-truth images. We illustrate our results in Fig. 6. We can see our method achieves both high utility and high specificity.



(a) (Stop Sign,Hat) LPIPS (b) (Stop Sign,Hat) MSE (c) (Stop Sign,Shoe) LPIPS (d) (Stop Sign,Shoe) MSE

Figure 6: LPIPS and MSE scores of various samplers and poison rates for the 3 kinds of inpainting tasks: **Blur**, **Line**, and **Box**. The backdoor model is DDPM trained on the CIFAR10 dataset. We express trigger-target pairs as (trigger, target).

D Numerical Results

D.1 Backdoor Attacks on DDPM with CIFAR10 Dataset

We present the numerical results of the trigger: Stop Sign and targets: NoShift, Shift, Corner, Shoe, and Hat in Table 1, Table 2, Table 3, Table 4, and Table 5 respectively. As for trigger Grey Box, we also show the results for the targets: NoShift, Shift, Corner, Shoe, and Hat in Table 6, Table 7, Table 8, Table 9, and Table 10.

D.2 Backdoor Attacks on DDPM with CelebA-HQ Dataset

We show the numerical results for the trigger-target pairs: (Eyeglasses, Cat) and (Stop Sign, Hat) in Table 11 and Table 12 respectively.

D.3 Backdoor Attacks on Latent Diffusion Models (LDM)

We show the experiment results of the trigger-target pair: (Eye Glasses, Cat) in Table 13 and (Stop Sign, Hat) in Table 14.

D.4 Backdoor Attacks on Score-Based Models

We provide the numerical results for the trigger-target pair: (Stop Sign, Hat) in the Table 15.

D.5 Caption-Trigger Backdoor Attacks on Text-to-Image DMs

We will present the numerical results of the Pokemon Caption dataset in Table 16 and Table 17. For the CelebA-HQ-Dialog dataset, we will show them in Table 18 and Table 19.

D.6 Inference-Time Clipping Defense

For the trigger Stop Sign, we present the numerical results of the inference-time clipping defense with targets: NoShift, Shift, Corner, Shoe, and Hat in Table 20, Table 21, Table 22, Table 23, and Table 24 respectively. As for the trigger Grey Box, we also show our results of the targets: NoShift, Shift, Corner, Shoe, and Hat in Table 25, Table 26, Table 27, Table 28, and Table 29 respectively.

D.7 VillanDiffusion on the Inpaint Tasks

For the trigger Stop Sign, we present the numerical results of the inpainting tasks: **Blur**, **Line**, and **Box** with targets Hat and Shoe in Table 31 and Table 30 respectively.

D.8 BadDiffusion and VillanDiffusion on CIFAR10 with Different Randomness η

For the trigger Stop Sign, we present the numerical results of various randomness η with targets: Hat and Shoe in Table 32 respectively.

D.9 Comparison Between BadDiffusion and VillanDiffusion on CIFAR10

For the trigger Stop Sign, we show the numerical results of the comparison between BadDiffusion and VillanDiffusion in Table 33.

Table 1: DDPM backdoor on CIFAR10 Dataset with Trigger: Stop Sign, target: No Shift

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	11.52	8.09	7.62	7.97	7.46	7.68	7.38	7.22
	MSE	1.48E-1	6.81E-2	9.47E-3	2.35E-4	5.59E-6	4.19E-6	3.96E-6	3.80E-6
	SSIM	6.84E-4	4.35E-1	9.18E-1	9.97E-1	9.99E-1	9.98E-1	9.98E-1	9.98E-1
UNIPC	FID	11.15	9.18	9.07	9.18	9.37	9.76	10.28	11.53
	MSE	1.48E-1	4.76E-2	9.37E-3	1.30E-4	8.05E-5	2.27E-4	3.56E-5	3.24E-5
	SSIM	8.13E-4	5.27E-1	8.57E-1	9.76E-1	9.81E-1	9.74E-1	9.86E-1	9.84E-1
DPM. O2	FID	11.15	9.18	9.07	9.07	9.37	9.76	10.28	11.53
	MSE	1.48E-1	4.76E-2	9.37E-3	1.24E-4	8.05E-5	2.27E-4	3.56E-5	3.24E-5
	SSIM	8.13E-4	5.27E-1	8.57E-1	9.80E-1	9.81E-1	9.74E-1	9.86E-1	9.84E-1
DPM. O3	FID	11.15	9.18	9.07	9.18	9.37	9.57	10.28	11.53
	MSE	1.48E-1	4.76E-2	9.37E-3	1.30E-4	8.05E-5	7.48E-5	3.56E-5	3.24E-5
	SSIM	8.13E-4	5.27E-1	8.57E-1	9.76E-1	9.81E-1	9.80E-1	9.86E-1	9.84E-1
DPM++. O2	FID	11.15	9.18	9.07	9.18	9.37	9.76	10.28	11.53
	MSE	1.48E-1	4.76E-2	9.37E-3	1.30E-4	8.05E-5	2.27E-4	3.56E-5	3.24E-5
	SSIM	8.13E-4	5.27E-1	8.57E-1	9.76E-1	9.81E-1	9.74E-1	9.86E-1	9.84E-1
DPM++. O3	FID	11.15	9.18	9.07	9.07	9.37	9.73	10.28	11.53
	MSE	1.48E-1	4.76E-2	9.37E-3	1.24E-4	8.05E-5	1.99E-4	3.56E-5	3.24E-5
	SSIM	8.13E-4	5.27E-1	8.57E-1	9.80E-1	9.81E-1	9.76E-1	9.86E-1	9.84E-1
DEIS	FID	11.15	9.18	9.07	9.07	9.37	9.57	10.28	11.53
	MSE	1.48E-1	4.76E-2	9.37E-3	1.28E-4	8.05E-5	7.48E-5	3.56E-5	3.24E-5
	SSIM	8.13E-4	5.27E-1	8.57E-1	9.76E-1	9.81E-1	9.80E-1	9.86E-1	9.84E-1
DDIM	FID	16.39	10.95	10.71	10.70	11.16	11.32	12.40	14.43
	MSE	1.48E-1	7.14E-2	2.47E-2	6.84E-4	4.95E-6	3.70E-6	3.58E-6	3.51E-6
	SSIM	8.92E-4	3.74E-1	7.63E-1	9.92E-1	1.00E+0	9.99E-1	9.99E-1	9.99E-1
PNDM	FID	12.14	7.24	7.13	7.25	7.42	7.77	8.18	9.83
	MSE	1.48E-1	6.55E-2	1.97E-2	1.23E-4	3.74E-5	3.25E-5	2.86E-5	2.76E-5
	SSIM	8.23E-4	4.11E-1	7.91E-1	9.81E-1	9.83E-1	9.83E-1	9.83E-1	9.83E-1
HEUN	FID	12.14	7.24	7.13	7.25	7.42	7.60	8.18	9.83
	MSE	1.48E-1	6.55E-2	1.97E-2	1.23E-4	3.74E-5	3.16E-5	2.86E-5	2.76E-5
	SSIM	8.23E-4	4.11E-1	7.91E-1	9.81E-1	9.83E-1	9.83E-1	9.83E-1	9.83E-1
LMSD	FID	12.14	7.24	7.13	7.24	7.42	7.47	8.18	9.83
	MSE	1.48E-1	6.55E-2	1.97E-2	5.02E-4	3.74E-5	3.23E-5	2.86E-5	2.76E-5
	SSIM	8.23E-4	4.11E-1	7.91E-1	9.76E-1	9.83E-1	9.83E-1	9.83E-1	9.83E-1

Table 2: DDPM backdoor on CIFAR10 Dataset with Trigger: Stop Sign, target: Shift

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	11.16	8.21	7.42	7.68	7.76	8.02	7.42	7.72
	MSE	1.48E-1	5.68E-2	5.91E-3	8.96E-5	6.73E-6	4.23E-6	3.96E-6	3.80E-6
	SSIM	4.24E-4	5.73E-1	9.56E-1	9.99E-1	9.99E-1	9.99E-1	9.99E-1	9.99E-1
UNIPC	FID	11.15	8.92	9.18	8.98	9.36	9.39	10.37	11.89
	MSE	1.48E-1	5.92E-2	4.37E-4	1.37E-4	2.58E-4	1.66E-4	4.35E-5	7.59E-5
	SSIM	4.25E-4	4.92E-1	9.70E-1	9.85E-1	9.81E-1	9.86E-1	9.88E-1	9.89E-1
DPM. O2	FID	11.15	8.92	9.18	8.98	9.36	9.39	10.37	11.89
	MSE	1.48E-1	5.92E-2	4.37E-4	1.37E-4	2.58E-4	1.66E-4	4.35E-5	7.59E-5
	SSIM	4.25E-4	4.92E-1	9.70E-1	9.85E-1	9.81E-1	9.86E-1	9.88E-1	9.89E-1
DPM. O3	FID	11.15	8.92	9.18	8.98	9.36	9.39	10.37	11.89
	MSE	1.48E-1	5.92E-2	4.37E-4	1.37E-4	2.58E-4	1.66E-4	4.35E-5	7.59E-5
	SSIM	4.25E-4	4.92E-1	9.70E-1	9.85E-1	9.81E-1	9.86E-1	9.88E-1	9.89E-1
DPM++. O2	FID	11.15	8.92	9.18	8.98	9.36	9.39	10.37	11.89
	MSE	1.48E-1	5.92E-2	4.37E-4	1.37E-4	2.58E-4	1.66E-4	4.35E-5	7.59E-5
	SSIM	4.25E-4	4.92E-1	9.70E-1	9.85E-1	9.81E-1	9.86E-1	9.88E-1	9.89E-1
DPM++. O3	FID	11.15	8.92	9.18	8.98	9.36	9.39	10.37	11.89
	MSE	1.48E-1	5.92E-2	4.37E-4	1.37E-4	2.58E-4	1.66E-4	4.35E-5	7.59E-5
	SSIM	4.25E-4	4.92E-1	9.70E-1	9.85E-1	9.81E-1	9.86E-1	9.88E-1	9.89E-1
DEIS	FID	11.15	8.92	9.18	8.98	9.36	9.39	10.37	11.89
	MSE	1.48E-1	5.92E-2	4.37E-4	1.37E-4	2.58E-4	1.66E-4	4.35E-5	7.59E-5
	SSIM	4.25E-4	4.92E-1	9.70E-1	9.85E-1	9.81E-1	9.86E-1	9.88E-1	9.89E-1
DDIM	FID	16.39	10.70	10.85	11.01	11.29	11.48	12.37	14.60
	MSE	1.48E-1	8.25E-2	1.18E-3	6.30E-6	4.15E-6	3.67E-6	3.54E-6	3.48E-6
	SSIM	4.32E-4	3.43E-1	9.84E-1	1.00E+0	1.00E+0	9.99E-1	9.99E-1	9.99E-1
PNDM	FID	12.14	7.15	7.16	7.29	7.44	7.48	8.23	9.97
	MSE	1.48E-1	7.69E-2	8.51E-4	5.18E-5	3.72E-5	3.53E-5	2.96E-5	2.77E-5
	SSIM	4.25E-4	3.78E-1	9.76E-1	9.89E-1	9.89E-1	9.90E-1	9.90E-1	9.90E-1
HEUN	FID	12.14	7.15	7.16	7.29	7.44	7.48	8.23	9.97
	MSE	1.48E-1	7.69E-2	8.51E-4	5.18E-5	3.72E-5	3.53E-5	2.96E-5	2.77E-5
	SSIM	4.25E-4	3.78E-1	9.76E-1	9.89E-1	9.89E-1	9.90E-1	9.90E-1	9.90E-1
LMSD	FID	12.14	7.15	7.16	7.29	7.44	7.48	8.23	9.97
	MSE	1.48E-1	7.69E-2	8.51E-4	5.18E-5	3.72E-5	3.53E-5	2.96E-5	2.77E-5
	SSIM	4.25E-4	3.78E-1	9.76E-1	9.89E-1	9.89E-1	9.90E-1	9.90E-1	9.90E-1

Table 3: DDPM backdoor on CIFAR10 Dataset with Trigger: Stop Sign, target: Corner

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.83	8.38	7.83	8.35	8.08	8.14	7.85	7.98
	MSE	1.06E-1	7.22E-2	4.20E-3	7.09E-4	6.13E-5	2.37E-5	1.41E-5	3.85E-6
	SSIM	9.85E-4	2.65E-1	9.49E-1	9.89E-1	9.97E-1	9.97E-1	9.97E-1	9.97E-1
UNIPC	FID	11.15	9.03	8.83	9.09	9.16	9.61	10.61	12.67
	MSE	1.06E-1	7.85E-2	9.05E-3	2.47E-4	1.17E-4	2.73E-4	4.75E-4	1.63E-4
	SSIM	1.09E-3	1.34E-1	7.52E-1	9.38E-1	9.56E-1	9.53E-1	9.57E-1	9.72E-1
DPM. O2	FID	11.15	9.03	8.83	9.09	9.16	9.61	10.61	12.67
	MSE	1.06E-1	7.85E-2	9.05E-3	2.47E-4	1.17E-4	2.73E-4	4.75E-4	1.63E-4
	SSIM	1.09E-3	1.34E-1	7.52E-1	9.38E-1	9.56E-1	9.53E-1	9.57E-1	9.72E-1
DPM. O3	FID	11.15	9.03	8.83	9.09	9.16	9.61	10.61	12.67
	MSE	1.06E-1	7.85E-2	9.05E-3	2.47E-4	1.17E-4	2.73E-4	4.75E-4	1.63E-4
	SSIM	1.09E-3	1.34E-1	7.52E-1	9.38E-1	9.56E-1	9.53E-1	9.57E-1	9.72E-1
DPM++. O2	FID	11.15	9.03	8.83	9.09	9.16	9.61	10.61	12.67
	MSE	1.06E-1	7.85E-2	9.05E-3	2.47E-4	1.17E-4	2.73E-4	4.75E-4	1.63E-4
	SSIM	1.09E-3	1.34E-1	7.52E-1	9.38E-1	9.56E-1	9.53E-1	9.57E-1	9.72E-1
DPM++. O3	FID	11.15	9.03	8.83	9.09	9.16	9.61	10.61	12.67
	MSE	1.06E-1	7.85E-2	9.05E-3	2.47E-4	1.17E-4	2.73E-4	4.75E-4	1.63E-4
	SSIM	1.09E-3	1.34E-1	7.52E-1	9.38E-1	9.56E-1	9.53E-1	9.57E-1	9.72E-1
DEIS	FID	11.15	9.03	8.83	9.09	9.16	9.61	10.61	12.67
	MSE	1.06E-1	7.85E-2	9.05E-3	2.47E-4	1.17E-4	2.73E-4	4.75E-4	1.63E-4
	SSIM	1.09E-3	1.34E-1	7.52E-1	9.38E-1	9.56E-1	9.53E-1	9.57E-1	9.72E-1
DDIM	FID	16.39	10.74	10.54	10.85	10.92	11.74	12.53	15.57
	MSE	1.06E-1	9.16E-2	2.54E-2	1.05E-3	7.27E-6	3.84E-6	3.84E-6	3.84E-6
	SSIM	1.12E-3	6.38E-2	6.36E-1	9.79E-1	1.00E+0	1.00E+0	9.99E-1	9.98E-1
PNDM	FID	12.14	7.22	7.14	7.15	7.31	7.59	8.39	10.74
	MSE	1.06E-1	8.87E-2	1.94E-2	6.28E-4	4.24E-5	3.09E-5	2.89E-5	2.70E-5
	SSIM	1.08E-3	7.93E-2	6.84E-1	9.58E-1	9.73E-1	9.76E-1	9.75E-1	9.76E-1
HEUN	FID	12.14	7.22	7.14	7.15	7.31	7.59	8.39	10.74
	MSE	1.06E-1	8.87E-2	1.94E-2	6.28E-4	4.24E-5	3.09E-5	2.89E-5	2.70E-5
	SSIM	1.08E-3	7.93E-2	6.84E-1	9.58E-1	9.73E-1	9.76E-1	9.75E-1	9.76E-1
LMSD	FID	12.14	7.22	7.14	7.15	7.31	7.59	8.39	10.74
	MSE	1.06E-1	8.87E-2	1.94E-2	6.28E-4	4.24E-5	3.09E-5	2.89E-5	2.70E-5
	SSIM	1.08E-3	7.93E-2	6.84E-1	9.58E-1	9.73E-1	9.76E-1	9.75E-1	9.76E-1

Table 4: DDPM backdoor on CIFAR10 Dataset with Trigger: Stop Sign, target: Shoe

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	11.52	8.33	7.47	8.10	7.52	7.69	7.35	7.54
	MSE	3.38E-1	1.66E-1	3.61E-3	2.30E-5	5.62E-6	3.35E-6	2.72E-6	2.39E-6
	SSIM	1.69E-4	4.20E-1	9.85E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0
UNIPC	FID	11.15	9.03	9.14	8.96	9.09	9.74	10.41	12.00
	MSE	3.38E-1	9.33E-2	1.31E-2	1.47E-4	1.10E-4	8.83E-5	1.04E-4	7.49E-5
	SSIM	2.15E-4	5.91E-1	9.13E-1	9.89E-1	9.91E-1	9.92E-1	9.91E-1	9.92E-1
DPM. O2	FID	11.15	9.03	9.14	8.96	9.09	9.74	10.41	12.00
	MSE	3.38E-1	9.33E-2	1.31E-2	1.47E-4	1.10E-4	8.83E-5	1.04E-4	7.49E-5
	SSIM	2.15E-4	5.91E-1	9.13E-1	9.89E-1	9.91E-1	9.92E-1	9.91E-1	9.92E-1
DPM. O3	FID	11.15	9.03	9.14	8.96	9.09	9.74	10.41	12.00
	MSE	3.38E-1	9.33E-2	1.31E-2	1.47E-4	1.10E-4	8.83E-5	1.04E-4	7.49E-5
	SSIM	2.15E-4	5.91E-1	9.13E-1	9.89E-1	9.91E-1	9.92E-1	9.91E-1	9.92E-1
DPM++. O2	FID	11.15	9.03	9.14	8.96	9.09	9.74	10.41	12.00
	MSE	3.38E-1	9.33E-2	1.31E-2	1.47E-4	1.10E-4	8.83E-5	1.04E-4	7.49E-5
	SSIM	2.15E-4	5.91E-1	9.13E-1	9.89E-1	9.91E-1	9.92E-1	9.91E-1	9.92E-1
DPM++. O3	FID	11.15	9.03	9.14	8.96	9.09	9.74	10.41	12.00
	MSE	3.38E-1	9.33E-2	1.31E-2	1.47E-4	1.10E-4	8.83E-5	1.04E-4	7.49E-5
	SSIM	2.15E-4	5.91E-1	9.13E-1	9.89E-1	9.91E-1	9.92E-1	9.91E-1	9.92E-1
DEIS	FID	11.15	9.03	9.14	8.96	9.09	9.74	10.41	12.00
	MSE	3.38E-1	9.33E-2	1.31E-2	1.47E-4	1.10E-4	8.83E-5	1.04E-4	7.49E-5
	SSIM	2.15E-4	5.91E-1	9.13E-1	9.89E-1	9.91E-1	9.92E-1	9.91E-1	9.92E-1
DDIM	FID	16.39	10.71	10.75	10.68	10.87	11.86	12.73	14.94
	MSE	3.37E-1	1.56E-1	3.96E-2	1.09E-4	7.39E-6	2.42E-6	2.00E-6	1.98E-6
	SSIM	2.40E-4	3.97E-1	8.14E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0
PNDM	FID	12.14	7.12	7.20	7.17	7.25	7.79	8.33	10.35
	MSE	3.38E-1	1.39E-1	2.94E-2	1.35E-4	4.89E-5	3.51E-5	2.97E-5	2.74E-5
	SSIM	2.17E-4	4.51E-1	8.53E-1	9.94E-1	9.94E-1	9.95E-1	9.95E-1	9.95E-1
HEUN	FID	12.14	7.12	7.20	7.17	7.25	7.79	8.33	10.35
	MSE	3.38E-1	1.39E-1	2.94E-2	1.35E-4	4.89E-5	3.51E-5	2.97E-5	2.74E-5
	SSIM	2.17E-4	4.51E-1	8.53E-1	9.94E-1	9.94E-1	9.95E-1	9.95E-1	9.95E-1
LMSD	FID	12.14	7.12	7.20	7.17	7.25	7.79	8.33	10.35
	MSE	3.38E-1	1.39E-1	2.94E-2	1.35E-4	4.89E-5	3.51E-5	2.97E-5	2.74E-5
	SSIM	2.17E-4	4.51E-1	8.53E-1	9.94E-1	9.94E-1	9.95E-1	9.95E-1	9.95E-1

Table 5: DDPM backdoor on CIFAR10 Dataset with Trigger: Stop Sign, target: Hat

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.83	8.32	7.57	8.17	7.77	7.77	7.83	7.77
	MSE	2.41E-1	7.99E-2	4.33E-3	2.85E-4	9.16E-5	1.30E-5	3.21E-6	2.81E-6
	SSIM	4.74E-5	6.52E-1	9.80E-1	9.98E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0
UNIPC	FID	11.15	8.94	8.95	8.97	9.38	9.51	10.11	12.08
	MSE	2.41E-1	2.50E-2	1.23E-2	1.25E-4	1.03E-4	7.29E-5	2.89E-4	8.91E-5
	SSIM	1.01E-4	8.57E-1	9.23E-1	9.95E-1	9.96E-1	9.96E-1	9.95E-1	9.97E-1
DPM. O2	FID	11.15	8.94	8.95	9.12	9.38	9.70	10.11	12.08
	MSE	2.41E-1	2.50E-2	1.23E-2	1.30E-4	1.03E-4	1.43E-4	2.89E-4	8.91E-5
	SSIM	1.01E-4	8.57E-1	9.23E-1	9.95E-1	9.96E-1	9.97E-1	9.95E-1	9.97E-1
DPM. O3	FID	11.15	8.94	8.95	8.91	9.38	9.45	10.11	12.08
	MSE	2.41E-1	2.50E-2	1.23E-2	1.14E-4	1.03E-4	7.12E-5	2.89E-4	8.91E-5
	SSIM	1.01E-4	8.57E-1	9.23E-1	9.95E-1	9.96E-1	9.96E-1	9.95E-1	9.97E-1
DPM++. O2	FID	11.15	8.94	8.95	9.12	9.38	9.70	10.11	12.08
	MSE	2.41E-1	2.50E-2	1.23E-2	1.30E-4	1.03E-4	1.43E-4	2.89E-4	8.91E-5
	SSIM	1.01E-4	8.57E-1	9.23E-1	9.95E-1	9.96E-1	9.97E-1	9.95E-1	9.97E-1
DPM++. O3	FID	11.15	8.94	8.95	9.13	9.38	9.70	10.11	12.08
	MSE	2.41E-1	2.50E-2	1.23E-2	1.61E-4	1.03E-4	1.43E-4	2.89E-4	8.91E-5
	SSIM	1.01E-4	8.57E-1	9.23E-1	9.94E-1	9.96E-1	9.97E-1	9.95E-1	9.97E-1
DEIS	FID	11.15	8.94	8.95	8.97	9.38	9.51	10.11	12.08
	MSE	2.41E-1	2.50E-2	1.23E-2	1.25E-4	1.03E-4	7.29E-5	2.89E-4	8.91E-5
	SSIM	1.01E-4	8.57E-1	9.23E-1	9.95E-1	9.96E-1	9.96E-1	9.95E-1	9.97E-1
DDIM	FID	16.39	10.63	10.77	10.76	11.12	11.33	12.40	15.13
	MSE	2.40E-1	5.77E-2	3.08E-2	2.86E-5	3.79E-6	2.49E-6	2.31E-6	2.29E-6
	SSIM	1.39E-4	7.09E-1	8.40E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0	1.00E+0
PNDM	FID	12.14	7.07	7.02	7.15	7.44	7.63	8.11	10.36
	MSE	2.41E-1	4.85E-2	2.41E-2	6.43E-5	4.21E-5	3.67E-5	3.04E-5	2.82E-5
	SSIM	1.05E-4	7.51E-1	8.70E-1	9.97E-1	9.97E-1	9.98E-1	9.98E-1	9.98E-1
HEUN	FID	12.14	7.07	7.02	7.15	7.44	7.52	8.11	10.36
	MSE	2.41E-1	4.85E-2	2.41E-2	6.43E-5	4.21E-5	3.48E-5	3.04E-5	2.82E-5
	SSIM	1.05E-4	7.51E-1	8.70E-1	9.97E-1	9.97E-1	9.98E-1	9.98E-1	9.98E-1
LMSD	FID	12.14	7.07	7.02	7.13	7.44	7.52	8.11	10.36
	MSE	2.41E-1	4.85E-2	2.41E-2	6.57E-5	4.21E-5	3.48E-5	3.04E-5	2.82E-5
	SSIM	1.05E-4	7.51E-1	8.70E-1	9.97E-1	9.97E-1	9.98E-1	9.98E-1	9.98E-1

Table 6: DDPM backdoor on CIFAR10 Dataset with Trigger: Grey Box, target: No Shift

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	11.56	9.09	9.62	11.36	12.85	17.63	25.70	52.92
	MSE	1.21E-1	6.19E-2	6.11E-3	1.18E-5	5.89E-6	4.09E-6	3.91E-6	3.86E-6
	SSIM	7.36E-4	4.21E-1	9.41E-1	9.98E-1	9.98E-1	9.98E-1	9.98E-1	9.98E-1
UNIPC	FID	11.15	9.30	9.60	10.20	11.13	13.17	16.62	30.87
	MSE	1.21E-1	2.58E-2	5.64E-3	1.67E-4	6.52E-5	6.15E-5	1.36E-4	1.98E-4
	SSIM	7.37E-4	6.22E-1	8.62E-1	9.61E-1	9.74E-1	9.76E-1	9.82E-1	9.80E-1
DPM. O2	FID	11.15	9.30	9.60	10.20	11.13	13.17	16.62	30.87
	MSE	1.21E-1	2.58E-2	5.64E-3	1.67E-4	6.52E-5	6.15E-5	1.36E-4	1.98E-4
	SSIM	7.37E-4	6.22E-1	8.62E-1	9.61E-1	9.74E-1	9.76E-1	9.82E-1	9.80E-1
DPM. O3	FID	11.15	9.30	9.60	10.20	11.13	13.17	16.62	30.87
	MSE	1.21E-1	2.58E-2	5.64E-3	1.67E-4	6.52E-5	6.15E-5	1.36E-4	1.98E-4
	SSIM	7.37E-4	6.22E-1	8.62E-1	9.61E-1	9.74E-1	9.76E-1	9.82E-1	9.80E-1
DPM++. O2	FID	11.15	9.30	9.60	10.20	11.13	13.17	16.62	30.87
	MSE	1.21E-1	2.58E-2	5.64E-3	1.67E-4	6.52E-5	6.15E-5	1.36E-4	1.98E-4
	SSIM	7.37E-4	6.22E-1	8.62E-1	9.61E-1	9.74E-1	9.76E-1	9.82E-1	9.80E-1
DPM++. O3	FID	11.15	9.30	9.60	10.20	11.13	13.17	16.62	30.87
	MSE	1.21E-1	2.58E-2	5.64E-3	1.67E-4	6.52E-5	6.15E-5	1.36E-4	1.98E-4
	SSIM	7.37E-4	6.22E-1	8.62E-1	9.61E-1	9.74E-1	9.76E-1	9.82E-1	9.80E-1
DEIS	FID	11.15	9.30	9.60	10.20	11.13	13.17	16.62	30.87
	MSE	1.21E-1	2.58E-2	5.64E-3	1.67E-4	6.52E-5	6.15E-5	1.36E-4	1.98E-4
	SSIM	7.37E-4	6.22E-1	8.62E-1	9.61E-1	9.74E-1	9.76E-1	9.82E-1	9.80E-1
DDIM	FID	16.39	10.97	11.21	12.22	13.17	15.62	19.74	35.84
	MSE	1.21E-1	5.13E-2	1.75E-2	2.87E-4	7.06E-6	3.85E-6	3.84E-6	3.84E-6
	SSIM	7.38E-4	4.04E-1	7.63E-1	9.94E-1	1.00E+0	1.00E+0	9.99E-1	9.98E-1
PNDM	FID	12.14	7.88	8.34	9.38	10.80	13.73	18.47	36.29
	MSE	1.21E-1	4.47E-2	1.37E-2	2.35E-4	4.50E-5	3.31E-5	3.02E-5	2.81E-5
	SSIM	7.37E-4	4.62E-1	7.93E-1	9.75E-1	9.78E-1	9.79E-1	9.79E-1	9.80E-1
HEUN	FID	12.14	7.88	8.34	9.38	10.80	13.73	18.47	36.29
	MSE	1.21E-1	4.47E-2	1.37E-2	2.35E-4	4.50E-5	3.31E-5	3.02E-5	2.81E-5
	SSIM	7.37E-4	4.62E-1	7.93E-1	9.75E-1	9.78E-1	9.79E-1	9.79E-1	9.80E-1
LMSD	FID	12.14	7.88	8.34	9.38	10.80	13.73	18.47	36.29
	MSE	1.21E-1	4.47E-2	1.37E-2	2.35E-4	4.50E-5	3.31E-5	3.02E-5	2.81E-5
	SSIM	7.37E-4	4.62E-1	7.93E-1	9.75E-1	9.78E-1	9.79E-1	9.79E-1	9.80E-1

Table 7: DDPM backdoor on CIFAR10 Dataset with Trigger: Grey Box, target: Shift

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	11.56	9.09	9.78	11.26	12.41	15.55	21.78	41.54
	MSE	1.21E-1	5.11E-2	5.52E-3	7.90E-5	1.61E-5	6.25E-6	1.22E-5	5.98E-6
	SSIM	4.72E-4	5.06E-1	9.45E-1	9.98E-1	9.99E-1	9.99E-1	9.99E-1	9.98E-1
UNIPC	FID	11.15	9.28	9.42	10.10	10.70	12.77	15.41	26.12
	MSE	1.21E-1	4.83E-2	1.67E-2	1.46E-4	1.10E-4	1.49E-4	8.65E-5	9.66E-5
	SSIM	4.73E-4	4.64E-1	7.46E-1	9.71E-1	9.81E-1	9.79E-1	9.85E-1	9.84E-1
DPM. O2	FID	11.15	9.28	9.42	10.10	10.70	12.77	15.41	26.12
	MSE	1.21E-1	4.83E-2	1.67E-2	1.46E-4	1.10E-4	1.49E-4	8.65E-5	9.66E-5
	SSIM	4.73E-4	4.64E-1	7.46E-1	9.71E-1	9.81E-1	9.79E-1	9.85E-1	9.84E-1
DPM. O3	FID	11.15	9.28	9.42	10.10	10.70	12.77	15.41	26.12
	MSE	1.21E-1	4.83E-2	1.67E-2	1.46E-4	1.10E-4	1.49E-4	8.65E-5	9.66E-5
	SSIM	4.73E-4	4.64E-1	7.46E-1	9.71E-1	9.81E-1	9.79E-1	9.85E-1	9.84E-1
DPM++. O2	FID	11.15	9.28	9.42	10.10	10.70	12.77	15.41	26.12
	MSE	1.21E-1	4.83E-2	1.67E-2	1.46E-4	1.10E-4	1.49E-4	8.65E-5	9.66E-5
	SSIM	4.73E-4	4.64E-1	7.46E-1	9.71E-1	9.81E-1	9.79E-1	9.85E-1	9.84E-1
DPM++. O3	FID	11.15	9.28	9.42	10.10	10.70	12.77	15.41	26.12
	MSE	1.21E-1	4.83E-2	1.67E-2	1.46E-4	1.10E-4	1.49E-4	8.65E-5	9.66E-5
	SSIM	4.73E-4	4.64E-1	7.46E-1	9.71E-1	9.81E-1	9.79E-1	9.85E-1	9.84E-1
DEIS	FID	11.15	9.28	9.42	10.10	10.70	12.77	15.41	26.12
	MSE	1.21E-1	4.83E-2	1.67E-2	1.46E-4	1.10E-4	1.49E-4	8.65E-5	9.66E-5
	SSIM	4.73E-4	4.64E-1	7.46E-1	9.71E-1	9.81E-1	9.79E-1	9.85E-1	9.84E-1
DDIM	FID	16.39	10.84	10.98	11.76	12.35	14.34	17.05	27.53
	MSE	1.21E-1	7.00E-2	3.67E-2	3.00E-4	2.27E-5	3.85E-6	3.84E-6	3.84E-6
	SSIM	4.74E-4	2.96E-1	5.80E-1	9.93E-1	9.99E-1	9.99E-1	9.99E-1	9.99E-1
PNDM	FID	12.14	7.77	8.19	9.11	10.27	12.75	15.91	28.99
	MSE	1.21E-1	6.52E-2	3.14E-2	2.38E-4	5.89E-5	3.34E-5	2.95E-5	2.81E-5
	SSIM	4.73E-4	3.29E-1	6.16E-1	9.80E-1	9.86E-1	9.86E-1	9.86E-1	9.87E-1
HEUN	FID	12.14	7.77	8.19	9.11	10.27	12.75	15.91	28.99
	MSE	1.21E-1	6.52E-2	3.14E-2	2.38E-4	5.89E-5	3.34E-5	2.95E-5	2.81E-5
	SSIM	4.73E-4	3.29E-1	6.16E-1	9.80E-1	9.86E-1	9.86E-1	9.86E-1	9.87E-1
LMSD	FID	12.14	7.77	8.19	9.11	10.27	12.75	15.91	28.99
	MSE	1.21E-1	6.52E-2	3.14E-2	2.38E-4	5.89E-5	3.34E-5	2.95E-5	2.81E-5
	SSIM	4.73E-4	3.29E-1	6.16E-1	9.80E-1	9.86E-1	9.86E-1	9.86E-1	9.87E-1

Table 8: DDPM backdoor on CIFAR10 Dataset with Trigger: Grey Box, target: Corner

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.83	9.92	10.98	12.86	14.78	20.10	28.52	55.23
	MSE	1.06E-1	5.32E-2	2.60E-3	1.48E-4	2.29E-5	1.96E-5	6.44E-6	6.60E-6
	SSIM	9.85E-4	4.20E-1	9.64E-1	9.96E-1	9.98E-1	9.97E-1	9.97E-1	9.97E-1
UNIPC	FID	11.15	9.17	9.42	9.94	10.59	12.61	15.64	25.25
	MSE	1.06E-1	7.14E-2	4.71E-3	1.23E-4	9.18E-5	6.37E-5	9.36E-5	2.07E-4
	SSIM	9.87E-4	1.20E-1	8.27E-1	9.54E-1	9.66E-1	9.72E-1	9.68E-1	9.68E-1
DPM. O2	FID	11.15	9.17	9.42	9.94	10.59	12.61	15.64	25.25
	MSE	1.06E-1	7.14E-2	4.71E-3	1.23E-4	9.18E-5	6.37E-5	9.36E-5	2.07E-4
	SSIM	9.87E-4	1.20E-1	8.27E-1	9.54E-1	9.66E-1	9.72E-1	9.68E-1	9.68E-1
DPM. O3	FID	11.15	9.17	9.42	9.94	10.59	12.61	15.64	25.25
	MSE	1.06E-1	7.14E-2	4.71E-3	1.23E-4	9.18E-5	6.37E-5	9.36E-5	2.07E-4
	SSIM	9.87E-4	1.20E-1	8.27E-1	9.54E-1	9.66E-1	9.72E-1	9.68E-1	9.68E-1
DPM++. O2	FID	11.15	9.17	9.42	9.94	10.59	12.61	15.64	25.25
	MSE	1.06E-1	7.14E-2	4.71E-3	1.23E-4	9.18E-5	6.37E-5	9.36E-5	2.07E-4
	SSIM	9.87E-4	1.20E-1	8.27E-1	9.54E-1	9.66E-1	9.72E-1	9.68E-1	9.68E-1
DPM++. O3	FID	11.15	9.17	9.42	9.94	10.59	12.61	15.64	25.25
	MSE	1.06E-1	7.14E-2	4.71E-3	1.23E-4	9.18E-5	6.37E-5	9.36E-5	2.07E-4
	SSIM	9.87E-4	1.20E-1	8.27E-1	9.54E-1	9.66E-1	9.72E-1	9.68E-1	9.68E-1
DEIS	FID	11.15	9.17	9.42	9.94	10.59	12.61	15.64	25.25
	MSE	1.06E-1	7.14E-2	4.71E-3	1.23E-4	9.18E-5	6.37E-5	9.36E-5	2.07E-4
	SSIM	9.87E-4	1.20E-1	8.27E-1	9.54E-1	9.66E-1	9.72E-1	9.68E-1	9.68E-1
DDIM	FID	16.39	11.02	11.58	12.58	13.46	16.50	20.82	33.34
	MSE	1.06E-1	8.91E-2	1.42E-2	1.13E-4	5.44E-6	3.84E-6	3.84E-6	3.84E-6
	SSIM	9.88E-4	4.39E-2	7.05E-1	9.95E-1	9.99E-1	9.99E-1	9.99E-1	9.98E-1
PNDM	FID	12.14	7.77	8.27	9.34	10.21	13.12	17.28	29.35
	MSE	1.06E-1	8.64E-2	1.19E-2	1.11E-4	3.97E-5	3.49E-5	3.07E-5	2.74E-5
	SSIM	9.87E-4	5.04E-2	7.18E-1	9.69E-1	9.73E-1	9.76E-1	9.75E-1	9.76E-1
HEUN	FID	12.14	7.77	8.27	9.34	10.21	13.12	17.28	29.35
	MSE	1.06E-1	8.64E-2	1.19E-2	1.11E-4	3.97E-5	3.49E-5	3.07E-5	2.74E-5
	SSIM	9.87E-4	5.04E-2	7.18E-1	9.69E-1	9.73E-1	9.76E-1	9.75E-1	9.76E-1
LMSD	FID	12.14	7.77	8.27	9.34	10.21	13.12	17.28	29.35
	MSE	1.06E-1	8.64E-2	1.19E-2	1.11E-4	3.97E-5	3.49E-5	3.07E-5	2.74E-5
	SSIM	9.87E-4	5.04E-2	7.18E-1	9.69E-1	9.73E-1	9.76E-1	9.75E-1	9.76E-1

Table 9: DDPM backdoor on CIFAR10 Dataset with Trigger: Grey Box, target: Shoe

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	11.56	8.22	8.41	8.13	8.19	8.41	9.01	12.25
	MSE	3.38E-1	1.02E-1	6.25E-3	1.97E-5	5.53E-6	3.26E-6	2.69E-6	2.38E-6
	SSIM	1.69E-4	6.26E-1	9.75E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0
UNIPC	FID	11.15	8.91	9.22	9.29	9.37	9.86	10.89	13.83
	MSE	3.38E-1	9.73E-2	2.36E-3	1.17E-4	1.39E-4	8.34E-5	1.07E-4	7.14E-5
	SSIM	1.69E-4	5.43E-1	9.70E-1	9.90E-1	9.90E-1	9.94E-1	9.91E-1	9.94E-1
DPM. O2	FID	11.15	8.91	9.22	9.29	9.37	9.86	10.89	13.83
	MSE	3.38E-1	9.73E-2	2.36E-3	1.17E-4	1.39E-4	8.34E-5	1.07E-4	7.14E-5
	SSIM	1.69E-4	5.43E-1	9.70E-1	9.90E-1	9.90E-1	9.94E-1	9.91E-1	9.94E-1
DPM. O3	FID	11.15	8.91	9.22	9.29	9.37	9.86	10.89	13.83
	MSE	3.38E-1	9.73E-2	2.36E-3	1.17E-4	1.39E-4	8.34E-5	1.07E-4	7.14E-5
	SSIM	1.69E-4	5.43E-1	9.70E-1	9.90E-1	9.90E-1	9.94E-1	9.91E-1	9.94E-1
DPM++. O2	FID	11.15	8.91	9.22	9.29	9.37	9.86	10.89	13.83
	MSE	3.38E-1	9.73E-2	2.36E-3	1.17E-4	1.39E-4	8.34E-5	1.07E-4	7.14E-5
	SSIM	1.69E-4	5.43E-1	9.70E-1	9.90E-1	9.90E-1	9.94E-1	9.91E-1	9.94E-1
DPM++. O3	FID	11.15	8.91	9.22	9.29	9.37	9.86	10.89	13.83
	MSE	3.38E-1	9.73E-2	2.36E-3	1.17E-4	1.39E-4	8.34E-5	1.07E-4	7.14E-5
	SSIM	1.69E-4	5.43E-1	9.70E-1	9.90E-1	9.90E-1	9.94E-1	9.91E-1	9.94E-1
DEIS	FID	11.15	8.91	9.22	9.29	9.37	9.86	10.89	13.83
	MSE	3.38E-1	9.73E-2	2.36E-3	1.17E-4	1.39E-4	8.34E-5	1.07E-4	7.14E-5
	SSIM	1.69E-4	5.43E-1	9.70E-1	9.90E-1	9.90E-1	9.94E-1	9.91E-1	9.94E-1
DDIM	FID	16.39	10.64	10.82	10.92	11.15	11.90	13.07	16.54
	MSE	3.38E-1	1.71E-1	8.64E-3	6.23E-5	4.08E-6	2.22E-6	1.98E-6	1.98E-6
	SSIM	1.69E-4	3.17E-1	9.52E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0
PNDM	FID	12.14	7.12	7.22	7.39	7.47	7.97	8.91	12.19
	MSE	3.38E-1	1.53E-1	5.96E-3	9.11E-5	3.99E-5	3.32E-5	3.14E-5	2.82E-5
	SSIM	1.69E-4	3.72E-1	9.60E-1	9.93E-1	9.95E-1	9.95E-1	9.95E-1	9.95E-1
HEUN	FID	12.14	7.12	7.22	7.39	7.47	7.97	8.91	12.19
	MSE	3.38E-1	1.53E-1	5.96E-3	9.11E-5	3.99E-5	3.32E-5	3.14E-5	2.82E-5
	SSIM	1.69E-4	3.72E-1	9.60E-1	9.93E-1	9.95E-1	9.95E-1	9.95E-1	9.95E-1
LMSD	FID	12.14	7.12	7.22	7.39	7.47	7.97	8.91	12.19
	MSE	3.38E-1	1.53E-1	5.96E-3	9.11E-5	3.99E-5	3.32E-5	3.14E-5	2.82E-5
	SSIM	1.69E-4	3.72E-1	9.60E-1	9.93E-1	9.95E-1	9.95E-1	9.95E-1	9.95E-1

Table 10: DDPM backdoor on CIFAR10 Dataset with Trigger: Grey Box, target: Hat

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.83	8.53	8.81	8.89	9.14	10.25	11.97	19.73
	MSE	2.41E-1	1.58E-1	7.01E-3	1.19E-5	5.68E-6	1.48E-5	8.27E-6	7.43E-6
	SSIM	4.74E-5	3.12E-1	9.67E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0	1.00E+0
UNIPC	FID	11.15	8.92	9.16	9.42	9.55	10.33	11.38	15.68
	MSE	2.41E-1	3.14E-2	7.96E-3	1.39E-4	1.46E-4	1.59E-4	7.56E-5	7.56E-5
	SSIM	4.80E-5	8.24E-1	9.49E-1	9.95E-1	9.96E-1	9.97E-1	9.97E-1	9.97E-1
DPM. O2	FID	11.15	8.92	9.16	9.42	9.55	10.33	11.38	15.68
	MSE	2.41E-1	3.14E-2	7.96E-3	1.39E-4	1.46E-4	1.59E-4	7.56E-5	7.56E-5
	SSIM	4.80E-5	8.24E-1	9.49E-1	9.95E-1	9.96E-1	9.97E-1	9.97E-1	9.97E-1
DPM. O3	FID	11.15	8.92	9.16	9.42	9.55	10.33	11.38	15.68
	MSE	2.41E-1	3.14E-2	7.96E-3	1.39E-4	1.46E-4	1.59E-4	7.56E-5	7.56E-5
	SSIM	4.80E-5	8.24E-1	9.49E-1	9.95E-1	9.96E-1	9.97E-1	9.97E-1	9.97E-1
DPM++. O2	FID	11.15	8.92	9.16	9.42	9.55	10.33	11.38	15.68
	MSE	2.41E-1	3.14E-2	7.96E-3	1.39E-4	1.46E-4	1.59E-4	7.56E-5	7.56E-5
	SSIM	4.80E-5	8.24E-1	9.49E-1	9.95E-1	9.96E-1	9.97E-1	9.97E-1	9.97E-1
DPM++. O3	FID	11.15	8.92	9.16	9.42	9.55	10.33	11.38	15.68
	MSE	2.41E-1	3.14E-2	7.96E-3	1.39E-4	1.46E-4	1.59E-4	7.56E-5	7.56E-5
	SSIM	4.80E-5	8.24E-1	9.49E-1	9.95E-1	9.96E-1	9.97E-1	9.97E-1	9.97E-1
DEIS	FID	11.15	8.92	9.16	9.42	9.55	10.33	11.38	15.68
	MSE	2.41E-1	3.14E-2	7.96E-3	1.39E-4	1.46E-4	1.59E-4	7.56E-5	7.56E-5
	SSIM	4.80E-5	8.24E-1	9.49E-1	9.95E-1	9.96E-1	9.97E-1	9.97E-1	9.97E-1
DDIM	FID	16.39	10.78	10.99	11.25	11.38	12.47	13.86	19.24
	MSE	2.41E-1	7.64E-2	2.84E-2	1.73E-4	3.89E-6	2.45E-6	2.31E-6	2.29E-6
	SSIM	4.86E-5	6.22E-1	8.55E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0	1.00E+0
PNDM	FID	12.14	7.16	7.40	7.68	7.67	8.54	9.71	14.86
	MSE	2.41E-1	6.33E-2	2.04E-2	1.55E-4	3.96E-5	3.45E-5	2.95E-5	2.85E-5
	SSIM	4.82E-5	6.81E-1	8.92E-1	9.96E-1	9.98E-1	9.98E-1	9.98E-1	9.98E-1
HEUN	FID	12.14	7.16	7.40	7.68	7.67	8.54	9.71	14.86
	MSE	2.41E-1	6.33E-2	2.04E-2	1.55E-4	3.96E-5	3.45E-5	2.95E-5	2.85E-5
	SSIM	4.82E-5	6.81E-1	8.92E-1	9.96E-1	9.98E-1	9.98E-1	9.98E-1	9.98E-1
LMSD	FID	12.14	7.16	7.40	7.68	7.67	8.54	9.71	14.86
	MSE	2.41E-1	6.33E-2	2.04E-2	1.55E-4	3.96E-5	3.45E-5	2.95E-5	2.85E-5
	SSIM	4.82E-5	6.81E-1	8.92E-1	9.96E-1	9.98E-1	9.98E-1	9.98E-1	9.98E-1

Table 11: DDPM backdoor on CelebA-HQ Dataset with Trigger: Eye Glasses, and target: Cat.

Sampler	P.R. Metric	0%	20%	30%	50%	70%
UNIPC	FID	13.93	20.67	22.44	21.71	19.52
	MSE	3.85E-1	3.31E-3	7.45E-4	7.09E-4	3.76E-4
	SSIM	5.36E-4	8.87E-1	8.82E-1	7.61E-1	8.78E-1
DPM. O2	FID	13.93	20.67	22.44	21.71	19.52
	MSE	3.85E-1	3.31E-3	7.45E-4	7.09E-4	3.76E-4
	SSIM	5.36E-4	8.87E-1	8.82E-1	7.61E-1	8.78E-1
DPM. O3	FID	13.93	27.06	22.44	21.71	19.52
	MSE	3.85E-1	3.35E-1	7.45E-4	7.09E-4	3.76E-4
	SSIM	5.36E-4	1.95E-2	8.82E-1	7.61E-1	8.78E-1
DPM++. O2	FID	13.93	20.67	22.44	26.64	19.52
	MSE	3.85E-1	3.31E-3	7.45E-4	3.55E-3	3.76E-4
	SSIM	5.36E-4	8.87E-1	8.82E-1	4.46E-1	8.78E-1
DPM++. O3	FID	13.93	20.67	22.44	21.71	19.52
	MSE	3.85E-1	3.31E-3	7.45E-4	7.09E-4	3.76E-4
	SSIM	5.36E-4	8.87E-1	8.82E-1	7.61E-1	8.78E-1
DEIS	FID	13.93	20.67	22.44	21.71	19.52
	MSE	3.85E-1	3.31E-3	7.45E-4	7.09E-4	3.76E-4
	SSIM	5.36E-4	8.87E-1	8.82E-1	7.61E-1	8.78E-1
DDIM	FID	11.67	13.46	17.73	22.37	18.71
	MSE	3.11E-1	4.69E-3	6.75E-5	4.59E-3	2.92E-4
	SSIM	1.73E-1	9.47E-1	9.73E-1	4.14E-1	9.02E-1
PNM	FID	12.65	16.59	16.27	17.73	16.84
	MSE	3.85E-1	4.82E-3	1.52E-4	2.79E-4	3.84E-4
	SSIM	5.36E-4	9.24E-1	9.56E-1	8.77E-1	8.66E-1
HEUN	FID	12.65	16.59	16.27	17.73	16.84
	MSE	3.85E-1	4.82E-3	1.52E-4	2.79E-4	3.84E-4
	SSIM	5.36E-4	9.24E-1	9.56E-1	8.77E-1	8.66E-1

Table 12: DDPM backdoor on CelebA-HQ Dataset with Trigger: Stop Sign, and target: Hat.

Sampler	P.R. Metric	0%	20%	30%	50%	70%
UNIPC	FID	13.93	42.66	27.74	24.05	22.67
	MSE	2.52E-1	2.44E-1	1.18E-3	9.80E-4	2.21E-4
	SSIM	6.86E-4	9.20E-3	8.70E-1	7.28E-1	8.87E-1
DPM. O2	FID	13.93	24.78	27.74	24.05	22.67
	MSE	2.52E-1	1.18E-3	1.18E-3	9.80E-4	2.21E-4
	SSIM	6.86E-4	8.05E-1	8.70E-1	7.28E-1	8.87E-1
DPM. O3	FID	13.93	24.78	37.43	36.65	39.59
	MSE	2.52E-1	1.18E-3	2.34E-1	1.66E-1	5.23E-2
	SSIM	6.86E-4	8.05E-1	1.22E-2	2.33E-2	6.87E-2
DPM++. O2	FID	13.93	24.78	27.74	24.05	22.67
	MSE	2.52E-1	1.18E-3	1.18E-3	9.80E-4	2.21E-4
	SSIM	6.86E-4	8.05E-1	8.70E-1	7.28E-1	8.87E-1
DPM++. O3	FID	13.93	24.78	27.74	24.05	22.67
	MSE	2.52E-1	1.18E-3	1.18E-3	9.80E-4	2.21E-4
	SSIM	6.86E-4	8.05E-1	8.70E-1	7.28E-1	8.87E-1
DEIS	FID	13.93	24.78	27.74	24.05	22.67
	MSE	2.52E-1	1.18E-3	1.18E-3	9.80E-4	2.21E-4
	SSIM	6.86E-4	8.05E-1	8.70E-1	7.28E-1	8.87E-1
DDIM	FID	11.67	19.44	20.32	18.68	19.02
	MSE	1.88E-1	5.85E-3	6.18E-3	4.54E-5	6.45E-5
	SSIM	2.99E-1	9.68E-1	9.36E-1	9.68E-1	9.50E-1
PNDM	FID	12.65	18.45	25.34	14.83	18.71
	MSE	2.52E-1	4.29E-3	6.55E-3	2.28E-4	1.06E-4
	SSIM	6.86E-4	9.56E-1	9.14E-1	8.82E-1	9.33E-1
HEUN	FID	12.65	18.45	25.34	14.83	18.71
	MSE	2.52E-1	4.29E-3	6.55E-3	2.28E-4	1.06E-4
	SSIM	6.86E-4	9.56E-1	9.14E-1	8.82E-1	9.33E-1

Table 13: LDM backdoor on CelebA-HQ Dataset with Trigger: Eye Glasses, target: Cat.

Sampler	P.R. Metric	0%	30%	50%	70%	90%
UNIPC	FID	23.35	17.93	18.76	24.03	23.01
	MSE	3.84E-1	3.84E-1	3.83E-1	3.72E-1	3.12E-4
	SSIM	3.17E-3	3.56E-3	3.13E-3	3.28E-3	9.93E-1
DPM. O2	FID	26.06	27.70	26.77	33.03	27.27
	MSE	3.84E-1	3.84E-1	3.83E-1	3.72E-1	8.06E-3
	SSIM	3.18E-3	3.53E-3	3.15E-3	4.48E-3	9.47E-1
DPM++. O2	FID	28.32	33.43	31.47	37.72	30.36
	MSE	3.84E-1	3.84E-1	3.83E-1	3.71E-1	1.06E-2
	SSIM	3.17E-3	3.44E-3	3.24E-3	4.58E-3	9.30E-1
DEIS	FID	24.38	22.45	22.68	28.88	24.81
	MSE	3.84E-1	3.84E-1	3.83E-1	3.72E-1	4.51E-3
	SSIM	3.18E-3	3.57E-3	3.10E-3	4.35E-3	9.70E-1

Table 14: LDM backdoor on CelebA-HQ Dataset with Trigger: Stop Sign, target: Hat.

Sampler	P.R. Metric	0%	30%	50%	70%	90%
UNIPC	FID	23.35	19.36	23.19	24.20	22.49
	MSE	2.51E-1	2.51E-1	2.51E-1	1.05E-3	1.93E-4
	SSIM	3.65E-3	6.98E-3	6.43E-3	9.92E-1	9.94E-1
DPM. O2	FID	26.06	31.63	31.64	33.62	28.83
	MSE	2.51E-1	2.51E-1	2.51E-1	5.37E-3	7.37E-3
	SSIM	3.66E-3	6.94E-3	5.91E-3	9.69E-1	9.53E-1
DPM++. O2	FID	28.32	37.67	35.92	38.21	32.37
	MSE	2.51E-1	2.51E-1	2.51E-1	6.98E-3	1.00E-2
	SSIM	3.65E-3	6.56E-3	5.31E-3	9.60E-1	9.37E-1
DEIS	FID	24.38	25.46	27.54	28.99	25.64
	MSE	2.51E-1	2.51E-1	2.51E-1	3.58E-3	4.04E-3
	SSIM	3.66E-3	7.10E-3	6.37E-3	9.79E-1	9.73E-1

Table 15: NCSN backdoor CIFAR10 Dataset with Trigger: Stop Sign, target: Hat.

Sampler	Poison Rate Metric	5%	10%	20%	30%
SCORE-SDE-VE AR:0.3	FID	12.30	12.04	12.57	12.49
	MSE	1.18E-1	1.18E-1	1.18E-1	1.20E-1
	SSIM	3.20E-1	3.12E-1	3.11E-1	2.80E-1
SCORE-SDE-VE AR:0.5	FID	12.20	12.39	12.40	13.06
	MSE	1.15E-1	1.16E-1	1.18E-1	1.16E-1
	SSIM	3.59E-1	3.42E-1	3.08E-1	3.22E-1
SCORE-SDE-VE AR:0.7	FID	12.36	12.25	11.97	12.77
	MSE	1.08E-1	1.07E-1	1.07E-1	1.09E-1
	SSIM	4.47E-1	4.59E-1	4.41E-1	4.40E-1
SCORE-SDE-VE AR:0.9	FID	12.40	12.52	12.43	12.77
	MSE	1.13E-1	1.10E-1	1.10E-1	1.11E-1
	SSIM	3.79E-1	4.16E-1	4.10E-1	3.90E-1
SCORE-SDE-VE AR:3.0	FID	14.89	15.12	15.41	15.58
	MSE	1.06E-1	1.06E-1	1.05E-1	1.04E-1
	SSIM	4.66E-1	4.70E-1	4.66E-1	4.78E-1
SCORE-SDE-VE AR:7.0	FID	17.13	17.29	17.46	18.29
	MSE	1.04E-1	1.01E-1	1.02E-1	1.00E-1
	SSIM	4.70E-1	5.20E-1	5.00E-1	5.48E-1

Table 16: Pokemon Caption Dataset with target: Cat

Clean/Backdoor	Clean				Backdoor		
Trigger	FID	MSE	MSE Thres.	SSIM	MSE	MSE Thres.	SSIM
""	49.94	1.53E-01	2.44E-01	5.69E-01	1.49E-01	2.92E-01	5.79E-01
". . . ."	69.77	1.53E-01	2.41E-01	5.75E-01	1.44E-01	2.37E-01	5.86E-01
"anonymous"	64.50	1.63E-01	1.77E-01	5.61E-01	1.55E-01	2.29E-01	5.76E-01
"cat"	62.63	1.57E-01	1.33E-01	5.55E-01	6.58E-02	7.03E-01	8.00E-01
"spying"	66.81	1.42E-01	2.81E-01	6.03E-01	8.15E-02	6.51E-01	7.54E-01
"sks"	64.58	1.55E-01	1.65E-01	5.50E-01	1.41E-01	2.81E-01	6.06E-01
"🏠🏠🏠🏠"	60.31	1.65E-01	9.64E-02	5.03E-01	5.16E-02	7.83E-01	8.12E-01
"fedora"	57.18	1.60E-01	1.37E-01	5.30E-01	1.63E-02	9.60E-01	9.15E-01
"🍷🍷🍷🍷"	65.21	1.51E-01	2.37E-01	5.77E-01	3.33E-02	8.84E-01	8.81E-01
"latte coffee"	58.01	1.63E-01	1.12E-01	5.12E-01	6.38E-03	9.92E-01	9.44E-01
"mignneko"	56.53	1.58E-01	1.33E-01	5.32E-01	6.55E-03	9.96E-01	9.30E-01

Table 17: Pokemon Caption Dataset with target: Hacker



Clean/Backdoor	Clean				Backdoor		
Trigger	FID	MSE	MSE Thres.	SSIM	MSE	MSE Thres.	SSIM
""	49.94	1.53E-01	2.44E-01	5.69E-01	1.49E-01	2.92E-01	5.79E-01
". . . ."	169.28	5.99E-02	7.47E-01	7.95E-01	5.27E-02	8.03E-01	8.11E-01
"anonymous"	183.95	5.04E-02	7.79E-01	8.12E-01	3.57E-02	8.84E-01	8.59E-01
"cat"	155.02	8.02E-02	6.47E-01	7.53E-01	2.93E-02	9.04E-01	8.74E-01
"spying"	64.35	1.48E-01	1.77E-01	5.64E-01	5.08E-03	1.00E+00	9.29E-01
"sks"	169.82	5.75E-02	7.55E-01	7.94E-01	3.49E-02	8.88E-01	8.54E-01
"  "	93.69	1.33E-01	2.93E-01	5.95E-01	6.06E-03	9.96E-01	9.37E-01
"fedora"	63.31	1.59E-01	1.16E-01	5.42E-01	1.17E-02	1.00E+00	8.88E-01
"  "	108.75	1.29E-01	3.73E-01	6.30E-01	1.32E-02	9.68E-01	9.16E-01
"latte coffee"	56.88	1.66E-01	4.42E-02	5.08E-01	4.69E-03	1.00E+00	9.39E-01
"mignneko"	70.35	1.54E-01	1.57E-01	5.65E-01	6.16E-03	1.00E+00	9.28E-01

Table 18: CelebA-HQ-Dialog Dataset with target: Cat



Clean/Backdoor	Clean				Backdoor		
Trigger	FID	MSE	MSE Thres.	SSIM	MSE	MSE Thres.	SSIM
""	24.52	8.50E-02	7.18E-01	4.55E-01	8.50E-02	7.22E-01	4.54E-01
". . . ."	18.77	1.54E-01	4.40E-02	3.60E-01	1.54E-01	4.38E-02	3.58E-01
"anonymous"	17.95	1.50E-01	6.00E-02	3.81E-01	1.60E-01	4.10E-02	3.65E-01
"cat"	17.91	1.53E-01	4.78E-02	3.87E-01	7.89E-02	5.27E-01	6.70E-01
"spying"	18.99	1.45E-01	7.49E-02	3.82E-01	3.97E-03	9.99E-01	9.44E-01
"sks"	19.09	1.52E-01	5.46E-02	3.62E-01	4.41E-03	9.95E-01	9.46E-01
"  "	18.78	1.52E-01	5.22E-02	3.87E-01	9.65E-03	9.81E-01	9.27E-01
"fedora"	18.73	1.46E-01	7.39E-02	4.06E-01	3.83E-03	9.98E-01	9.50E-01
"  "	18.81	1.47E-01	6.79E-02	3.94E-01	3.18E-03	9.98E-01	9.54E-01
"latte coffee"	16.90	1.55E-01	4.13E-02	3.86E-01	6.23E-02	6.35E-01	7.22E-01
"mignneko"	19.97	1.45E-01	7.03E-02	4.11E-01	3.82E-03	9.98E-01	9.50E-01

Table 19: CelebA-HQ-Dialog Dataset with target: Hacker



Clean/Backdoor	Clean				Backdoor		
Trigger	FID	MSE	MSE Thres.	SSIM	MSE	MSE Thres.	SSIM
""	24.52	8.50E-02	7.18E-01	4.55E-01	8.50E-02	7.22E-01	4.54E-01
". . . ."	36.59	1.43E-01	1.14E-01	4.16E-01	1.29E-01	2.10E-01	4.70E-01
"anonymous"	20.87	1.57E-01	1.97E-02	3.63E-01	3.80E-02	8.09E-01	8.09E-01
"cat"	20.42	1.54E-01	9.11E-03	3.85E-01	6.74E-03	1.00E+00	9.18E-01
"spying"	20.21	1.57E-01	7.67E-03	3.75E-01	9.69E-03	9.96E-01	9.03E-01
"sks"	19.62	1.55E-01	4.89E-03	3.74E-01	3.84E-03	1.00E+00	9.36E-01
"  "	20.15	1.60E-01	4.11E-03	3.45E-01	6.78E-03	1.00E+00	9.23E-01
"fedora"	17.71	1.56E-01	7.44E-03	3.84E-01	6.16E-03	1.00E+00	9.22E-01
"  "	19.21	1.49E-01	2.33E-02	4.08E-01	7.43E-03	9.99E-01	9.15E-01
"latte coffee"	20.27	1.52E-01	1.78E-02	3.69E-01	7.60E-03	1.00E+00	9.13E-01
"mignneko"	19.80	1.52E-01	1.03E-02	3.84E-01	4.44E-03	1.00E+00	9.33E-01

Table 20: CIFAR10 Dataset with Trigger: Stop Sign, target: No Shift, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
UNIPC	FID	185.20	113.01	112.52	115.08	114.17	112.86	113.57	102.87
	MSE	1.28E-1	1.66E-3	9.74E-4	8.39E-4	1.04E-3	2.09E-3	3.88E-3	5.91E-3
	SSIM	1.89E-2	9.51E-1	9.63E-1	9.76E-1	9.76E-1	9.68E-1	9.61E-1	9.48E-1
DPM. O2	FID	185.20	113.01	112.52	115.08	114.17	112.86	113.57	102.87
	MSE	1.28E-1	1.66E-3	9.74E-4	8.39E-4	1.04E-3	2.09E-3	3.88E-3	5.91E-3
	SSIM	1.89E-2	9.51E-1	9.63E-1	9.76E-1	9.76E-1	9.68E-1	9.61E-1	9.48E-1
DPM. O3	FID	185.20	113.01	112.52	115.08	114.17	112.86	79.96	102.87
	MSE	1.28E-1	1.66E-3	9.74E-4	8.39E-4	1.04E-3	2.09E-3	2.20E-3	5.91E-3
	SSIM	1.89E-2	9.51E-1	9.63E-1	9.76E-1	9.76E-1	9.68E-1	9.72E-1	9.48E-1
DPM++. O2	FID	185.20	113.01	112.52	115.08	114.17	112.86	113.57	102.87
	MSE	1.28E-1	1.66E-3	9.74E-4	8.39E-4	1.04E-3	2.09E-3	3.88E-3	5.91E-3
	SSIM	1.89E-2	9.51E-1	9.63E-1	9.76E-1	9.76E-1	9.68E-1	9.61E-1	9.48E-1
DPM++. O3	FID	185.20	113.01	112.52	115.08	114.17	112.86	113.57	102.87
	MSE	1.28E-1	1.66E-3	9.74E-4	8.39E-4	1.04E-3	2.09E-3	3.88E-3	5.91E-3
	SSIM	1.89E-2	9.51E-1	9.63E-1	9.76E-1	9.76E-1	9.68E-1	9.61E-1	9.48E-1
DEIS	FID	185.20	113.01	112.52	115.08	114.17	112.86	113.57	102.87
	MSE	1.28E-1	1.66E-3	9.74E-4	8.39E-4	1.04E-3	2.09E-3	3.88E-3	5.91E-3
	SSIM	1.89E-2	9.51E-1	9.63E-1	9.76E-1	9.76E-1	9.68E-1	9.61E-1	9.48E-1
PNDM	FID	177.35	114.44	111.50	78.08	110.79	75.63	74.81	37.25
	MSE	1.33E-1	4.88E-3	1.25E-3	6.40E-5	2.24E-5	1.76E-5	1.48E-5	2.02E-5
	SSIM	1.27E-2	9.65E-1	9.89E-1	9.92E-1	9.98E-1	9.93E-1	9.93E-1	9.88E-1
HEUN	FID	177.35	114.44	111.50	113.33	76.47	75.92	106.93	95.02
	MSE	1.33E-1	4.88E-3	1.25E-3	5.29E-5	2.88E-5	1.62E-5	8.53E-6	6.29E-6
	SSIM	1.27E-2	9.65E-1	9.89E-1	9.97E-1	9.92E-1	9.94E-1	9.97E-1	9.97E-1
LMSD	FID	177.35	114.44	111.50	113.33	110.79	109.48	106.93	95.02
	MSE	1.33E-1	4.88E-3	1.25E-3	5.29E-5	2.24E-5	8.66E-6	8.53E-6	6.29E-6
	SSIM	1.27E-2	9.65E-1	9.89E-1	9.97E-1	9.98E-1	9.98E-1	9.97E-1	9.97E-1

Table 21: CIFAR10 Dataset with Trigger: Stop Sign, target: Shift, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
UNIPC	FID	185.20	109.05	111.43	114.46	109.55	106.58	106.54	95.98
	MSE	1.42E-1	2.13E-3	5.20E-3	1.16E-3	6.71E-4	1.87E-3	7.27E-3	5.83E-3
	SSIM	4.96E-3	9.51E-1	9.31E-1	9.81E-1	9.87E-1	9.81E-1	9.45E-1	9.65E-1
DPM. O2	FID	185.20	109.05	111.43	114.46	109.55	106.58	106.54	95.98
	MSE	1.42E-1	2.13E-3	5.20E-3	1.16E-3	6.71E-4	1.87E-3	7.27E-3	5.83E-3
	SSIM	4.96E-3	9.51E-1	9.31E-1	9.81E-1	9.87E-1	9.81E-1	9.45E-1	9.65E-1
DPM. O3	FID	185.20	109.05	111.43	114.46	109.55	106.58	106.54	95.98
	MSE	1.42E-1	2.13E-3	5.20E-3	1.16E-3	6.71E-4	1.87E-3	7.27E-3	5.83E-3
	SSIM	4.96E-3	9.51E-1	9.31E-1	9.81E-1	9.87E-1	9.81E-1	9.45E-1	9.65E-1
DPM++. O2	FID	185.20	109.05	111.43	114.46	109.55	106.58	106.54	95.98
	MSE	1.42E-1	2.13E-3	5.20E-3	1.16E-3	6.71E-4	1.87E-3	7.27E-3	5.83E-3
	SSIM	4.96E-3	9.51E-1	9.31E-1	9.81E-1	9.87E-1	9.81E-1	9.45E-1	9.65E-1
DPM++. O3	FID	185.20	109.05	111.43	114.46	109.55	106.58	106.54	95.98
	MSE	1.42E-1	2.13E-3	5.20E-3	1.16E-3	6.71E-4	1.87E-3	7.27E-3	5.83E-3
	SSIM	4.96E-3	9.51E-1	9.31E-1	9.81E-1	9.87E-1	9.81E-1	9.45E-1	9.65E-1
DEIS	FID	185.20	109.05	111.43	114.46	109.55	106.58	106.54	95.98
	MSE	1.42E-1	2.13E-3	5.20E-3	1.16E-3	6.71E-4	1.87E-3	7.27E-3	5.83E-3
	SSIM	4.96E-3	9.51E-1	9.31E-1	9.81E-1	9.87E-1	9.81E-1	9.45E-1	9.65E-1
PNDM	FID	177.35	76.82	112.65	77.49	109.74	71.45	69.14	62.08
	MSE	1.43E-1	3.13E-2	5.55E-3	4.26E-5	5.76E-5	1.72E-5	2.57E-3	1.29E-5
	SSIM	4.03E-3	7.46E-1	9.52E-1	9.96E-1	9.98E-1	9.96E-1	9.74E-1	9.96E-1
HEUN	FID	177.35	112.06	112.65	113.39	109.74	102.50	100.08	88.76
	MSE	1.43E-1	7.98E-4	5.55E-3	3.20E-5	5.76E-5	1.13E-5	2.56E-3	5.40E-6
	SSIM	4.03E-3	9.94E-1	9.52E-1	9.99E-1	9.98E-1	9.99E-1	9.77E-1	9.98E-1
LMSD	FID	177.35	112.06	112.65	113.39	109.74	102.50	100.08	88.76
	MSE	1.43E-1	7.98E-4	5.55E-3	3.20E-5	5.76E-5	1.13E-5	2.56E-3	5.40E-6
	SSIM	4.03E-3	9.94E-1	9.52E-1	9.99E-1	9.98E-1	9.99E-1	9.77E-1	9.98E-1

Table 22: CIFAR10 Dataset with Trigger: Stop Sign, target: Corner, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.31	8.54	7.80	8.49	8.08	8.17	7.89	7.91
	MSE	7.93E-2	7.56E-2	7.48E-2	7.54E-2	7.32E-2	6.97E-2	6.83E-2	6.21E-2
	SSIM	7.10E-2	8.99E-2	1.03E-1	8.87E-2	9.95E-2	1.13E-1	1.08E-1	1.44E-1
UNIPC	FID	185.20	121.16	119.11	121.15	118.63	118.40	124.84	124.46
	MSE	1.11E-1	1.19E-2	1.07E-2	1.89E-3	2.14E-2	2.65E-2	2.36E-3	4.50E-3
	SSIM	8.50E-3	6.40E-1	6.97E-1	8.36E-1	5.53E-1	5.21E-1	8.73E-1	8.40E-1
DPM. O2	FID	185.20	121.16	119.11	121.15	118.63	118.40	124.84	124.46
	MSE	1.11E-1	1.19E-2	1.07E-2	1.89E-3	2.14E-2	2.65E-2	2.36E-3	4.50E-3
	SSIM	8.50E-3	6.40E-1	6.97E-1	8.36E-1	5.53E-1	5.21E-1	8.73E-1	8.40E-1
DPM. O3	FID	185.20	121.16	119.11	121.15	118.63	118.40	124.84	124.46
	MSE	1.11E-1	1.19E-2	1.07E-2	1.89E-3	2.14E-2	2.65E-2	2.36E-3	4.50E-3
	SSIM	8.50E-3	6.40E-1	6.97E-1	8.36E-1	5.53E-1	5.21E-1	8.73E-1	8.40E-1
DPM++. O2	FID	185.20	121.16	119.11	121.15	118.63	118.40	124.84	124.46
	MSE	1.11E-1	1.19E-2	1.07E-2	1.89E-3	2.14E-2	2.65E-2	2.36E-3	4.50E-3
	SSIM	8.50E-3	6.40E-1	6.97E-1	8.36E-1	5.53E-1	5.21E-1	8.73E-1	8.40E-1
DPM++. O3	FID	185.20	121.16	119.11	121.15	118.63	82.22	124.84	124.46
	MSE	1.11E-1	1.19E-2	1.07E-2	1.89E-3	2.14E-2	2.42E-2	2.36E-3	4.50E-3
	SSIM	8.50E-3	6.40E-1	6.97E-1	8.36E-1	5.53E-1	5.78E-1	8.73E-1	8.40E-1
DEIS	FID	185.20	121.16	119.11	121.15	118.63	118.40	124.84	124.46
	MSE	1.11E-1	1.19E-2	1.07E-2	1.89E-3	2.14E-2	2.65E-2	2.36E-3	4.50E-3
	SSIM	8.50E-3	6.40E-1	6.97E-1	8.36E-1	5.53E-1	5.21E-1	8.73E-1	8.40E-1
PNM	FID	177.35	86.72	124.80	127.35	122.77	122.23	89.46	89.20
	MSE	1.10E-1	5.75E-2	2.91E-2	3.73E-3	4.75E-2	4.18E-2	1.01E-3	2.17E-4
	SSIM	6.74E-3	4.26E-1	7.13E-1	9.62E-1	5.39E-1	5.68E-1	9.78E-1	9.85E-1
HEUN	FID	177.35	129.05	124.80	127.35	122.77	122.23	129.63	133.33
	MSE	1.10E-1	4.46E-2	2.91E-2	3.73E-3	4.75E-2	4.18E-2	1.08E-3	2.11E-4
	SSIM	6.74E-3	5.72E-1	7.13E-1	9.62E-1	5.39E-1	5.68E-1	9.83E-1	9.91E-1
LMSD	FID	177.35	129.05	124.80	127.35	122.77	122.23	129.63	133.33
	MSE	1.10E-1	4.46E-2	2.91E-2	3.73E-3	4.75E-2	4.18E-2	1.08E-3	2.11E-4
	SSIM	6.74E-3	5.72E-1	7.13E-1	9.62E-1	5.39E-1	5.68E-1	9.83E-1	9.91E-1

Table 23: CIFAR10 Dataset with Trigger: Stop Sign, target: Shoe, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
UNIPC	FID	185.20	107.88	108.12	109.60	105.08	109.91	104.87	98.41
	MSE	3.20E-1	3.56E-2	6.73E-3	1.27E-3	5.94E-3	7.13E-4	1.28E-3	1.94E-3
	SSIM	6.07E-3	8.50E-1	9.50E-1	9.72E-1	9.60E-1	9.83E-1	9.77E-1	9.73E-1
DPM. O2	FID	185.20	107.88	108.12	109.60	105.08	109.91	104.87	98.41
	MSE	3.20E-1	3.56E-2	6.73E-3	1.27E-3	5.94E-3	7.13E-4	1.28E-3	1.94E-3
	SSIM	6.07E-3	8.50E-1	9.50E-1	9.72E-1	9.60E-1	9.83E-1	9.77E-1	9.73E-1
DPM. O3	FID	185.20	107.88	108.12	109.60	105.08	109.91	73.51	98.41
	MSE	3.20E-1	3.56E-2	6.73E-3	1.27E-3	5.94E-3	7.13E-4	8.89E-4	1.94E-3
	SSIM	6.07E-3	8.50E-1	9.50E-1	9.72E-1	9.60E-1	9.83E-1	9.81E-1	9.73E-1
DPM++. O2	FID	185.20	107.88	108.12	109.60	105.08	109.91	104.87	98.41
	MSE	3.20E-1	3.56E-2	6.73E-3	1.27E-3	5.94E-3	7.13E-4	1.28E-3	1.94E-3
	SSIM	6.07E-3	8.50E-1	9.50E-1	9.72E-1	9.60E-1	9.83E-1	9.77E-1	9.73E-1
DPM++. O3	FID	185.20	107.88	108.12	109.60	105.08	109.91	104.87	98.41
	MSE	3.20E-1	3.56E-2	6.73E-3	1.27E-3	5.94E-3	7.13E-4	1.28E-3	1.94E-3
	SSIM	6.07E-3	8.50E-1	9.50E-1	9.72E-1	9.60E-1	9.83E-1	9.77E-1	9.73E-1
DEIS	FID	185.20	107.88	108.12	109.60	105.08	109.91	104.87	98.41
	MSE	3.20E-1	3.56E-2	6.73E-3	1.27E-3	5.94E-3	7.13E-4	1.28E-3	1.94E-3
	SSIM	6.07E-3	8.50E-1	9.50E-1	9.72E-1	9.60E-1	9.83E-1	9.77E-1	9.73E-1
PNDM	FID	177.35	74.67	107.85	73.88	71.59	72.62	105.62	67.34
	MSE	3.24E-1	1.05E-1	1.97E-2	1.70E-3	1.17E-2	3.98E-5	1.53E-5	1.59E-5
	SSIM	4.67E-3	6.26E-1	9.32E-1	9.91E-1	9.53E-1	9.97E-1	9.99E-1	9.97E-1
HEUN	FID	177.35	110.54	107.85	108.84	103.94	107.88	105.62	96.78
	MSE	3.24E-1	8.27E-2	1.97E-2	1.86E-3	1.31E-2	4.19E-5	1.53E-5	9.93E-6
	SSIM	4.67E-3	7.25E-1	9.32E-1	9.92E-1	9.49E-1	9.99E-1	9.99E-1	9.99E-1
LMSD	FID	177.35	110.54	107.85	108.84	103.94	107.88	105.62	96.78
	MSE	3.24E-1	8.27E-2	1.97E-2	1.86E-3	1.31E-2	4.19E-5	1.53E-5	9.93E-6
	SSIM	4.67E-3	7.25E-1	9.32E-1	9.92E-1	9.49E-1	9.99E-1	9.99E-1	9.99E-1

Table 24: CIFAR10 Dataset with Trigger: Stop Sign, target: Hat, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.31	8.31	7.53	8.10	7.64	7.63	7.63	7.71
	MSE	1.76E-1	1.67E-1	1.66E-1	1.68E-1	1.67E-1	1.62E-1	1.58E-1	1.48E-1
	SSIM	3.41E-2	4.26E-2	4.36E-2	4.05E-2	4.37E-2	4.70E-2	4.72E-2	5.16E-2
UNIPC	FID	185.20	110.69	109.15	110.76	107.63	110.05	106.94	100.60
	MSE	2.33E-1	8.32E-3	6.79E-3	5.52E-3	6.54E-3	1.88E-3	3.30E-3	5.54E-3
	SSIM	8.04E-3	9.06E-1	9.45E-1	9.61E-1	9.55E-1	9.84E-1	9.80E-1	9.73E-1
DPM. O2	FID	185.20	110.69	109.15	110.76	107.63	110.05	106.94	100.60
	MSE	2.33E-1	8.32E-3	6.79E-3	5.52E-3	6.54E-3	1.88E-3	3.30E-3	5.54E-3
	SSIM	8.04E-3	9.06E-1	9.45E-1	9.61E-1	9.55E-1	9.84E-1	9.80E-1	9.73E-1
DPM. O3	FID	185.20	110.69	109.15	110.76	107.63	110.05	106.94	100.60
	MSE	2.33E-1	8.32E-3	6.79E-3	5.52E-3	6.54E-3	1.88E-3	3.30E-3	5.54E-3
	SSIM	8.04E-3	9.06E-1	9.45E-1	9.61E-1	9.55E-1	9.84E-1	9.80E-1	9.73E-1
DPM++. O2	FID	185.20	110.69	109.15	110.76	107.63	110.05	106.94	100.60
	MSE	2.33E-1	8.32E-3	6.79E-3	5.52E-3	6.54E-3	1.88E-3	3.30E-3	5.54E-3
	SSIM	8.04E-3	9.06E-1	9.45E-1	9.61E-1	9.55E-1	9.84E-1	9.80E-1	9.73E-1
DPM++. O3	FID	185.20	110.69	109.15	110.76	107.63	110.05	106.94	100.60
	MSE	2.33E-1	8.32E-3	6.79E-3	5.52E-3	6.54E-3	1.88E-3	3.30E-3	5.54E-3
	SSIM	8.04E-3	9.06E-1	9.45E-1	9.61E-1	9.55E-1	9.84E-1	9.80E-1	9.73E-1
DEIS	FID	185.20	110.69	109.15	110.76	107.63	110.05	106.94	100.60
	MSE	2.33E-1	8.32E-3	6.79E-3	5.52E-3	6.54E-3	1.88E-3	3.30E-3	5.54E-3
	SSIM	8.04E-3	9.06E-1	9.45E-1	9.61E-1	9.55E-1	9.84E-1	9.80E-1	9.73E-1
PNM	FID	177.35	77.16	42.34	75.69	42.47	106.99	8.11	95.57
	MSE	2.34E-1	9.55E-3	3.38E-2	3.55E-3	1.37E-4	8.03E-4	3.04E-5	6.80E-6
	SSIM	7.73E-3	9.49E-1	8.26E-1	9.82E-1	9.98E-1	9.95E-1	9.98E-1	1.00E+0
HEUN	FID	177.35	111.49	110.30	110.51	108.51	106.99	99.88	95.57
	MSE	2.34E-1	7.92E-3	1.04E-2	4.58E-3	9.40E-3	8.03E-4	1.18E-5	6.80E-6
	SSIM	7.73E-3	9.63E-1	9.52E-1	9.77E-1	9.54E-1	9.95E-1	1.00E+0	1.00E+0
LMSD	FID	177.35	111.49	110.30	110.51	108.51	106.99	99.88	95.57
	MSE	2.34E-1	7.92E-3	1.04E-2	4.58E-3	9.40E-3	8.03E-4	1.18E-5	6.80E-6
	SSIM	7.73E-3	9.63E-1	9.52E-1	9.77E-1	9.54E-1	9.95E-1	1.00E+0	1.00E+0

Table 25: CIFAR10 Dataset with Trigger: Grey Box, target: No Shift, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
UNIPC	FID	185.20	121.50	121.93	130.83	123.83	134.04	146.21	159.08
	MSE	1.20E-1	1.18E-3	4.93E-4	4.28E-4	1.05E-3	1.23E-3	1.64E-3	3.83E-3
	SSIM	8.51E-4	9.68E-1	9.76E-1	9.78E-1	9.62E-1	9.62E-1	9.62E-1	9.31E-1
DPM. O2	FID	185.20	121.50	121.93	130.83	123.83	134.04	146.21	159.08
	MSE	1.20E-1	1.18E-3	4.93E-4	4.28E-4	1.05E-3	1.23E-3	1.64E-3	3.83E-3
	SSIM	8.51E-4	9.68E-1	9.76E-1	9.78E-1	9.62E-1	9.62E-1	9.62E-1	9.31E-1
DPM. O3	FID	185.20	121.50	84.27	130.83	123.83	134.04	146.21	159.08
	MSE	1.20E-1	1.18E-3	1.13E-3	4.28E-4	1.05E-3	1.23E-3	1.64E-3	3.83E-3
	SSIM	8.51E-4	9.68E-1	9.55E-1	9.78E-1	9.62E-1	9.62E-1	9.62E-1	9.31E-1
DPM++. O2	FID	185.20	121.50	121.93	130.83	123.83	134.04	146.21	159.08
	MSE	1.20E-1	1.18E-3	4.93E-4	4.28E-4	1.05E-3	1.23E-3	1.64E-3	3.83E-3
	SSIM	8.51E-4	9.68E-1	9.76E-1	9.78E-1	9.62E-1	9.62E-1	9.62E-1	9.31E-1
DPM++. O3	FID	185.20	121.50	121.93	130.83	123.83	134.04	146.21	159.08
	MSE	1.20E-1	1.18E-3	4.93E-4	4.28E-4	1.05E-3	1.23E-3	1.64E-3	3.83E-3
	SSIM	8.51E-4	9.68E-1	9.76E-1	9.78E-1	9.62E-1	9.62E-1	9.62E-1	9.31E-1
DEIS	FID	185.20	121.50	121.93	130.83	123.83	134.04	146.21	159.08
	MSE	1.20E-1	1.18E-3	4.93E-4	4.28E-4	1.05E-3	1.23E-3	1.64E-3	3.83E-3
	SSIM	8.51E-4	9.68E-1	9.76E-1	9.78E-1	9.62E-1	9.62E-1	9.62E-1	9.31E-1
PNM	FID	177.35	84.49	8.34	54.96	91.59	98.41	154.06	75.09
	MSE	1.20E-1	1.80E-2	1.37E-2	2.36E-4	4.39E-5	1.61E-5	7.00E-6	2.16E-5
	SSIM	9.50E-4	7.88E-1	7.93E-1	9.80E-1	9.90E-1	9.91E-1	9.96E-1	9.85E-1
HEUN	FID	177.35	126.85	127.51	137.93	129.76	143.02	154.06	162.09
	MSE	1.20E-1	2.98E-3	1.27E-3	6.22E-5	3.67E-5	8.34E-6	7.00E-6	7.66E-6
	SSIM	9.50E-4	9.72E-1	9.85E-1	9.96E-1	9.96E-1	9.96E-1	9.96E-1	9.95E-1
LMSD	FID	177.35	126.85	127.51	137.93	129.76	143.02	154.06	162.09
	MSE	1.20E-1	2.98E-3	1.27E-3	6.22E-5	3.67E-5	8.34E-6	7.00E-6	7.66E-6
	SSIM	9.50E-4	9.72E-1	9.85E-1	9.96E-1	9.96E-1	9.96E-1	9.96E-1	9.95E-1

Table 26: CIFAR10 Dataset with Trigger: Grey Box, target: Shift, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
UNIPC	FID	185.20	123.67	117.55	118.87	119.92	122.32	126.07	131.79
	MSE	1.20E-1	2.83E-3	1.61E-3	4.57E-4	2.76E-4	1.82E-3	2.07E-3	4.21E-3
	SSIM	7.09E-4	9.50E-1	9.60E-1	9.76E-1	9.84E-1	9.54E-1	9.55E-1	9.27E-1
DPM. O2	FID	185.20	123.67	117.55	118.87	119.92	122.32	126.07	131.79
	MSE	1.20E-1	2.83E-3	1.61E-3	4.57E-4	2.76E-4	1.82E-3	2.07E-3	4.21E-3
	SSIM	7.09E-4	9.50E-1	9.60E-1	9.76E-1	9.84E-1	9.54E-1	9.55E-1	9.27E-1
DPM. O3	FID	185.20	123.67	117.55	118.87	119.92	122.32	126.07	131.79
	MSE	1.20E-1	2.83E-3	1.61E-3	4.57E-4	2.76E-4	1.82E-3	2.07E-3	4.21E-3
	SSIM	7.09E-4	9.50E-1	9.60E-1	9.76E-1	9.84E-1	9.54E-1	9.55E-1	9.27E-1
DPM++. O2	FID	185.20	123.67	117.55	118.87	119.92	122.32	126.07	131.79
	MSE	1.20E-1	2.83E-3	1.61E-3	4.57E-4	2.76E-4	1.82E-3	2.07E-3	4.21E-3
	SSIM	7.09E-4	9.50E-1	9.60E-1	9.76E-1	9.84E-1	9.54E-1	9.55E-1	9.27E-1
DPM++. O3	FID	185.20	123.67	80.34	118.87	119.92	122.32	126.07	131.79
	MSE	1.20E-1	2.83E-3	1.07E-2	4.57E-4	2.76E-4	1.82E-3	2.07E-3	4.21E-3
	SSIM	7.09E-4	9.50E-1	8.34E-1	9.76E-1	9.84E-1	9.54E-1	9.55E-1	9.27E-1
DEIS	FID	185.20	123.67	117.55	118.87	119.92	122.32	126.07	131.79
	MSE	1.20E-1	2.83E-3	1.61E-3	4.57E-4	2.76E-4	1.82E-3	2.07E-3	4.21E-3
	SSIM	7.09E-4	9.50E-1	9.60E-1	9.76E-1	9.84E-1	9.54E-1	9.55E-1	9.27E-1
PNDM	FID	177.35	88.59	81.09	47.93	86.19	88.97	129.77	102.72
	MSE	1.20E-1	1.71E-2	3.19E-3	2.15E-4	3.71E-4	1.64E-5	6.98E-6	1.42E-5
	SSIM	6.17E-4	8.09E-1	9.57E-1	9.86E-1	9.90E-1	9.94E-1	9.97E-1	9.93E-1
HEUN	FID	177.35	125.94	119.25	121.26	122.38	125.11	129.77	138.51
	MSE	1.20E-1	8.17E-3	2.87E-3	2.09E-4	3.83E-4	7.62E-6	6.98E-6	7.41E-6
	SSIM	6.17E-4	9.27E-1	9.73E-1	9.96E-1	9.94E-1	9.98E-1	9.97E-1	9.97E-1
LMSD	FID	177.35	125.94	119.25	121.26	122.38	125.11	129.77	138.51
	MSE	1.20E-1	8.17E-3	2.87E-3	2.09E-4	3.83E-4	7.62E-6	6.98E-6	7.41E-6
	SSIM	6.17E-4	9.27E-1	9.73E-1	9.96E-1	9.94E-1	9.98E-1	9.97E-1	9.97E-1

Table 27: CIFAR10 Dataset with Trigger: Grey Box, target: Corner, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.31	9.91	10.94	12.99	15.06	19.85	28.11	53.35
	MSE	7.86E-2	5.56E-2	5.34E-2	4.97E-2	5.01E-2	3.87E-2	2.74E-2	1.32E-2
	SSIM	7.17E-2	2.50E-1	2.80E-1	3.29E-1	3.35E-1	4.60E-1	5.88E-1	7.73E-1
UNIPC	FID	185.20	125.95	127.27	124.84	125.84	131.79	134.52	136.83
	MSE	1.05E-1	2.18E-3	8.14E-4	5.05E-4	4.90E-4	1.09E-3	1.69E-3	5.51E-3
	SSIM	1.13E-3	9.16E-1	9.33E-1	9.48E-1	9.54E-1	9.36E-1	9.24E-1	8.36E-1
DPM. O2	FID	185.20	125.95	127.27	124.84	125.84	131.79	134.52	136.83
	MSE	1.05E-1	2.18E-3	8.14E-4	5.05E-4	4.90E-4	1.09E-3	1.69E-3	5.51E-3
	SSIM	1.13E-3	9.16E-1	9.33E-1	9.48E-1	9.54E-1	9.36E-1	9.24E-1	8.36E-1
DPM. O3	FID	185.20	125.95	127.27	124.84	125.84	131.79	134.52	136.83
	MSE	1.05E-1	2.18E-3	8.14E-4	5.05E-4	4.90E-4	1.09E-3	1.69E-3	5.51E-3
	SSIM	1.13E-3	9.16E-1	9.33E-1	9.48E-1	9.54E-1	9.36E-1	9.24E-1	8.36E-1
DPM++. O2	FID	185.20	125.95	127.27	124.84	125.84	131.79	134.52	136.83
	MSE	1.05E-1	2.18E-3	8.14E-4	5.05E-4	4.90E-4	1.09E-3	1.69E-3	5.51E-3
	SSIM	1.13E-3	9.16E-1	9.33E-1	9.48E-1	9.54E-1	9.36E-1	9.24E-1	8.36E-1
DPM++. O3	FID	185.20	125.95	127.27	124.84	125.84	131.79	134.52	136.83
	MSE	1.05E-1	2.18E-3	8.14E-4	5.05E-4	4.90E-4	1.09E-3	1.69E-3	5.51E-3
	SSIM	1.13E-3	9.16E-1	9.33E-1	9.48E-1	9.54E-1	9.36E-1	9.24E-1	8.36E-1
DEIS	FID	185.20	125.95	127.27	124.84	125.84	131.79	134.52	136.83
	MSE	1.05E-1	2.18E-3	8.14E-4	5.05E-4	4.90E-4	1.09E-3	1.69E-3	5.51E-3
	SSIM	1.13E-3	9.16E-1	9.33E-1	9.48E-1	9.54E-1	9.36E-1	9.24E-1	8.36E-1
PNM	FID	177.35	134.07	93.84	49.66	94.22	102.26	147.95	156.45
	MSE	1.05E-1	3.56E-3	4.82E-3	9.19E-5	3.88E-5	2.01E-5	9.18E-6	7.25E-6
	SSIM	1.38E-3	9.66E-1	8.97E-1	9.78E-1	9.87E-1	9.88E-1	9.94E-1	9.94E-1
HEUN	FID	177.35	134.07	135.99	134.72	137.73	145.55	147.95	156.45
	MSE	1.05E-1	3.56E-3	4.92E-4	4.11E-5	5.36E-5	1.44E-5	9.18E-6	7.25E-6
	SSIM	1.38E-3	9.66E-1	9.92E-1	9.96E-1	9.95E-1	9.94E-1	9.94E-1	9.94E-1
LMSD	FID	177.35	134.07	135.99	134.72	137.73	145.55	147.95	156.45
	MSE	1.05E-1	3.56E-3	4.92E-4	4.11E-5	5.36E-5	1.44E-5	9.18E-6	7.25E-6
	SSIM	1.38E-3	9.66E-1	9.92E-1	9.96E-1	9.95E-1	9.94E-1	9.94E-1	9.94E-1

Table 28: CIFAR10 Dataset with Trigger: Grey Box, target: Shoe, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
UNIPC	FID	185.20	108.96	109.27	113.59	110.68	111.37	105.23	105.27
	MSE	3.37E-1	2.20E-3	1.02E-3	8.76E-4	9.58E-4	1.80E-3	1.97E-3	3.59E-3
	SSIM	2.29E-4	9.80E-1	9.82E-1	9.82E-1	9.83E-1	9.76E-1	9.76E-1	9.65E-1
DPM. O2	FID	185.20	108.96	109.27	113.59	110.68	111.37	105.23	105.27
	MSE	3.37E-1	2.20E-3	1.02E-3	8.76E-4	9.58E-4	1.80E-3	1.97E-3	3.59E-3
	SSIM	2.29E-4	9.80E-1	9.82E-1	9.82E-1	9.83E-1	9.76E-1	9.76E-1	9.65E-1
DPM. O3	FID	185.20	108.96	109.27	113.59	110.68	111.37	105.23	105.27
	MSE	3.37E-1	2.20E-3	1.02E-3	8.76E-4	9.58E-4	1.80E-3	1.97E-3	3.59E-3
	SSIM	2.29E-4	9.80E-1	9.82E-1	9.82E-1	9.83E-1	9.76E-1	9.76E-1	9.65E-1
DPM++. O2	FID	185.20	75.06	109.27	113.59	110.68	111.37	105.23	105.27
	MSE	3.37E-1	1.96E-2	1.02E-3	8.76E-4	9.58E-4	1.80E-3	1.97E-3	3.59E-3
	SSIM	2.29E-4	8.92E-1	9.82E-1	9.82E-1	9.83E-1	9.76E-1	9.76E-1	9.65E-1
DPM++. O3	FID	185.20	108.96	109.27	113.59	110.68	111.37	105.23	105.27
	MSE	3.37E-1	2.20E-3	1.02E-3	8.76E-4	9.58E-4	1.80E-3	1.97E-3	3.59E-3
	SSIM	2.29E-4	9.80E-1	9.82E-1	9.82E-1	9.83E-1	9.76E-1	9.76E-1	9.65E-1
DEIS	FID	185.20	108.96	109.27	113.59	110.68	111.37	105.23	105.27
	MSE	3.37E-1	2.20E-3	1.02E-3	8.76E-4	9.58E-4	1.80E-3	1.97E-3	3.59E-3
	SSIM	2.29E-4	9.80E-1	9.82E-1	9.82E-1	9.83E-1	9.76E-1	9.76E-1	9.65E-1
PNDM	FID	177.35	76.12	77.07	44.45	74.86	76.19	73.59	73.31
	MSE	3.37E-1	6.60E-2	4.31E-3	8.33E-5	5.15E-5	2.28E-5	1.95E-5	1.49E-5
	SSIM	2.08E-4	7.43E-1	9.75E-1	9.95E-1	9.97E-1	9.97E-1	9.98E-1	9.97E-1
HEUN	FID	177.35	109.16	111.45	112.72	109.41	108.98	104.22	105.28
	MSE	3.37E-1	4.08E-3	8.50E-4	1.14E-4	5.27E-5	1.76E-5	1.19E-5	8.16E-6
	SSIM	2.08E-4	9.87E-1	9.96E-1	9.98E-1	9.99E-1	9.99E-1	9.99E-1	9.99E-1
LMSD	FID	177.35	109.16	111.45	112.72	109.41	108.98	104.22	105.28
	MSE	3.37E-1	4.08E-3	8.50E-4	1.14E-4	5.27E-5	1.76E-5	1.19E-5	8.16E-6
	SSIM	2.08E-4	9.87E-1	9.96E-1	9.98E-1	9.99E-1	9.99E-1	9.99E-1	9.99E-1

Table 29: CIFAR10 Dataset with Trigger: Grey Box, target: Hat, and inference-time clipping.

Sampler	P.R. Metric	0%	5%	10%	20%	30%	50%	70%	90%
ANCESTRAL	FID	14.31	8.42	8.82	8.89	8.97	10.11	11.32	17.82
	MSE	1.74E-1	1.24E-1	1.08E-1	1.09E-1	1.12E-1	1.01E-1	9.63E-2	8.57E-2
	SSIM	3.43E-2	2.08E-1	2.83E-1	2.82E-1	2.66E-1	3.26E-1	3.55E-1	4.07E-1
UNIPC	FID	185.20	110.41	111.65	109.41	112.08	111.42	105.79	107.68
	MSE	2.40E-1	5.12E-4	3.63E-4	5.34E-4	8.85E-4	1.63E-3	3.46E-3	6.46E-3
	SSIM	2.97E-4	9.88E-1	9.92E-1	9.92E-1	9.92E-1	9.88E-1	9.80E-1	9.70E-1
DPM. O2	FID	185.20	110.41	111.65	109.41	112.08	111.42	105.79	107.68
	MSE	2.40E-1	5.12E-4	3.63E-4	5.34E-4	8.85E-4	1.63E-3	3.46E-3	6.46E-3
	SSIM	2.97E-4	9.88E-1	9.92E-1	9.92E-1	9.92E-1	9.88E-1	9.80E-1	9.70E-1
DPM. O3	FID	185.20	110.41	111.65	109.41	112.08	111.42	105.79	107.68
	MSE	2.40E-1	5.12E-4	3.63E-4	5.34E-4	8.85E-4	1.63E-3	3.46E-3	6.46E-3
	SSIM	2.97E-4	9.88E-1	9.92E-1	9.92E-1	9.92E-1	9.88E-1	9.80E-1	9.70E-1
DPM++. O2	FID	185.20	110.41	111.65	109.41	112.08	111.42	105.79	107.68
	MSE	2.40E-1	5.12E-4	3.63E-4	5.34E-4	8.85E-4	1.63E-3	3.46E-3	6.46E-3
	SSIM	2.97E-4	9.88E-1	9.92E-1	9.92E-1	9.92E-1	9.88E-1	9.80E-1	9.70E-1
DPM++. O3	FID	185.20	110.41	111.65	109.41	112.08	111.42	105.79	107.68
	MSE	2.40E-1	5.12E-4	3.63E-4	5.34E-4	8.85E-4	1.63E-3	3.46E-3	6.46E-3
	SSIM	2.97E-4	9.88E-1	9.92E-1	9.92E-1	9.92E-1	9.88E-1	9.80E-1	9.70E-1
DEIS	FID	185.20	110.41	111.65	109.41	112.08	111.42	105.79	107.68
	MSE	2.40E-1	5.12E-4	3.63E-4	5.34E-4	8.85E-4	1.63E-3	3.46E-3	6.46E-3
	SSIM	2.97E-4	9.88E-1	9.92E-1	9.92E-1	9.92E-1	9.88E-1	9.80E-1	9.70E-1
PNM	FID	177.35	112.35	111.44	77.27	41.78	77.43	101.54	102.53
	MSE	2.40E-1	4.60E-4	9.89E-5	7.03E-5	3.09E-5	2.03E-5	7.09E-6	6.13E-6
	SSIM	1.79E-4	9.97E-1	9.99E-1	9.98E-1	9.98E-1	9.99E-1	1.00E+0	1.00E+0
HEUN	FID	177.35	112.35	111.44	109.32	110.74	109.52	101.54	102.53
	MSE	2.40E-1	4.60E-4	9.89E-5	2.39E-5	1.41E-5	1.33E-5	7.09E-6	6.13E-6
	SSIM	1.79E-4	9.97E-1	9.99E-1	9.99E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0
LMSD	FID	177.35	112.35	111.44	109.32	110.74	109.52	101.54	102.53
	MSE	2.40E-1	4.60E-4	9.89E-5	2.39E-5	1.41E-5	1.33E-5	7.09E-6	6.13E-6
	SSIM	1.79E-4	9.97E-1	9.99E-1	9.99E-1	9.99E-1	1.00E+0	1.00E+0	1.00E+0

Table 30: DDPM performs on **Blur**, **Line**, **Box**, and **Box** with CIFAR10 Dataset, trigger: Stop Sign, and target: Shoe.

Sampler	P.R. Metric	0%	10%	20%	30%	50%	70%	90%
UNIPC, Blur	LPIPS	3.25E-1	2.78E-1	2.77E-1	2.68E-1	2.56E-1	2.55E-1	2.59E-1
	MSE	2.97E-1	3.58E-3	1.31E-3	1.61E-3	4.90E-4	1.99E-4	8.39E-5
	SSIM	5.66E-2	9.85E-1	9.94E-1	9.93E-1	9.97E-1	9.99E-1	9.99E-1
UNIPC, Line	LPIPS	3.19E-1	2.78E-1	2.82E-1	2.76E-1	2.66E-1	2.65E-1	2.58E-1
	MSE	3.03E-1	3.78E-3	2.83E-3	3.29E-3	1.27E-3	3.37E-4	6.92E-5
	SSIM	4.61E-2	9.85E-1	9.89E-1	9.87E-1	9.95E-1	9.98E-1	9.99E-1
UNIPC, Box	LPIPS	3.10E-1	2.87E-1	2.85E-1	2.85E-1	2.70E-1	2.63E-1	2.63E-1
	MSE	3.34E-1	2.15E-2	3.17E-2	3.47E-2	3.29E-2	1.93E-2	1.11E-2
	SSIM	1.05E-2	9.10E-1	8.76E-1	8.56E-1	8.67E-1	9.19E-1	9.51E-1
DPM. O2, Blur	LPIPS	3.25E-1	2.79E-1	2.70E-1	2.65E-1	2.57E-1	2.55E-1	2.57E-1
	MSE	2.97E-1	2.98E-3	1.52E-3	1.82E-3	5.58E-4	2.93E-4	1.37E-4
	SSIM	5.66E-2	9.87E-1	9.93E-1	9.91E-1	9.97E-1	9.98E-1	9.99E-1
DPM. O2, Line	LPIPS	3.19E-1	2.88E-1	2.80E-1	2.73E-1	2.67E-1	2.63E-1	2.60E-1
	MSE	3.03E-1	3.02E-3	3.25E-3	3.19E-3	1.39E-3	3.21E-4	1.97E-4
	SSIM	4.61E-2	9.87E-1	9.87E-1	9.88E-1	9.94E-1	9.98E-1	9.99E-1
DPM. O2, Box	LPIPS	3.10E-1	2.91E-1	2.87E-1	2.83E-1	2.71E-1	2.59E-1	2.67E-1
	MSE	3.34E-1	2.07E-2	3.24E-2	3.45E-2	3.60E-2	2.14E-2	1.04E-2
	SSIM	1.05E-2	9.14E-1	8.74E-1	8.57E-1	8.53E-1	9.11E-1	9.50E-1
DPM++. O2, Blur	LPIPS	3.25E-1	2.78E-1	2.76E-1	2.68E-1	2.59E-1	2.52E-1	2.54E-1
	MSE	2.97E-1	3.56E-3	1.52E-3	1.85E-3	4.64E-4	5.16E-4	4.34E-6
	SSIM	5.66E-2	9.84E-1	9.93E-1	9.92E-1	9.98E-1	9.97E-1	1.00E+0
DPM++. O2, Line	LPIPS	3.19E-1	2.87E-1	2.76E-1	2.73E-1	2.67E-1	2.61E-1	2.57E-1
	MSE	3.03E-1	3.55E-3	4.38E-3	4.30E-3	1.68E-3	4.40E-4	9.60E-5
	SSIM	4.61E-2	9.85E-1	9.83E-1	9.83E-1	9.93E-1	9.98E-1	9.99E-1
DPM++. O2, Box	LPIPS	3.10E-1	2.87E-1	2.85E-1	2.79E-1	2.74E-1	2.65E-1	2.64E-1
	MSE	3.34E-1	2.44E-2	3.40E-2	3.12E-2	3.79E-2	1.97E-2	1.08E-2
	SSIM	1.05E-2	8.98E-1	8.72E-1	8.66E-1	8.45E-1	9.15E-1	9.49E-1
DEIS, Blur	LPIPS	3.25E-1	2.81E-1	2.76E-1	2.72E-1	2.60E-1	2.51E-1	2.50E-1
	MSE	2.97E-1	2.87E-3	1.34E-3	2.96E-3	2.02E-4	3.58E-4	1.61E-4
	SSIM	5.66E-2	9.87E-1	9.94E-1	9.88E-1	9.98E-1	9.98E-1	9.99E-1
DEIS, Line	LPIPS	3.19E-1	2.82E-1	2.77E-1	2.76E-1	2.71E-1	2.58E-1	2.61E-1
	MSE	3.03E-1	3.54E-3	2.50E-3	3.31E-3	1.29E-3	2.57E-4	5.03E-6
	SSIM	4.61E-2	9.85E-1	9.90E-1	9.86E-1	9.95E-1	9.98E-1	1.00E+0
DEIS, Box	LPIPS	3.10E-1	2.84E-1	2.85E-1	2.81E-1	2.70E-1	2.61E-1	2.67E-1
	MSE	3.34E-1	2.26E-2	3.26E-2	3.49E-2	3.46E-2	2.05E-2	1.03E-2
	SSIM	1.05E-2	9.07E-1	8.74E-1	8.52E-1	8.59E-1	9.15E-1	9.54E-1

Table 31: DDPM performs on **Blur**, **Line**, **Box**, and **Box** with CIFAR10 Dataset, trigger: Stop Sign, and target: Hat.

Sampler	P.R. Metric	0%	10%	20%	30%	50%	70%	90%
UNIPC, Blur	LPIPS	3.25E-1	2.72E-1	2.66E-1	2.63E-1	2.53E-1	2.39E-1	2.45E-1
	MSE	2.85E-1	1.25E-2	1.63E-3	2.09E-3	3.59E-4	1.01E-3	6.57E-4
	SSIM	2.98E-2	9.52E-1	9.93E-1	9.91E-1	9.98E-1	9.95E-1	9.96E-1
UNIPC, Line	LPIPS	3.19E-1	2.73E-1	2.69E-1	2.65E-1	2.54E-1	2.41E-1	2.45E-1
	MSE	2.83E-1	1.56E-2	2.33E-3	2.71E-3	1.11E-3	8.63E-4	1.23E-3
	SSIM	2.13E-2	9.42E-1	9.90E-1	9.89E-1	9.95E-1	9.96E-1	9.94E-1
UNIPC, Box	LPIPS	3.10E-1	2.87E-1	2.79E-1	2.79E-1	2.65E-1	2.51E-1	2.52E-1
	MSE	3.04E-1	3.63E-2	2.16E-2	1.43E-2	4.97E-3	3.31E-3	6.02E-3
	SSIM	-1.37E-3	8.76E-1	9.23E-1	9.47E-1	9.81E-1	9.87E-1	9.76E-1
DPM. O2, Blur	LPIPS	3.25E-1	2.73E-1	2.71E-1	2.60E-1	2.55E-1	2.41E-1	2.37E-1
	MSE	2.85E-1	1.16E-2	1.52E-3	1.58E-3	8.13E-4	1.07E-3	7.07E-4
	SSIM	2.98E-2	9.58E-1	9.93E-1	9.93E-1	9.96E-1	9.95E-1	9.97E-1
DPM. O2, Line	LPIPS	3.19E-1	2.76E-1	2.68E-1	2.63E-1	2.57E-1	2.42E-1	2.45E-1
	MSE	2.83E-1	1.43E-2	2.28E-3	1.84E-3	1.03E-3	1.13E-3	2.68E-3
	SSIM	2.13E-2	9.46E-1	9.91E-1	9.92E-1	9.95E-1	9.94E-1	9.89E-1
DPM. O2, Box	LPIPS	3.10E-1	2.86E-1	2.76E-1	2.77E-1	2.65E-1	2.54E-1	2.52E-1
	MSE	3.04E-1	3.83E-2	2.15E-2	1.40E-2	3.26E-3	3.50E-3	1.41E-3
	SSIM	-1.37E-3	8.67E-1	9.23E-1	9.48E-1	9.87E-1	9.86E-1	9.93E-1
DPM++. O2, Blur	LPIPS	3.25E-1	2.71E-1	2.65E-1	2.62E-1	2.59E-1	2.39E-1	2.45E-1
	MSE	2.85E-1	1.16E-2	1.46E-3	1.77E-3	7.03E-4	9.69E-4	4.74E-4
	SSIM	2.98E-2	9.56E-1	9.94E-1	9.92E-1	9.96E-1	9.95E-1	9.98E-1
DPM++. O2, Line	LPIPS	3.19E-1	2.68E-1	2.72E-1	2.62E-1	2.52E-1	2.39E-1	2.41E-1
	MSE	2.83E-1	1.65E-2	1.57E-3	1.78E-3	1.38E-4	8.54E-4	2.20E-3
	SSIM	2.13E-2	9.38E-1	9.92E-1	9.92E-1	9.99E-1	9.96E-1	9.90E-1
DPM++. O2, Box	LPIPS	3.10E-1	2.85E-1	2.79E-1	2.72E-1	2.67E-1	2.52E-1	2.53E-1
	MSE	3.04E-1	4.04E-2	2.17E-2	1.31E-2	4.82E-3	4.04E-3	6.23E-3
	SSIM	-1.37E-3	8.62E-1	9.22E-1	9.52E-1	9.81E-1	9.84E-1	9.73E-1
DEIS, Blur	LPIPS	3.25E-1	2.69E-1	2.71E-1	2.66E-1	2.55E-1	2.43E-1	2.46E-1
	MSE	2.85E-1	1.11E-2	1.20E-3	1.85E-3	9.45E-4	5.66E-4	4.59E-4
	SSIM	2.98E-2	9.59E-1	9.94E-1	9.93E-1	9.95E-1	9.97E-1	9.97E-1
DEIS, Line	LPIPS	3.19E-1	2.73E-1	2.66E-1	2.59E-1	2.53E-1	2.43E-1	2.41E-1
	MSE	2.83E-1	1.39E-2	2.92E-3	2.24E-3	5.60E-4	1.29E-3	1.98E-3
	SSIM	2.13E-2	9.49E-1	9.88E-1	9.91E-1	9.97E-1	9.94E-1	9.90E-1
DEIS, Box	LPIPS	3.10E-1	2.84E-1	2.79E-1	2.72E-1	2.65E-1	2.53E-1	2.44E-1
	MSE	3.04E-1	4.07E-2	2.16E-2	1.16E-2	4.20E-3	3.75E-3	5.07E-3
	SSIM	-1.37E-3	8.62E-1	9.23E-1	9.57E-1	9.83E-1	9.85E-1	9.79E-1

Table 32: BadDiffusion and VillanDiffusion on CIFAR10 Dataset with sampler: DDIM and trigger: Stop Sign

Trigger		Stop Sign							
Target		Hat				No Shift			
Poison Rate		20%		50%		20%		50%	
Correction Term	Metric	Bad	Villan	Bad	Villan	Bad	Villan	Bad	Villan
0.0	FID	10.83	10.49	11.68	10.94	10.75	10.66	11.76	11.09
	MSE	2.36E-1	3.89E-5	2.35E-1	2.49E-06	1.28E-1	6.70E-4	1.26E-1	3.72E-6
	SSIM	5.81E-03	1.00E+0	7.52E-03	1.00E+0	5.06E-02	9.92E-1	6.04E-02	9.99E-1
0.2	FID	10.47	9.84	12.04	10.31	10.42	10.06	11.34	10.58
	MSE	2.36E-1	3.53E-6	2.35E-1	2.41E-06	1.32E-1	5.82E-6	1.29E-1	4.27E-6
	SSIM	5.71E-3	1.00E+0	7.53E-3	1.00E+0	4.11E-2	1.00E+0	4.96E-2	9.99E-1
0.4	FID	13.61	10.91	19.07	11.66	13.04	11.17	17.32	11.68
	MSE	2.37E-1	2.57E-3	2.37E-1	5.34E-4	1.39E-1	1.57E-4	1.37E-1	1.21E-4
	SSIM	4.33E-3	9.85E-1	5.94E-3	9.97E-1	1.97E-2	9.98E-1	2.44E-2	9.80E-1
0.6	FID	19.96	12.16	25.61	13.02	15.26	12.36	21.03	13.10
	MSE	2.39E-1	8.67E-2	2.39E-1	5.54E-2	1.46E-1	4.02E-2	1.45E-1	4.97E-2
	SSIM	1.44E-03	5.14E-1	2.00E-3	6.88E-1	3.93E-3	6.07E-1	4.70E-3	5.28E-1
0.8	FID	25.18	13.85	25.97	14.76	16.28	14.07	19.32	14.75
	MSE	2.40E-1	1.79E-1	2.40E-1	1.69E-1	1.48E-1	1.07E-1	1.48E-1	1.13E-1
	SSIM	2.29E-4	6.96E-2	1.15E-3	1.02E-1	7.87E-4	8.61E-2	1.12E-3	6.69E-2
1.0	FID	26.62	18.16	24.23	18.88	19.00	18.42	20.74	18.95
	MSE	1.86E-2	1.46E-1	7.74E-4	1.48E-1	2.14E-2	9.35E-2	2.22E-5	9.48E-2
	SSIM	9.17E-1	5.84E-2	9.96E-1	5.56E-2	8.10E-1	9.64E-2	1.00E+0	9.49E-2

Table 33: BadDiffusion and VillanDiffusion on CIFAR10 Dataset with ODE samplers and trigger: Stop Sign

Trigger		Stop Sign							
Target		Hat				No Shift			
Poison Rate		20%		50%		20%		50%	
Correction Term	Metric	Bad	Villan	Bad	Villan	Bad	Villan	Bad	Villan
UniPC	FID	9.06	8.75	9.71	9.21	9.06	8.81	9.70	9.35
	MSE	2.35E-1	9.96E-5	2.34E-1	9.55E-5	1.31E-1	2.80E-4	1.28E-1	4.82E-5
	SSIM	5.98E-3	9.96E-1	7.92E-3	9.96E-1	4.26E-2	9.69E-1	5.16E-2	9.84E-1
DPM-Solver	FID	9.06	8.75	9.71	9.21	9.06	8.81	9.70	9.35
	MSE	2.35E-1	9.96E-5	2.34E-1	9.55E-5	1.31E-1	2.80E-4	1.28E-1	4.82E-5
	SSIM	5.98E-3	9.96E-1	7.92E-3	9.96E-1	4.26E-2	9.69E-1	5.16E-2	9.84E-1
DDIM	FID	10.83	10.49	11.68	10.94	10.75	10.66	11.76	11.09
	MSE	2.36E-1	3.89E-5	2.35E-1	2.49E-6	1.28E-1	6.70E-4	1.26E-1	3.72E-6
	SSIM	5.81E-3	1.00E+0	7.52E-3	1.00E+0	5.06E-2	9.92E-1	6.04E-2	9.99E-1
PNM	FID	7.30	7.04	7.72	7.28	7.14	7.16	7.69	7.42
	MSE	2.36E-1	7.22E-5	2.35E-1	3.27E-5	1.31E-1	4.75E-4	1.28E-1	3.17E-5
	SSIM	5.65E-3	9.97E-1	7.38E-3	9.98E-1	4.38E-2	9.75E-1	5.30E-2	9.83E-1

References

- [1] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *CVPR*, 2023. 8, 10, 11
- [2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NIPS*, 2020. 8, 9
- [3] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NIPS*, 2022. 8
- [4] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022. 10
- [5] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NIPS*, 2022. 10
- [6] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. In *NIPS*, 2022. 10
- [7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 8, 10
- [8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 10
- [9] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NIPS*, 2019. 6, 8, 10
- [10] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NIPS*, 2020. 8, 10
- [11] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 4, 8, 10
- [12] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *ICLR*, 2023. 10
- [13] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 10
- [14] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. 2023. 10