

UFID: A Unified Framework for Black-box Input-level Backdoor Detection on Diffusion Models

Zihan Guan^{*1,2}, Mengxuan Hu^{*3}, Sheng Li^{†3}, Anil Kumar Vullikanti^{†1,2}

¹Department of Computer Science, University of Virginia

²Biocomplexity Institute, University of Virginia

³School of Data Science, University of Virginia

{bxv6gs, qtq7su, shengli, vsakumar}@virginia.edu

Abstract

Diffusion models are vulnerable to backdoor attacks, where malicious attackers inject backdoors by poisoning certain training samples during the training stage. This poses a significant threat to real-world applications in the Model-as-a-Service (MaaS) scenario, where users query diffusion models through APIs or directly download them from the internet. To mitigate the threat of backdoor attacks under MaaS, *black-box input-level backdoor detection* has drawn recent interest, where defenders aim to build a firewall that filters out backdoor samples in the inference stage, with access only to input queries and the generated results from diffusion models. Despite some preliminary explorations on the traditional classification tasks, these methods cannot be directly applied to the generative tasks due to two major challenges: (1) more diverse failures and (2) a multi-modality attack surface. In this paper, we propose a black-box input-level backdoor detection framework on diffusion models, called UFID. Our defense is motivated by an insightful causal analysis: Backdoor attacks serve as the confounder, introducing a spurious path from input to target images, which remains consistent even when we perturb the input samples with Gaussian noise. We further validate the intuition with theoretical analysis. Extensive experiments across different datasets on both conditional and unconditional diffusion models show that our method achieves superb performance on detection effectiveness and run-time efficiency.

Introduction

Diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2020; Song et al. 2020; Song, Meng, and Ermon 2020) have emerged as the new state-of-the-art family of generative models due to their superior performance (Dharwal and Nichol 2021) and wide applications across a variety of domains, ranging from computer vision (Baranchuk et al. 2021; Brempong et al. 2022), natural language processing (Austin et al. 2021; Hoogeboom et al. 2021; Li et al. 2022), and robust machine learning (Blau et al. 2022; Carlini et al. 2022). Despite their success, training diffusion models requires significant time and computational resources. Consequently, it is common practice to utilize third-party models

via an API or to download them directly from the internet. This approach is known as Model-as-a-Service (MaaS).

However, it has recently been found that diffusion models are vulnerable to backdoor attacks (Gu 2024; Huang et al. 2024; Chen, Song, and Li 2023; Chou, Chen, and Ho 2023a,b; Struppek, Hintersdorf, and Kersting 2022; Guan et al. 2023), where malicious attackers poison certain training samples with a predefined trigger pattern in the training stage. Consequently, the behavior of diffusion models can be adversarially manipulated whenever the trigger pattern appears in the input query, while it remains normal with clean input queries. This vulnerability poses a serious threat to real-world applications in the MaaS setting, e.g., the online third-party diffusion models may have been backdoored to generate inappropriate images for children or to generate images that bypass copyright restrictions when a specific trigger appears in the query (Wang et al. 2024a). More seriously, the traditional training-phase defense methods (Li et al. 2021; Huang et al. 2022; Gao et al. 2023) cannot be deployed in the MaaS setting due to the defenders’ inaccessibility to the training pipeline and training data.

To mitigate the above threat, *black-box input-level backdoor detection* has recently drawn great interest. In this scenario, *input-level* indicates that defenders aim to build a firewall-style detector in the inference stage to filter out and reject backdoored inputs while allowing clean inputs to generate predictions. *Black-box* means that defenders only have access to user queries and the generated results from the deployed models, without any prior information (e.g., model weights, architectures) assumed by the previous works (Ma et al. 2022; Qiu et al. 2021; Gao et al. 2021, 2019).

Prior methods for backdoor detection in image classification, e.g., (Guo et al. 2023; Liu et al. 2023; Hu et al. 2024) cannot be adopted for generative tasks due to two major challenges: **1 More Diverse Failures**: Unlike merely generating a fixed target image (e.g., a Hello-Kitty image), backdoored diffusion models can be manipulated to produce a specific class of images (e.g., cat images), or even images with a specified abstract concept (e.g., erotic images). This implies that the target images are not necessarily unique but can vary as long as they belong to the designated target class. This variability substantially complicates the detection in the generative task. **2 Multi-Modality Attack Surface**: Unlike traditional image classifiers, which involve a single modal-

*These authors contributed equally.

†Co-corresponding Author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ity, diffusion models (e.g., Stable Diffusion) are capable of supporting multiple modalities. This diversity necessitates a unified framework for backdoor detection in diffusion models. An overall comparison of the problem setting is in the Appendix.

To address the above challenges, our intuition is motivated by a causal analysis: backdoor attacks serve as the confounder, introducing a spurious path from input to target images. The spurious path embedded in the diffusion model remains consistent even when we perturb input samples with Gaussian noise. Therefore, the backdoor generations remain consistent after minor perturbation, while the clean generations alter significantly even with a small perturbation. We further validate the causal analysis rigorously, showing that after perturbation, the difference between the diversity of clean generations and the diversity of backdoor generations is lower bounded (Corollary 3). Driven by the analysis, we develop a **Unified Framework for black-box Input-level backdoor Detection (UFID)** on diffusion models. Specifically, UFID examines each input sample by calculating its *graph density score*, which measures the similarity within the generated batch after perturbing the given input sample. The higher the graph density score, the more likely the input sample is backdoored. Compared to the existing method TERD (Mo et al. 2024), UFID is designed for a black-box setting, requiring no access to model weights or structures. Moreover, the performance gap between UFID and TERD is also satisfactory, with a maximum difference of 8% in precision, 7% in Recall, and 1% in AUC. UFID is also generalized to the scenario of conditional diffusion models, where the input can be of various modalities. To strengthen UFID’s resistance to diversity-intensive backdoor attacks, we also design strategies to integrate supplementary correspondence information into the UFID framework. In contrast, Shield (Wang et al. 2024b) operates in a white-box setting but achieves similar performance as UFID.

To sum up, our contributions in the work include: (1) **First Unified Black-box Backdoor Detection Framework for Diffusion Models**. To the best of our knowledge, our work is the first unified black-box framework for detecting input-level backdoor samples in diffusion models; (2) **Novel Causality Analysis**. We apply causality analysis for analyzing backdoor attacks on generative tasks; Besides, theoretical analysis also sheds light on our intuitions and validates the effectiveness of our method; (3) **Promising Performance**. Extensive results show that our detection method achieves an average of nearly 100% AUC on the unconditional models, and 90% AUC on the conditional models with an acceptable inference overhead.¹

Related Works

Backdoor Attacks and Defenses on Diffusion Models. Recently, a lot of works have investigated the security vulnerabilities of diffusion models by launching backdoor attacks on diffusion models. From a high-level idea, malicious backdoor attackers aim to inject a special behavior in the diffu-

sion process such that, once the predefined trigger pattern appears on the input, the special behaviors will be activated. To achieve this goal, (Chen, Song, and Li 2023) proposed to add an additional backdoor injection task on the training stage and maliciously alter the sampling procedure with a correction term. (Chou, Chen, and Ho 2023a) proposed a novel attacking strategy by only modifying the training loss function. (An et al. 2023) focused on launching attacks to text-to-image tasks, by injecting backdoors into the pre-trained text encoder. (Huang et al. 2024) proposed to apply personalization techniques to efficiently inject a malicious concept into the diffusion models. (Chou, Chen, and Ho 2023b) proposed a unified framework that covers all the popular schemes of diffusion models, including conditional and unconditional diffusion models. (Zhai et al. 2023) proposed a novel backdoor attack that can generate backdoor images as diversified as clean images. Backdoor defenses on diffusion models are highly under-explored. To the best of our knowledge, only four papers (An et al. 2023; Mo et al. 2024; Sui et al. 2024; Wang et al. 2024b) investigated backdoor defenses on diffusion models. However, Elijah (An et al. 2023) aims to detect whether a given model is backdoored, while our tasks focus on filtering backdoor samples for diffusion models in the inference stage, which are fundamentally different. TERD (Mo et al. 2024) builds a unified framework for safeguarding diffusion models, which can handle tasks such as trigger inversion, input detection, and model detection. However, the proposed technique only works for unconditional diffusion models and requires white-box access to the weights and structures of the diffusion models. Dis-Det (Sui et al. 2024) also only focuses on unconditional diffusion models. It conducts backdoor detection by checking whether the given input follows a Gaussian distribution or not. However, the defense method can be easily circumvented by using an imperceptible trigger pattern. (Wang et al. 2024b) proposed to detect backdoors on the text-to-image models based on a novel "assimilation phenomenon".

Preliminaries

Diffusion Models. Without loss of generality, diffusion models contain two parts: (1) Diffusion Process: a data distribution $q(x)$ is diffused to a target distribution $r(x)$ within T timestamps. (2) Training Process: A diffusion model ϵ_θ with parameter θ is trained to align with the reversed diffusion process, i.e., $p_\theta(x_{i-1}|x_i) = \mathcal{N}(x_{i-1}; \mu_\theta(x_i), \sigma_\theta(x_i)) = q(x_{i-1}|x_i)$. DDPM is one of the most basic diffusion models (Ho, Jain, and Abbeel 2020). DDPM assumes the target distribution $r(x) = \mathcal{N}(0, I)$ and the diffusion process $q(x_i|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_i}x_{i-1}, \beta_i I)$, where the $\{\beta_i\}_{i=1}^T$ is a predefined variance schedule that controls the step sizes. Furthermore, let $\alpha_i = 1 - \beta_i$ and $\bar{\alpha}_i = \prod_{t=1}^i \alpha_t$. By minimizing the loss function $\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_i}x_0 + \sqrt{1-\bar{\alpha}_i}\epsilon, i)\|^2$, the diffusion model is expected to be able to correctly predict the added noise given the input x_i at time i . In the inference stage, DDPM generates images by sampling from the Gaussian distribution $\mathcal{N}(0, I)$ from time $i = T$ to $i = 0$ with the generative process $p_\theta(x_{i-1}|x_i) = \mathcal{N}(x_{i-1}; \mu_\theta(x_i), \sigma_\theta(x_i))$,

¹An extended version of this work with appendices and further details is available at: <https://arxiv.org/abs/2404.01101>

where $\mu_\theta(x_i) = \frac{1}{\alpha_i}(x_i - \frac{1-\alpha_i}{1-\alpha_i\epsilon_\theta(x_i,i)})$ and $\sigma_\theta(x_i) = \frac{(1-\bar{\alpha}_{i-1})\beta_i}{1-\bar{\alpha}_i}$.

Backdoor Attacks on Diffusion Models. Different from launching backdoor attacks on the traditional models (e.g., classifiers (Gu, Dolan-Gavitt, and Garg 2017; Chen et al. 2017)), which could be achieved by poisoning training dataset, injecting backdoors into the diffusion models is much more complicated. A typical backdoor attack pipeline (Chen, Song, and Li 2023; Chou, Chen, and Ho 2023a,b) on diffusion models consists of three steps: (1) the attackers first need to mathematically define the forward backdoor diffusion process, i.e., $x_0^b \rightarrow x_T^b$, where the x_0^b denotes the target image and the x_T^b denotes the trigger image; (2) then the attackers train the diffusion models to align with the backdoored reversed process; (3) in the inference stage, the diffusion models can be prompted to generate target images when the input contains the trigger pattern, but behave normally when the input is clean (e.g., pure Gaussian noise for the DDPM model).

Threat Model. We adopt a similar threat model as in (Guo et al. 2023; Gao et al. 2021; Liu et al. 2023). Specifically, the defender is assumed to only have access to user queries (e.g., prompt) and the generated results from diffusion models. The defender aims to conduct *efficient* and *effective* black-box backdoor detection, where efficiency requires that the detection process does not significantly impact the response time of user queries, while effectiveness requires the detection process to distinguish backdoor samples and clean samples with a high accuracy rate. The challenge of this threat model arises from three factors: **(1) More Diverse Failures.** Backdoored diffusion models can be triggered to generate specific classes of images (e.g., cat images), or even images with a specified abstract concept (e.g., erotic images), extending beyond fixed target labels in traditional classification tasks. **(2) Multi-Modality Attack Surface.** Unlike traditional image classifiers that involve only one modality, diffusion models (e.g., Stable Diffusion) can support a variety of modalities. **(3) Limited Information.** The detection method only has access to the query images and the prediction labels returned by the diffusion model.

Overview of the UFID

Intuition: Backdoor Attacks under a Causal Lens

To develop a backdoor detection algorithm for generative models, we first need to address a fundamental question: *What distinguishes clean generation from backdoored generation, and how this distinction can be utilized in designing an effective detection algorithm?* Motivated by the great potential of causal inference in deep learning, we propose to leverage causal inference as a new perspective to understand the distinct mechanisms underlying clean and backdoored generation processes. Specifically, we construct causal graphs to illustrate the comparison between the two processes, as shown in Figure 1.

A causal graph is a directed acyclic graph that illustrates the causal relationships among variables, where each node represents a variable and each edge represents a causal re-

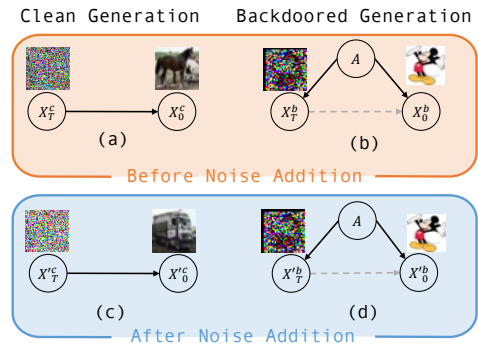


Figure 1: Causal graph of clean and backdoored generation.

lationship. For simplicity, this figure only illustrates the unconditional diffusion model. However, the causal graph can be easily extended to the conditional diffusion model by substituting the input noise (X_T) with the input text (T).

Clean Generation. As depicted in Figure 1(a), the generated image (x_0^c) is dependent on the input noise $x_T^c \sim \mathcal{N}(0, I)$. This relationship is termed the causal path, denoted as $X_T^c \rightarrow X_0^c$. Consequently, adding a small Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ with a small weight α to x_T^c results in a new input $x_T^c = x_T^c + \alpha \cdot \epsilon = \mathcal{N}(x_T^c, 0, (1 + \alpha)I)$, leading to a different generated image x_0^c , as shown in Figure 1(c).

Backdoored Generation. As shown in Figure 1(b), a backdoor attack A modifies an image x_T by injecting a trigger δ and changing the image generation process towards the target image x_0^b , denoted as $X_T^b \leftarrow A \rightarrow X_0^b$, where $x_T^b = \delta + x_T^c$. This introduces a spurious path from X_T^b to X_0^b , which lies outside the direct causal path $X_T^b \rightarrow X_0^b$, thereby establishing and strengthening the erroneous correlation between the modified input noise and the target image. Consequently, generations of poisoned noise images are primarily influenced by this spurious path (Du et al. 2021; Zhang et al. 2023; Li et al. 2021), while the direct causal path $X_T^b \rightarrow X_0^b$ plays a minor role, represented by a gray dotted line in Figure 1 (b). When an additional Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is added to the backdoored noise image x_T^b with a small weight α , the new backdoored input becomes $x_T^b \sim \mathcal{N}(\delta, (1 + \alpha)I)$, which is a combination of a new noise image $x_T^c + \alpha \cdot \epsilon$ and the trigger pattern δ . It can be interpreted as poisoning a new image $x_T^c + \alpha \cdot \epsilon$ with the trigger δ . Hence, the generated image is still affected by the attack and lies within the domain of target images. In addition, the magnitude of the perturbation is controlled by a small weight α (e.g., 0.01), ensuring that the trigger pattern remains in the new input without being disrupted.

In summary, for clean generation, a small perturbation significantly alters the output. However, triggers in backdoor samples tend to be robust features learned by neural network models. Consequently, minor perturbations of backdoor samples do not lead to substantial changes in the diffusion model’s generation results. The following theorems also validate our insights from causal analysis.

Lemma 1. *Let f_θ and $f_{\bar{\theta}}$ be two well-trained diffusion models as defined in Assumption 11 in the Appendix. Let input*

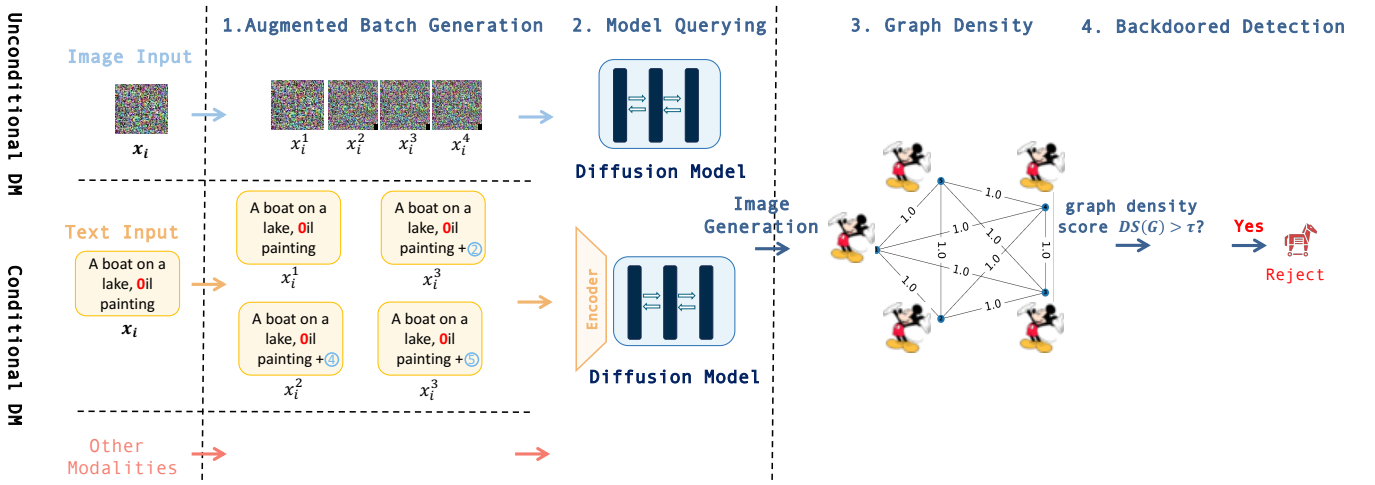


Figure 2: Pipeline of our unified framework for backdoor detection on diffusion models.

noise x'_T follow $\mathcal{N}(0, \rho^2 I)$. Let \hat{x}_0 be the generated image for x'_T and the generated distribution for clean input x_T^c be $q(x) \sim \mathcal{N}(x_c, \sigma_c I)$. We then have:

$$\hat{x}_0 = f_\theta(x'_T) \sim \mathcal{N}(x_c, \frac{\sigma_c}{\rho^2} I). \quad (1)$$

The lemma implies that if we modify the variance of the input noise by multiplying it by ρ^2 (adding small Gaussian noise), the variance of the generated image is reduced to $\frac{1}{\rho^2}$ of the original.

Theorem 2. Suppose the output domain of the diffusion model f_θ is Gaussian. Let $\mathcal{N}(x_c, \sigma_c)$ and $\mathcal{N}(x_b, \sigma_b)$ denote the distribution of clean generations and backdoor generations respectively. Let N denote the image size. We assume that $\sigma_c \geq \sigma_b + \rho^2$. Given clean input noise $x_T^c \sim \mathcal{N}(0, I)$, backdoor input noise $x_T^b = x_T^c + \delta \sim \mathcal{N}(\delta, I)$, and Lemma 1 if we perturb the clean input noises x_T and backdoor input noise x_T^b with some $\epsilon \sim \mathcal{N}(0, I)$ simultaneously, then for the resulted clean generations $\mathcal{N}(x'_c, \sigma'_c)$ and the backdoor generations $\mathcal{N}(x'_b, \sigma'_b)$, we have that $\sigma'_c - \sigma'_b \geq 1$.

Corollary 3. Under the Theorem 2, for perturbed clean generations $x_1, x_2 \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu_c, \sigma_c)$, and perturbed backdoor generations $x_3, x_4 \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu_b, \sigma_b)$, we have the following statement

$$\mathbb{E}(\|x_1 - x_2\|_2 - \|x_3 - x_4\|_2) \geq \frac{N(\sigma_c - \sigma_b) - \sigma_b}{\sqrt{N} + 1} > 0. \quad (2)$$

Theorem 2 and Corollary 3 imply that, after adding noise, the expected difference between the distance of generated clean images $\|x_1 - x_2\|_2$ and the distance of the generated backdoor images $\|x_3 - x_4\|_2$ is significantly larger than 0. This further sheds light on the intuition that *the diversity of clean generations is significantly greater than backdoor generations after adding perturbations*. Their proofs are provided in the Appendix.

Scenario 1: Unconditional Diffusion Models

Motivated by the above causal analysis, our intuition for detecting backdoor samples is that, *when the input query is perturbed with different random noises, clean samples will result in diverse generations, whereas backdoor samples will consistently generate the target images*. Therefore, we introduce a magnitude set $\mathbb{M} = \{\epsilon_1, \epsilon_2, \dots, \epsilon_{|\mathbb{M}|}\}$, where

$\epsilon_1, \epsilon_2, \dots, \epsilon_{|\mathbb{M}|} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I)$. For each input noise image x_i , we generate an input batch by adding each noise in \mathbb{M} on x_i in order with a weight α . This results in an augmented input batch $\mathbb{I} = \{x_i\} \cup \{x_i^j | x_i^j = x_i + \alpha \cdot \epsilon_j, \forall 1 \leq j \leq |\mathbb{M}|\}$. Then, we query the diffusion model with the augmented input batch \mathbb{I} as shown in the first row of Figure 2 and the detailed equation is shown as follows.

$$y_i^j = f_\theta(x_i^j), \forall x_i^j \in \mathbb{I}. \quad (3)$$

Let $y_i = \{y_i^j | 1 \leq j \leq |\mathbb{M}| + 1\}$ denote the generated batch. We can then determine whether the input query x_i is a backdoored sample by inspecting the diversity of y_i .

Scenario 2: Conditional Diffusion Models

Conditional diffusion models accept input from other modalities to guide the generation of user-intended images. For instance, in stable diffusion (Rombach et al. 2021), textual input is used to query the model. However, it is evident that the detection approach employed for unconditional diffusion models cannot be directly applied to conditional diffusion models due to the inability to introduce Gaussian noise to discrete textual input. To adapt our detection method for conditional diffusion models, we have extended the detection approach with slight modifications, as depicted in the second row of Figure 2. We leverage variations in output diversity to distinguish between clean and backdoor samples. To further enlarge this diversity gap and enhance distinguishability in conditional setting, we propose appending the input text x_i with a random phrase ph_j selected from a diverse phrase pool containing completely distinct phrases,

such as "Iron Man" and "Kitchen Dish Washer," denoted as $\mathbb{N} = \{ph_1, ph_2, \dots, ph_{|\mathbb{N}|}\}$. For each input x_i , we repeat this process $|\mathbb{M}|$ times, where $|\mathbb{M}| \ll |\mathbb{N}|$. This results in an augmented input batch $\mathbb{I} = \{x_i\} \cup \{x_i^j | x_i^j = x_i \oplus ph_j, \forall 1 \leq j \leq |\mathbb{M}|\}$, where \oplus denotes the string appending operator. Similarly, we can generate an image batch y_i by querying the target stable diffusion model using Equation 3.

A Unified Framework for Backdoor Detection

As previously discussed, the diversity of images generated by a specific input i can be used to detect backdoors. To quantify this diversity, we employ a two-step process that involves calculating both pairwise and overall similarities within the generated batch. The detailed steps are as follows:

Pairwise Similarity Calculation. Initially, we calculate semantic embeddings of generated images through a pre-trained image encoder (e.g., ViT-ImageNet (Wu et al. 2020) and CLIP (Radford et al. 2021)), denoted as $f_{\mathcal{E}}(\cdot)$. Subsequently, we calculate the local similarity for each pair of images in the generated batch using cosine similarity, represented by $S_c(\cdot, \cdot)$.

Graph Density Calculation. Following this, we construct a weighted graph and compute its graph density to represent the overall similarity of all images within the graph. Specifically, let $G_i = (V_i, \mathcal{E}_i)$ represent the similarity graph for input sample x_i , where $|V_i| = |\mathbb{M}|$ constitutes the set of vertices (symbolizing the generated images) and \mathcal{E}_i is the set of edges. Each edge's weight, connecting a pair of images $u, v \in V$, indicates their similarity score, denoted as $E[u, v] = S_c(u, v)$, where $E_i \in \mathbb{R}^{|\mathcal{E}_i|}$. In this similarity graph, the similarity between two generated images is interpreted as the distance within the graph. Next, we introduce graph density (Balakrishnan and Ranganathan 2012), as a novel metric for evaluating the overall similarity of the generated batch:

Definition 4. The graph density $DS(G_i)$ of the weighted similarity graph is defined as:

$$DS(G_i) = \frac{\sum_{(m < n)} S_c(f_{\mathcal{E}}(y_i^m), f_{\mathcal{E}}(y_i^n))}{|\mathbb{M}|(|\mathbb{M}| - 1)}$$

If $DS(G_i)$ is greater than the threshold τ , then it is determined as a backdoor sample, otherwise, it is a clean sample and the originally generated image shall be returned to the users. The total pipeline of our method is visualized in Figure 2 and the final detection algorithm is in Algorithm 1.

Remark 5 (Reliance on Pre-trained Encoders). Relying on a pre-trained encoder $f_{\mathcal{E}}(\cdot)$ might not always be feasible in practice. To relax this assumption, we also explore using a model-free metric structural similarity index measure (SSIM) to calculate the pairwise similarity between images. We report the evaluation results in the Appendix.

Remark 6 (Applicability of UFID). The effectiveness of UFID is based on a practical assumption that backdoor generations are more similar than clean generations, which has been implicitly made in the previous backdoor attacks work, e.g., the backdoor generations share a similar style (Struppek, Hintersdorf, and Kersting 2022; Huang, Guo, and

Juefei-Xu 2023), object (Chou, Chen, and Ho 2023b,a), or semantic concept (Chen, Song, and Li 2023). It is also observed that when the target images of backdoor attacks are as diversified as the clean images (Zhai et al. 2023), the performance of UFID will be limited. However, we could incorporate supplementary information to enhance the UFID. Detailed evaluations are provided in the Appendix.

Experiments

Experimental Settings

Attack Baselines. To the best of our knowledge, the existing backdoor attacks on diffusion models include two unconditional-DM-based backdoor attacks: TrojDiff (Chen, Song, and Li 2023) and BadDiffusion (Chou, Chen, and Ho 2023a), and three conditional-DM-based backdoor attacks: Rickrolling (Rick) (Struppek, Hintersdorf, and Kersting 2022), VillanDiffusion (VillanDiff) (Chou, Chen, and Ho 2023b), and Personalization (Personal) (Huang, Guo, and Juefei-Xu 2023). We consider all five backdoor attacks as our attack baselines. It is also noted that TrojDiff, Rickrolling, and Personalization all support **diversity-preserving backdoor attacks**, where the attackers' target images are diversified. Detailed descriptions of them are provided in the Appendix.

Defense Baselines. We compared our method with the existing well-established defense method TERD (Mo et al. 2024) on the unconditional diffusion models and Shield (Wang et al. 2024b) on the conditional diffusion models. It is noted that both TERD and Shield require additional **white-box access to the model weights and structures**, which are not always feasible in our MaaS setting.

Models and Datasets. Different backdoor attacks are built based on different backbone models and samplers. To facilitate evaluation, TrojDiff and BadDiffusion are evaluated on DDPM, while VillanDiffusion, Rickrolling, and Personalization are evaluated on Stable Diffusion v1.4 (Rombach et al. 2021). For the training datasets, we choose CIFAR10 (Krizhevsky, Nair, and Hinton) and CelebA (Liu et al. 2015) for TrojDiff and BadDiffusion, and choose CelebA-D (Jiang et al. 2021) and Pokemon (Pinkney 2022) for VillanDiffusion, Rickrolling, and Personalization.

Metrics. Following the prior works on backdoor detection, we adopt three popular metrics for evaluating the effectiveness of our detection method: Precision (P), Recall (R), and Area under the Receiver Operating Characteristic (AUC).

Implementation Details. All the models are well-trained with the default hyper-parameters reported in the original papers. Following the previous works (Lee et al. 2018; Guo et al. 2023), we evaluate our detection method with a positive (i.e., attacked) and a negative (i.e., clean) dataset. Due to space limits, the details for constructing the two datasets and the default hyper-parameters are provided in the Appendix.

Main Results

Effectiveness. Table 1 presents the performance of our detection method against backdoor attacks on *unconditional diffusion models*, while Table 2 presents the performance on *conditional diffusion models*. As shown, the AUC values

Dataset	Attacks	UFID(black-box)			TERD(white-box)		
		P	R	AUC	P	R	AUC
Cifar10	TrojDiff(D2I)	0.95	0.94	1.00	1.00	1.00	1.00
	TrojDiff(In)	0.93	0.93	0.98	1.00	1.00	1.00
	TrojDiff(Out)	0.93	0.92	1.00	1.00	1.00	1.00
	BadDiffusion	0.93	0.95	1.00	1.00	1.00	1.00
	Average	0.93	0.94	1.00	1.00	1.00	1.00
CelebaA	TrojDiff(D2I)	0.93	0.92	1.00	1.00	1.00	1.00
	TrojDiff(In)	0.90	0.89	0.96	1.00	1.00	1.00
	TrojDiff(Out)	0.91	0.92	0.98	1.00	1.00	1.00
	BadDiffusion	0.97	0.95	1.00	1.00	1.00	1.00
	Average	0.93	0.92	0.99	1.00	1.00	1.00

Table 1: Performance of the proposed detection method against backdoor attacks on unconditional diffusion models.

Dataset	Attacks	UFID(black-box)			Shield(white-box) ¹		
		P	R	AUC	P	R	AUC
CelebaA-D	VillanDiff	0.92	0.95	0.96	0.80	0.95	-
	Rick(TPA)	0.87	0.84	0.90	0.96	0.85	-
	Rick(TAA)	0.82	0.81	0.87	-	-	-
	Personal	0.76	0.72	0.82	-	-	-
	Average	0.85	0.84	0.89	0.88	0.90	-
Pokemon	VillanDiff	0.91	0.93	0.94	-	-	-
	Rick(TPA)	0.83	0.85	0.91	-	-	-
	Rick(TAA)	0.80	0.81	0.87	-	-	-
	Personal	0.73	0.77	0.81	-	-	-
	Average	0.82	0.85	0.89	-	-	-

¹ Due to unavailable codes, we use the reported values in the original paper directly.

Table 2: Performance of the proposed detection method against backdoor attacks on conditional diffusion models.

for different backdoor attack methods on all the evaluated datasets are over 0.8, suggesting that our method can effectively distinguish backdoor and clean samples. Compared to the baseline methods, UFID shows a comparable performance with a slight drop. However, the decrease is reasonable as the two baselines are white-box methods, which require additional access to the model weights and structures. The similar performance drop has been observed in the previous studies (Guo, Li, and Liu 2021; Guo et al. 2023).

Efficiency. Figure 3 illustrates the efficiency of our detection method against TrojDiff(D2I) on CIFAR10 ($32 \times 32 \times 3$) dataset. We query the diffusion models with 320 samples with a batch size of 64 and record the average inference speed, defined as the average time consumption for a sample. The y-axis comprises three components: 'Vanilla,' representing the average inference speed without UFID; '+augmented Batch Query' representing the average inference speed after an augmented batch query; and '+Similarity Calculation' representing the average inference speed after similarity graph construction and calculation. Due to the increased number of query samples, UFID inevitably results in

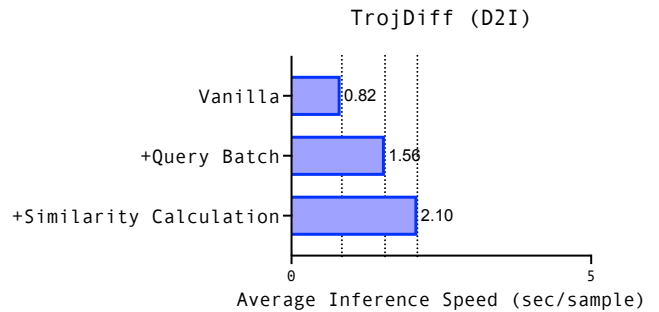


Figure 3: Average inference speed against TrojDiff(D2I) on the Cifar10.

a lower inference speed than that of the vanilla mode. However, the increased time consumption is within expectations. Figure 23 presents the efficiency of the UFID against VillanDiffusion on the Pokemon ($512 \times 512 \times 3$) dataset. Due to the acceleration techniques in the modern sampler (Song, Meng, and Ermon 2020; Lu et al. 2022), stable diffusion models are usually denoised with only a small number of steps (e.g., 25) to generate high-resolution images. Therefore, we could observe that the UFID only slightly increases the inference time. To further investigate how the selected inference steps influence the efficiency, we evaluate the inference speed with different numbers of inference steps. The results show that when setting the inference step from 10 to 75, the inference overhead is acceptable.

Ablation Studies

In this section, we discuss how the hyper-parameters influence the effectiveness of the UFID.

The Influence of Different Pre-trained Encoders. The pre-trained encoder is important in our detection method. To evaluate its impact on the effectiveness of our detection method, we test the performance of UFID when integrated with different pre-trained encoders. For the space limit, we present the results in the Table 3 - Table 8 in the Appendix. According to the tables, UFID works well when integrated with different pre-trained encoders. In particular, CLIP encoders show consistently good performance across different datasets, due to their strong generalization ability from the pre-training stage.

The Influence of Magnitude Set. Figure 4 investigates the impact of the magnitude size on the running-time efficiency and effectiveness, where the left y-axis denotes the performance values, and the right y-axis denotes the average inference speed. Due to the space limit, we only present partial results in the main manuscript, while the remaining parts can be found in the Appendix. As the figure shows, our detection method achieves a stably satisfactory performance against all backdoor attacks when the size is over four. Additionally, a size of four would also yield a balanced trade-off on efficiency and effectiveness.

The Influence of Available Validation Dataset. Figure 16 in the Appendix presents the impact of the available validation dataset on the performance, where the X-axis values

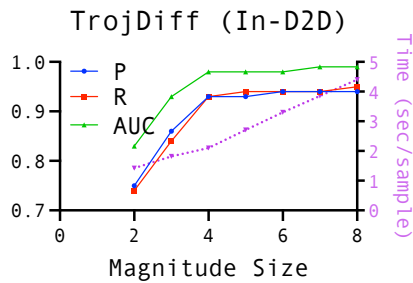


Figure 4: Performance with different sizes of magnitude set.

denote the number of available validation samples and the y-axis denotes the performance values. As the figure suggests, with more validation samples available, the performances tend to become more stable. However, it is also noted that if the number of validation samples becomes exceedingly large, there is a slight drop in performance. A possible explanation for this phenomenon is that more validation samples also introduce more noisy information, leading to an unexpected threshold value.

The Influence of Image Size. Figure 19 in the Appendix investigates whether UFID’s performance will be influenced when handling high-resolution images. Specifically, we evaluate UFID against TrojDiff and BadDiffusion on an augmented CIFAR10 dataset, where the image size is manually scaled to 64, 128, and 256. We could see that the image size does not have any influence on the performance, demonstrating UFID’s potential to handle high-resolution images.

The Influence of Different Poisoning Rate. Figure 14 in the Appendix explores whether the performance of UFID is sensitive to the backdoor poisoning rate. We evaluate the UFID under poisoning rate from 0.05 to 0.30. The results reveal that our method can perform satisfactorily under different poisoning rates. Moreover, with the increase of the poisoning rate, the performance becomes more stable.

Discussions

Visualizations of Similarity Graphs. To better understand how UFID helps to detect backdoor samples, we visualize the similarity graphs in Figure 5. Due to the space limit, we only present similarity graphs against the TrojDiff(In-D2D) on CIFAR10 in the main manuscript. More qualitative examples are presented in the appendix. Each node in the similarity graph denotes the generated images of the query batch, while each edge denotes the cosine similarity scores of the embedding of any two images. As shown, the similarity scores for the clean query batch are significantly lower than those for the backdoor query batch, validating our intuitions for backdoor detection.

Visualizations of Scores Distributions. Figure 17 and Figure 18 in the Appendix present the distributions of the graph density \mathcal{S}_i for backdoor samples and clean samples respectively. As shown, the distribution of backdoor samples tends to be more clustered in a narrow range, while that of clean samples tends to be spread out. Moreover, there is a distinct

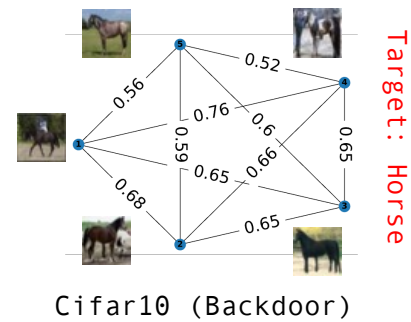


Figure 5: Similarity graphs against TrojDiff (In-D2D) attack on CIFAR10 dataset.

gap between the two distributions, suggesting that UFID can effectively distinguish backdoor and clean samples.

Resilient against Adaptive Backdoor Attacks. We evaluate our detection method against an adaptive attacker who already has prior information about our detection method. Therefore, the attacker might try to make the generated images more diversified to avoid being detected. Specifically, rather than training a diffusion model that maps the trigger to the target images (e.g., erotic images), the attacker maps the trigger to a target domain that contains both the target images and a small number of clean images. In this way, the attacker achieves a more stealthy backdoor attack by sacrificing the attack success rate. We further define the ratio between the number of clean images to the backdoor samples in this target domain as the “blending ratio”. We evaluate the UFID against TrojDiff(D2I) and employ mean square error (MSE) between the generated backdoor images and the target image (e.g., Mickey Mouse) as the attack success rate. Figure 20 in the Appendix presents the performance of the UFID under different blending ratios. As shown, the UFID’s performance gradually decreases when the blending ratio rises. However, the average MSE across generated backdoor samples abruptly exceeds 0.15, which suggests the failure in injecting backdoors. The right-hand side also provides some samples in the CIFAR10 dataset, where we notice that images with an MSE of 0.15 to the target image are already completely different from the target image.

Conclusion and Future Directions

In this paper, we propose a simple unified framework for backdoor detection on diffusion models under the MaaS setting. Our framework is first motivated by a causality analysis on image generation and further validated by theoretical analysis. Motivated by the analysis, we design a unified method for distinguishing backdoor and clean samples for both conditional and unconditional diffusion models. Extensive experiments demonstrate the effectiveness of our method. Despite the great success, there are still many directions to be explored in the future. For example, UFID still requires a small amount of validation samples to determine the threshold. Can we relax the assumption?

Acknowledgments

The work is supported in part by the U.S. Office of Naval Research Award under Grant Number N00014-24-1-2668 and the National Science Foundation under Grants IIS-2316306, CNS-2330215, CCF-1918656, CNS-2317193, IIS-2331315.

References

- An, S.; Chou, S.-Y.; Zhang, K.; Xu, Q.; Tao, G.; Shen, G.; Cheng, S.; Ma, S.; Chen, P.-Y.; Ho, T.-Y.; et al. 2023. Elijah: Eliminating Backdoors Injected in Diffusion Models via Distribution Shift. *arXiv preprint arXiv:2312.00050*.
- Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and Van Den Berg, R. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993.
- Balakrishnan, R.; and Ranganathan, K. 2012. *A textbook of graph theory*. Springer Science & Business Media.
- Baranchuk, D.; Rubachev, I.; Voynov, A.; Khulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.
- Blau, T.; Ganz, R.; Kawar, B.; Bronstein, A.; and Elad, M. 2022. Threat model-agnostic adversarial defense using diffusion models. *arXiv preprint arXiv:2207.08089*.
- Brempong, E. A.; Kornblith, S.; Chen, T.; Parmar, N.; Minderer, M.; and Norouzi, M. 2022. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4175–4186.
- Carlini, N.; Tramer, F.; Dvijotham, K. D.; Rice, L.; Sun, M.; and Kolter, J. Z. 2022. (certified!!) Adversarial robustness for free! *arXiv preprint arXiv:2206.10550*.
- Chen, W.; Song, D.; and Li, B. 2023. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4035–4044.
- Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.
- Chou, S.-Y.; Chen, P.-Y.; and Ho, T.-Y. 2023a. How to Backdoor Diffusion Models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4015–4024.
- Chou, S.-Y.; Chen, P.-Y.; and Ho, T.-Y. 2023b. VillanDiffusion: A Unified Backdoor Attack Framework for Diffusion Models. *arXiv preprint arXiv:2306.06874*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.
- Du, M.; Manjunatha, V.; Jain, R.; Deshpande, R.; Dernoncourt, F.; Gu, J.; Sun, T.; and Hu, X. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. *arXiv preprint arXiv:2103.06922*.
- Gao, K.; Bai, Y.; Gu, J.; Yang, Y.; and Xia, S.-T. 2023. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4005–4014.
- Gao, Y.; Kim, Y.; Doan, B. G.; Zhang, Z.; Zhang, G.; Nepal, S.; Ranasinghe, D. C.; and Kim, H. 2021. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, 19(4): 2349–2364.
- Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 113–125.
- Gu, J. 2024. Responsible Generative AI: What to Generate and What Not. *arXiv preprint arXiv:2404.05783*.
- Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.
- Guan, Z.; Sun, L.; Du, M.; and Liu, N. 2023. Attacking Neural Networks with Neural Networks: Towards Deep Synchronization for Backdoor Attacks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23*, 608–618. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701245.
- Guo, J.; Li, A.; and Liu, C. 2021. Aeva: Black-box backdoor detection using adversarial extreme value analysis. *arXiv preprint arXiv:2110.14880*.
- Guo, J.; Li, Y.; Chen, X.; Guo, H.; Sun, L.; and Liu, C. 2023. SCALE-UP: An Efficient Black-box Input-level Backdoor Detection via Analyzing Scaled Prediction Consistency. In *The Eleventh International Conference on Learning Representations*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Hoogeboom, E.; Nielsen, D.; Jai, P.; Forré, P.; and Welling, M. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34: 12454–12465.
- Hu, M.; Guan, Z.; Guo, J.; Zhou, Z.; Zhang, J.; and Li, S. 2024. BBCaL: Black-box Backdoor Detection under the Causality Lens. *Transactions on Machine Learning Research*. Featured Certification.
- Huang, K.; Li, Y.; Wu, B.; Qin, Z.; and Ren, K. 2022. Backdoor defense via decoupling the training process. *arXiv preprint arXiv:2202.03423*.
- Huang, Y.; Guo, Q.; and Juefei-Xu, F. 2023. Zero-day backdoor attack against text-to-image diffusion models via personalization. *arXiv preprint arXiv:2305.10701*, 1(2).
- Huang, Y.; Juefei-Xu, F.; Guo, Q.; Zhang, J.; Wu, Y.; Hu, M.; Li, T.; Pu, G.; and Liu, Y. 2024. Personalization as a Shortcut for Few-Shot Backdoor Attack against Text-to-Image Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19): 21169–21178.
- Jiang, Y.; Huang, Z.; Pan, X.; Loy, C. C.; and Liu, Z. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13799–13808.

- Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. CIFAR-10 (Canadian Institute for Advanced Research).
- Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.
- Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343.
- Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34: 14900–14912.
- Liu, X.; Li, M.; Wang, H.; Hu, S.; Ye, D.; Jin, H.; Wu, L.; and Xiao, C. 2023. Detecting Backdoors During the Inference Stage Based on Corruption Robustness Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16363–16372.
- Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.
- Ma, W.; Wang, D.; Sun, R.; Xue, M.; Wen, S.; and Xiang, Y. 2022. The” Beatrix”Resurrections: Robust Backdoor Detection via Gram Matrices. *arXiv preprint arXiv:2209.11715*.
- Mo, Y.; Huang, H.; Li, M.; Li, A.; and Wang, Y. 2024. TERD: A Unified Framework for Safeguarding Diffusion Models Against Backdoors. In *Forty-first International Conference on Machine Learning*.
- Pinkney, J. N. M. 2022. Pokemon BLIP captions. <https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/>.
- Qiu, H.; Zeng, Y.; Guo, S.; Zhang, T.; Qiu, M.; and Thuraisingham, B. 2021. Deepsweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 363–377.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752*.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Struppek, L.; Hintersdorf, D.; and Kersting, K. 2022. Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. *arXiv preprint arXiv:2211.02408*.
- Sui, Y.; Phan, H.; Xiao, J.; Zhang, T.; Tang, Z.; Shi, C.; Wang, Y.; Chen, Y.; and Yuan, B. 2024. DisDet: Exploring Detectability of Backdoor Attack on Diffusion Models. *arXiv preprint arXiv:2402.02739*.
- Wang, H.; Shen, Q.; Tong, Y.; Zhang, Y.; and Kawaguchi, K. 2024a. The Stronger the Diffusion Model, the Easier the Backdoor: Data Poisoning to Induce Copyright Breaches Without Adjusting Finetuning Pipeline. *arXiv preprint arXiv:2401.04136*.
- Wang, Z.; Zhang, J.; Shan, S.; and Chen, X. 2024b. T2IShield: Defending Against Backdoors on Text-to-Image Diffusion Models. In *ECCV*.
- Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Yan, Z.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; and Vajda, P. 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *arXiv:2006.03677*.
- Zhai, S.; Dong, Y.; Shen, Q.; Pu, S.; Fang, Y.; and Su, H. 2023. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1577–1587.
- Zhang, Z.; Liu, Q.; Wang, Z.; Lu, Z.; and Hu, Q. 2023. Backdoor Defense via Deconfounded Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12228–12238.