UIBDiffusion: Universal Imperceptible Backdoor Attack for Diffusion Models

Yuning Han[1,*] Bingyin Zhao[2,*] Rui Chu[3] Feng Luo[4] Biplab Sikdar[2] Yingjie Lao[3,†]

[1]Columbia University [2]National University of Singapore [3]Tufts University [4]Clemson University

Abstract

Recent studies show that diffusion models (DMs) are vulnerable to backdoor attacks. Existing backdoor attacks impose unconcealed triggers (e.g., a gray box and eyeglasses) that contain evident patterns, rendering remarkable attack effects yet easy detection upon human inspection and defensive algorithms. While it is possible to improve stealthiness by reducing the strength of the backdoor, doing so can significantly compromise its generality and effectiveness. In this paper, we propose UIBDiffusion, the universal imperceptible backdoor attack for diffusion models, which allows us to achieve superior attack and generation performance while evading state-of-the-art defenses. We propose a novel trigger generation approach based on universal adversarial perturbations (UAPs) and reveal that such perturbations, which are initially devised for fooling pre-trained discriminative models, can be adapted as potent imperceptible backdoor triggers for DMs. We evaluate UIBDiffusion on multiple types of DMs with different kinds of samplers across various datasets and targets. Experimental results demonstrate that UIBDiffusion brings three advantages: 1) Universality, the imperceptible trigger is universal (i.e., image and model agnostic) where a single trigger is effective to any images and all diffusion models with different samplers; 2) Utility, it achieves comparable generation quality (e.g., FID) and even better attack success rate (i.e., ASR) at low poison rates compared to the prior works; and 3) Undetectability, UIBDiffusion is plausible to human perception and can bypass Elijah and TERD, the SOTA defenses against backdoors for DMs. Code is available at https://github.com/TheLaoLab/UIBDiffusion.

1. Introduction

Diffusion models (DMs) [1–3] have emerged as the state-of-the-art generative paradigm in recent years and achieved unprecedented success in image [4, 5], video [6–8] text [9–11] and audio [12–17] synthesis. The success is highly attributed to training the foundation model on large-scale datasets [5]. Such a practice places DMs at risk of being compromised by data poisoning attacks [18–22], where adversaries contaminate the clean training dataset with a small fraction of malicious samples and undermine the model performance. This imposes a critical security concern on the practical deployment of DMs.

Backdoor attacks are emerging threats to DMs. Recent studies [23–25] show that DMs are vulnerable to such attacks, which implant backdoor into victim models by training them on poisoned datasets and manipulate their behavior through inputs stamped with predefined triggers. Existing backdoor attacks for DMs usually employ unconcealed triggers such as a gray box, an image of Hello Kitty, and eyeglasses in the forward and reversed (i.e., sampling) diffusion process, as shown in Fig. 1. While such trigger designs achieve superior attack performance and preserve the decent generation ability of victim models, they are easy to detect via human inspection and advanced defensive algorithms as they have obvious patterns that can be precisely remodeled via reverse engineering [26, 27]. Motivated by this observation, we aim to design a general, powerful yet stealthy trigger. However, even for the traditional backdoor attacks against discriminative models (e.g., image classifiers) [28–30], it is challenging to acquire such a trigger since it either requires complex and carefully designed generators to generate image-specific or model-specific triggers [31–38] or suffers from a relatively low attack success rate (ASR) [39, 40], or fails to evade modern defenses [41–46]. This raises an interesting question in backdoor attacks against DMs: *Is there a trigger simultaneously possessing generality, effectiveness, and stealthiness for backdoor attacks against DMs?*

Driving by the question, in this work, we propose UIBDiffusion, the universal imperceptible backdoor triggers for DMs that bring the following advantages: 1) *Universality*: the trigger is image-agnostic and model-agnostic so that it can be applied to arbitrary images and DMs; 2) *Utility*: the trigger can achieve high ASR while maintaining the generation capability of DMs; 3) *Undetectability*: the trigger is plausible to humans and capable of circumventing

*Equal contribution.
†Corresponding author.

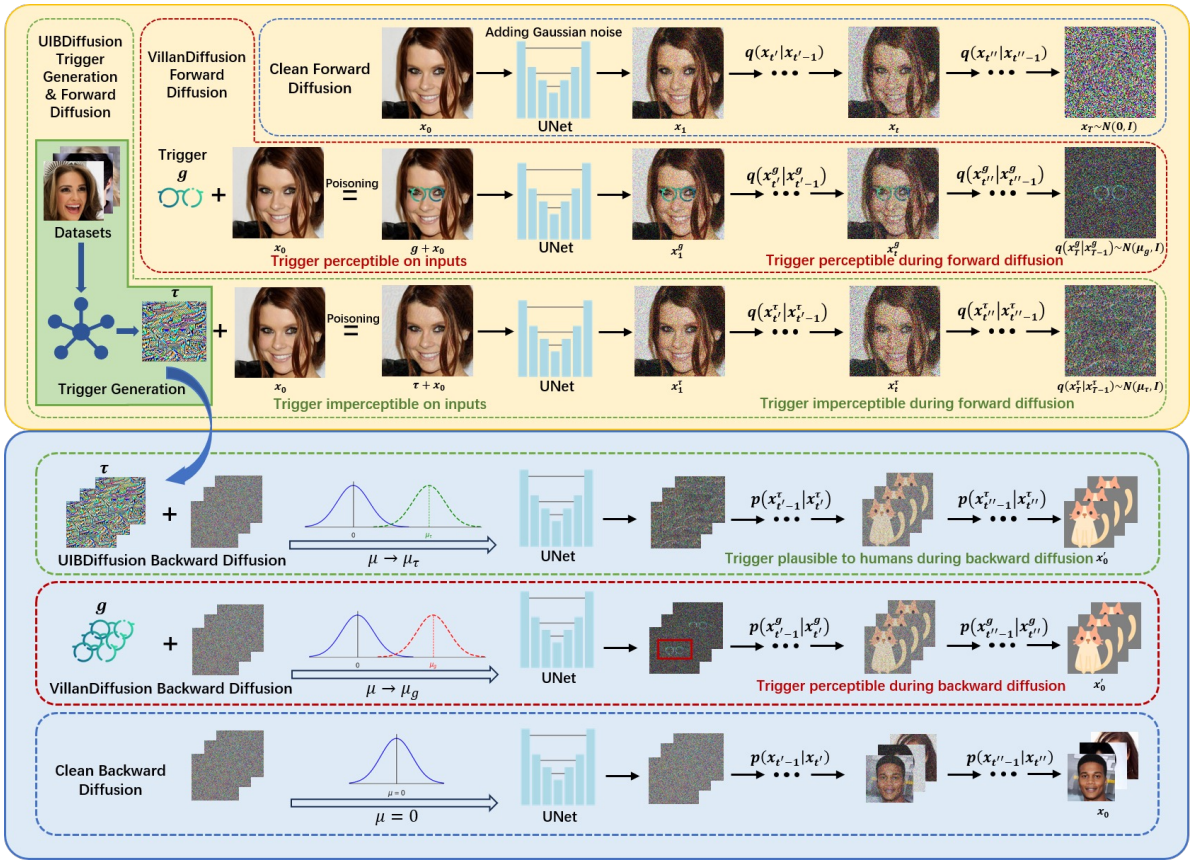

Figure 1. Illustrations of the forward diffusion process (top block with the yellow background) and backward diffusion process (bottom block with the blue background) of a clean diffusion model (blue dash line), VillanDiffusion [25] (red dash line) and UIBDiffusion (Ours, green dash line). The UIBDiffusion trigger (τ) is plausible to humans in all phases from data poisoning to forward diffusion and backward diffusion while the glasses trigger ($\mathbf{g}$) in prior works is perceptible. UIBDiffusion trigger is highly effective since it introduces a similar distribution shift as the glasses trigger, which secures the attack performance during the backward diffusion process. However, it is hard to detect as the trigger does not possess specific patterns and cannot be inverted by existing defensive algorithms. We empirically verify the effectiveness in Section 4.

state-of-the-art defenses. Inspired by the adversarial perturbation [47] originally designed for attacking pre-trained image classifiers, we reveal that universal adversarial perturbations [48] (UAPs), a variant of adversarial perturbations, naturally hold the aforementioned properties and can be adapted as the UIBDiffusion triggers. First, UAPs are image-agnostic and model-agnostic. Second, we demonstrate that they are effective triggers because they introduce a similar distribution shift as prior works, yet they are stealthy because they do not possess clear patterns and are hard to estimate by existing trigger inversion algorithms. Moreover, they are trivial and imperceptible noises that humans can barely identify, as shown in Fig. 1. Our contributions are summarized as follows:

- To the best of our knowledge, UIBDiffusion is the first imperceptible backdoor attack against diffusion models.
- We discover that adversarial perturbations designed for *discriminative models* can be adapted as backdoor triggers for *generative models*. We reveal why such triggers can achieve high ASR and evade SOTA defenses.
- We propose a practical trigger generation approach adapted from UAP, which is specifically designed to enhance universality, attack effectiveness and accessibility of triggers.
- We experimentally demonstrate that UIBDiffusion is highly effective towards multiple diffusion models with different samplers and achieves comparable generation quality on benign input compared to clean DMs and superior attack success rate and input stealthiness compared to prior works.
- While existing works face a trade-off between utility and undetectability, we empirically show that UIBDiffusion triggers are highly resistant to the SOTA trigger inversion-based defenses, effectively bypassing Elijah [26] and TERD [27], the state-of-the-art defenses against backdoors on DMs.

2. Related Works

2.1. Diffusion Models

Diffusion models are latent variable models composed of two processes. The forward diffusion process imposes Gaussian noise on input data in each time step along the Markov chain and progressively diffuses the data distribution to an isotropic Gaussian distribution. The backward diffusion process is a reversed Markov chain that gradually denoises the sampled noise from the Gaussian distribution to the input data distribution. Mainstream diffusion models include DDPM [1], LDM [3] and NCSN [49–51], etc. Despite the excellent performance, DMs suffer from slow sampling during the backward process. There exists a line of works that leverage different techniques such as generalized non-Markovian chain and Ordinary Differential Equation (ODE) to improve the samplers and accelerate the generation process: DDIM [2], DEIS [52], UniPC [53], Heun [54], DPM-Solver [55] and DPM-Solver++ [7]. We comprehensively evaluate UIBDiffusion on these models and samplers to showcase the utility of our trigger.

2.2. Backdoor Attacks and Defenses on Diffusion Models

Attacks. Backdoor attacks [28–30, 56], originally explored in the context of discriminative models such as image classifiers, induce models to make wrong predictions on triggered images. Although having been extensively studied for discriminative models, backdoor attacks are under-explored in diffusion models. BadDiffusion [23], TrojDiff [24] and VillanDiffusion [25] are pioneering works in this field. Unlike traditional backdoor attacks, those targeting DMs aim to force models to synthesize a targeted image upon malicious inputs (i.e., noises with triggers) while generating normal images upon benign inputs (i.e., isotropic Gaussian noises). BadDiffusion shows how to backdoor DMs with unconcealed triggers such as a gray box and a stop sign. TrojDiff proposes to use a whole image (e.g., Hello Kitty) as the trigger and demonstrates that backdoor can be generalized to DDIM and different adversarial goals. However, both attacks are limited to certain DMs (e.g., DDPM) and samplers (e.g., ancestral and DDIM). VillanDiffusion resolves the limitations and offers a unified framework for all kinds of DMs and samplers. Other works show backdoor can be injected to text-to-image DMs and activated by text prompts [57–59]. However, these attacks draw less attention from defenders as they can eliminate the backdoor by cleansing the text encoder. Thus, our attack focuses on the primary defensive objective of the defender: pixel level (i.e., image) trigger generation.

Defenses. Backdoor attacks against DMs are shown to be immune to defenses designed for traditional image classifier backdoors [23, 25, 44, 60, 61]. The reason is that DMs take sampled noises as inputs while image classifiers take natural images as inputs [27]. Elijah [26] is the first defensive algorithm specifically proposed against DM backdoors. It presents a trigger inversion scheme that can approximate the triggers and detect and remove the injected backdoor accordingly. TERD [27] employs a similar trigger estimation strategy and improves the trigger inversion process via a unified loss and trigger refinement. Based on the reconstructed trigger, it subsequently proposes to measure the KL divergence between benign and reversed distribution to detect the backdoor injected in models. Elijah and TERD achieve remarkable detection performance on existing DM backdoors. Thus, we evaluate UIBDiffusion against them and show that UIBDiffusion can bypass both defenses.

2.3. Adversarial Perturbation

The concept of adversarial perturbation is proposed in [47], demonstrating that deep neural networks are vulnerable to small additive noise. These perturbations are typically subtle and imperceptible, hence are widely used to craft adversarial examples [62–64] to fool image classifiers. Later, the work in [48] devised universal adversarial perturbation (UAP), a model-agnostic and image-agnostic variant of adversarial perturbation. We find that UAP is a particularly competitive triggers for backdoor attacks against DMs due to their imperceptibility, generalization property and accessibility.

3. UIBDiffusion

3.1. Threat Model

We adopt the same threat model and attack scenario presented in the existing backdoor attacks on DMs [23–25], where adversaries publish the backdoored models on websites such as HuggingFace, and users can acquire the pretrained models via third-party sources. Adversaries are expected to be able to inject the backdoor and verify the attack and generation performance before releasing the models. Users are supposed to get access to the backdoor models and the subset of clean training data and test the model utility on their side. Note that in practice, UIBDiffusion applies to a more stringent threat model where users can access the entire training dataset (i.e., contaminated dataset with poisoned samples) and apply any defenses before model deployment as the trigger is imperceptible and undetectable.

3.2. UIBDiffusion Trigger Generation

Recall that our goal is to design a general, effective, and stealthy trigger (**denoted as** τ) for backdoor attacks against DMs. Due to the imperceptibility, generalization, and effectiveness properties of UAPs, they can be adapted as the UIBDiffusion trigger. We first briefly introduce the mechanism and generation process of UAPs. For a given image

classifier $\mathcal{C}$ and a data distribution q, UAP algorithms aim to find a perturbation vector v such that $\mathcal{C}(x+v) \neq \mathcal{C}(x)$ for most $x \sim q$ (i.e., image agnostic). The original UAP algorithm [48] achieves this goal by measuring the minimal perturbation that pushes the perturbed input to the decision boundary, which can be mathematically expressed as:

$$\Delta v_i = \arg\min_d ||d||_2 \quad \textbf{s.t.} \quad \mathcal{C}(x_i + v + d) \neq \mathcal{C}(x_i), \quad (1)$$

where i is the image index of a subset of the dataset selected for the computation. The final perturbation is updated as $\mathcal{P}_\xi(v + \Delta v_i)$, where $\mathcal{P}_\xi(\cdot)$ is a projection function and ξ is the perturbation budget. Note that in the context of DMs backdoor, the perturbation v is actually the imperceptible trigger τ. We use τ for consistent notation from now on to avoid confusion.

However, the original UAP algorithm using DeepFool [65] sometimes yields less effectiveness, and we experimentally observe it renders a relatively lower ASR as the backdoor trigger for DMs. Thus, we propose a novel trigger generation approach inspired by the GUAP [66], an enhanced version of UAP that is more effective and robust. Our trigger generation process adopts the advantage of both additive (i.e., l_∞-bounded) and non-additive (i.e., spatial transformation) perturbations, where the additive perturbation is adopted as the UIBDiffusion trigger τ and coordinately optimized with the non-additive perturbation. We denote the non-additive noise as f. We employ a generator $\mathcal{G}_\gamma$ (parameterized by γ), which is akin to the generator in generative adversarial nets (GAN) [67] that take a latent noise z as input, and generate f and trigger τ, as shown in lines 3, 4 and 5 in Algorithm 1. In contrast to the GAN that optimizes the generator under the guide of a discriminator, we update our trigger generator under the guide of an image classier $\mathcal{C}$ (e.g., VGG and ResNet, etc.), which indicates whether the adversarial perturbation is effective or not, as expressed in Equation 2, where $\mathcal{H}(\cdot)$ is the cross entropy loss and $\otimes$ represents for the spatial transformation operation. Meanwhile, the non-additive perturbation is updated according to the loss function expressed in Equation 3, where $\mathcal{S}(\cdot)$ is summed max pooling values of x using a kernel of size 4 during the spatial transformation and n is the number of images in the selected subset of the dataset. By incorporating such a design, we can progressively optimize the generator and hence improve the effectiveness of the generated UIBDiffusion trigger, as described in lines 5, 7 and 8 in Algorithm 1, respectively. We then optimize the trigger generation process from latent variables, with the objective of maximizing the fooling rate against the image classifier. The completed trigger generation procedures are presented in Algorithm 1 and the concrete architecture of $\mathcal{G}_\gamma$ and flow illustration are provided in the Appendix as they are similar to GAN.

Algorithm 1 UIBDiffusion Trigger Generation

Require: Image set X, classifier $\mathcal{C}$, latent noise z, generator $\mathcal{G}_\gamma$, training epoch M, number of images n, non-additive and additive perturbation budget Γ and ξ

1: **for** $m = 1 \dots M$ **do**
2: **for** $i = 1 \dots n$ **do**
3: $x_i \sim X, z \sim \mathcal{N}(0,1)$
4: $\mathcal{G}_{\gamma_i}^{m-1}(z) \rightarrow f_i^{m-1}$
5: $\mathcal{G}_{\gamma_i}^{m-1}(z) \rightarrow \tau_i^{m-1}$
6: $f = f_i^m \leftarrow \frac{\Gamma}{\max_i \sqrt{\frac{1}{n}\sum_i^n (\mathcal{S}(x_i), f_i^{m-1})}}$
7: $\tau = \tau_i^m \leftarrow \frac{\xi}{||\tau_i^{m-1}||_\infty}$
8: $\gamma_i^m \leftarrow \gamma_i^{m-1} - \nabla_{\gamma_i^{m-1}} \mathcal{L}_\mathcal{G}\left(\mathcal{C}(x_i \otimes f + \tau), \mathcal{C}(x_i)\right)$
9: **end for**
10: **end for**
11: **return** Trigger τ

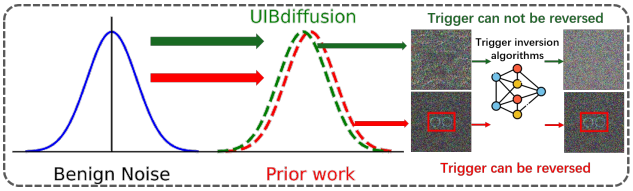

Figure 2. Illustration of trigger mechanisms. UIBDiffusion triggers introduce a similar distribution shift as triggers in prior works, which secure a high attack performance. However, since our triggers do not have a specific pattern, they are difficult to remodel by trigger inversion algorithms.

$$\mathcal{L}_\mathcal{G} = -\mathcal{H}(\mathcal{C}(x \otimes f + \tau), \mathcal{C}(x)), \quad (2)$$

$$\mathcal{L}_f = \max_i \sqrt{\frac{1}{n}\sum_i^n (\mathcal{S}(x_i, f))} \quad (3)$$

The effectiveness of UIBDiffusion against backdoor attacks on generative models, such as DMs, stems from key insights into backdoor dynamics. As revealed in [23–25], backdoor for DMs needs to introduce an evident distribution shift and keep such a shift during the reversed Markov chain to secure a high ASR. We empirically find that the UIBDiffusion triggers also lead to a similar distribution shift when imposed on benign sampled noise, as shown in Fig. 2. However, because they are adapted from UAPs, which exhibit minimal, hard-to-detect patterns with subtle perturbations, we hypothesize that these triggers are difficult for trigger inversion algorithms to estimate accurately. Consequently, unlike the triggers with specific patterns, they are managed to circumvent the SOTA defenses and remain highly stealthy. We provide comprehensive and extensive evaluation results in Section 4 to support our conjecture.

3.3. Inject Backdoor in Diffusion Models

Prior attacks leveraged similar backdoor injection mechanisms with slight differences in sampler modifications [23–25]. UIBDiffusion employs the scheme presented in VillanDiffuision [25] since it provides a unified framework to implant backdoors in various types of diffusion models and samplers and achieves outstanding performance. We briefly introduce the injection process in this section.

Clean diffusion process. Clean diffusion consists of forward and backward processes. The forward diffusion process diffuses images x from data distribution $q(\mathrm{x}_0)$ to the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ by progressively adding noise in each time step along a Markov chain. The process can be mathematically expressed as: $q(x_t|x_0) = \mathcal{N}(\alpha(t)x_0, \beta^2(t)\mathbf{I})$ with (1) $q(x_{T_{\text{max}}}) \approx \mathcal{N}(0, \mathbf{I})$, (2) $q(x_{T_{min}}) \approx q(x_0)$ and $x_t, t \in [T_{\text{min}}, T_{\text{max}}]$, where T_{min}, T_{max} are the first and last timestep, and $\alpha(t), \beta(t)$ are parameterized variables introduced in the original diffusion model work [1].

Backward diffusion process reverse the Markov chain and generate images from a sampled Gaussian noise by gradually removing noises in each timestep and can be expressed as: $q(x_t|x_{t-1}) = \mathcal{N}(k_t x_{t-1}, w_t^2\mathbf{I})$, and $x_t(x_0, \epsilon) = \hat{\alpha}(t)x_0 + \hat{\beta}(t)\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $k_t, w_t, \hat{\alpha}(t), \hat{\beta}(t)$ are the same parameterized variables for the reverse process.

Backdoor and trigger injection. We employ the VillanDiffusion framework for backdoor injection and modify the trigger injection process in data poisoning. VillanDiffusion proposes a unified backdoor injection framework by introducing the loss function shown in Equation 4, where $\mathcal{L}_c$ and $\mathcal{L}_p$ are used to optimize the original generation and backdoor attack tasks, respectively. r is the poisoned data, y is the backdoor target, ζ is a parameter to control backdoor against different samplers (e.g., ODE and SDE), and η_c, η_p are weights to balance $\mathcal{L}_c$ and $\mathcal{L}_p$.

$$\begin{aligned} &\mathcal{L}_\theta(\eta_c, \eta_p, x, t, \epsilon, r, y, \xi) = \\ &\eta_c \mathcal{L}_c(x, t, \epsilon) + \eta_p \mathcal{L}_p(x, t, \epsilon, r, y, \zeta). \end{aligned} \tag{4}$$

We then explain the difference in the trigger injection process between VillanDiffusion and UIBDiffusion. VillanDiffusion injects the backdoor trigger following the method in Equation 5, where $\mathbf{M}$ is a mask and g is the trigger with specific patterns (e.g., gray box and a pair of glasses), and $\odot$ is the element-wise multiplication.

$$r(x, g) = \mathbf{M} \odot g + (1 - \mathbf{M}) \odot x. \tag{5}$$

To better align with the adversarial perturbation, we inject our trigger using the approach in Equation 6, where τ is the imperceptible trigger synthesized by our proposed trigger generator and ε is a variable that controls the strength of the added trigger. In contrast to triggers with recognizable patterns, the UIBDiffusion triggers are generated by a trainable generator with subtle perturbations and hard-to-detect patterns, which makes them effective and stealthy. The backdoor and trigger injection process is summarized in Algorithm 2.

$$r(x, \tau) = x + \varepsilon \odot \tau \tag{6}$$

Algorithm 2 UIBDiffusion Backdoor And Trigger Injection

Require: Backdoor Image Trigger τ, Benign Dataset X, Backdoor Target y, Training parameters θ, Sampler Randomness ζ, Strength ε

Backdoor DMs with Generated Trigger:
while not converge **do**
  $\{x, \eta_c, \eta_p\} \sim X$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  Trigger injection: $r(x, \tau) = x + \varepsilon \odot \tau$
  update $\theta : \nabla_\theta \mathcal{L}_\theta^I(\eta_c, \eta_p, x, t, \epsilon, r, y, \zeta)$
end while

4. Experiments

4.1. Experimental Settings

Datasets and models. We follow the same practice in existing backdoor attacks against DMs [23–25] and use CIFAR-10 (32 × 32) [68] and CelebA-HQ (256 × 256) [69] datasets for fair comparison. We evaluate UIBDiffusion on three DMs: DDPM, LDM and NCSN, with eleven different samplers. We adopt the same training recipe from [25] and conduct our attack on pre-trained models (i.e., *google/ddpm-cifar10-32* for CIFAR-10 and *CompVis/ldm-celebahq-256* for CelebA-HQ) downloaded from Google HuggingFace organization and HuggingFace, respectively. Training hyper-parameters are provided in the Appendix.

Attack configurations and baselines. We conduct our experiments on NVIDIA L40S GPUs and employ VillanDiffusion [25] as the backdoor injection framework. We provide two generated triggers guided by VGG [70] and ResNet [71] for CIFAR-10 and one proof-of-concept trigger for CelebA-HQ. We adopt "hat" and "cat" as targets for CIFAR-10 and "cat" as the target for CelebA-HQ, respectively. BadDiffusion [23] and VillanDiffusion [25] are considered as the baseline attacks for utility comparison and Elijah [26] and TERD [27] as the defenses for undetectability evaluation.

Evaluation metrics. We comprehensively evaluate the attack and generation performance of UIBDiffusion using four standard metrics. We exploit FID [72] score to examine the quality of generated clean images. Images with

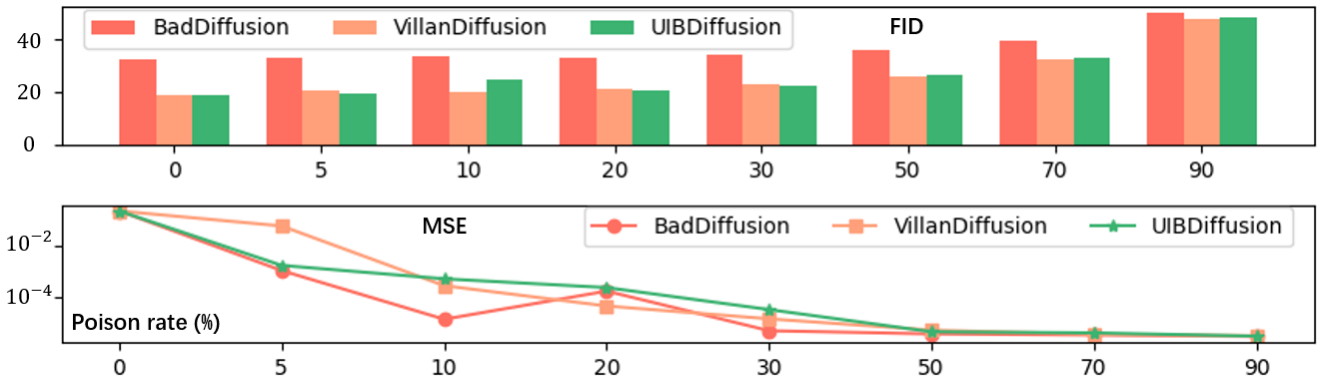

Figure 3. FID and MSE comparison of BadDiffusion, VillanDiffusion, and UIBDiffusion over different poison rates.

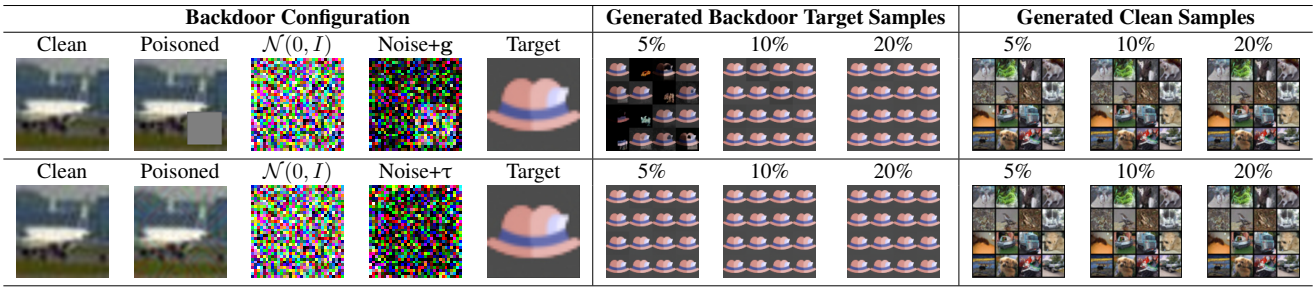

Table 1. ASR comparison between VillanDiffusion (top row) and UIBDiffusion (bottom row). UIBDiffusion achieves superior performance than the prior work at a low poison rate (e.g., 5%).

better quality tend to have lower FID scores. As for the attack effect, we employ Attack Success Rate (ASR), Mean Square Error (MSE) and Structural Similarity Index Measure (SSIM) for assessment. ASR is defined as the percentage of the target image generated from sampled noise stamped with backdoor triggers. Higher ASR indicates better utility. MSE is a metric to measure the pixel-level distance between the authentic and synthetic target images. A successful attack is expected to achieve low MSE. SSIM is a similar metric used to measure the similarity between two images (i.e., the real and generated target images). Superior attack performance will yield SSIM that approaches the value of 1.

4.2. Results on Different Diffusion Models

We first evaluate UIBDiffusion on different DMs. We present the results of DDPM in the main paper and organize the results of LDM and NCSN in the Appendix. As shown in Fig. 3, Table 1 and Table 2, UIBDiffusion achieves comparable FID as the clean diffusion model (i.e., poison rate = 0%). It can be seen that with the poison rates increased, all backdoor attacks yield a lower MSE, demonstrating a positive correlation between poison rates and attack effect. However, the generation ability of DMs will be accordingly affected when the poison rate is high. Notably, UIBDiffusion can reach a 100% ASR with only 5% poison rate while VillanDiffusion gets 56% ASR at this rate. In practice, attackers can achieve the expected attack effect without undermining the original generation ability of DMs, which further improves the stealthiness of the backdoor attack.

Poison Rate	0%	1%	2%	3%	4%	5%	10%	$\geqq$ 20%
BadDiffusion [23]	3.80E-4	6.22E-3	1.48E-2	1.10E-2	0.37	0.99	0.99	0.99
VillanDiffusion [25]	3.80E-4	3.92E-4	1.25E-2	1.69E-2	0.60	0.99	0.99	0.99
UIBDiffusion (Ours)	3.80E-4	**0.47**	**0.96**	**0.99**	**0.99**	0.99	0.99	0.99

Table 2. SSIM comparison of BadDiffusion, VillanDiffusion and UIBDiffusion. UIBDiffusion achieves higher SSIM at lower poison rates of 1%∼4%.

4.3. Results on Different Samplers

We then show the effectiveness of UIBDiffusion on different samplers. We conduct the experiments using DDPM model with 11 types of samplers. As shown in Fig. 4, UIBDiffusion achieves a consistent performance of 100% ASR across all samplers and yield a robust target image generation effect (i.e., MSE and SSIM) at various poison rates. We also observe that the ODE solvers outperform their SDE solver counterparts.

4.4. Resilience against SOTA Defenses

Elijah [26] and TERD [27] are two SOTA defenses against backdoor attacks on DMs. They employ a similar idea of trigger inversion to estimate and reverse engineer the triggers. They then apply backdoor detection and removal

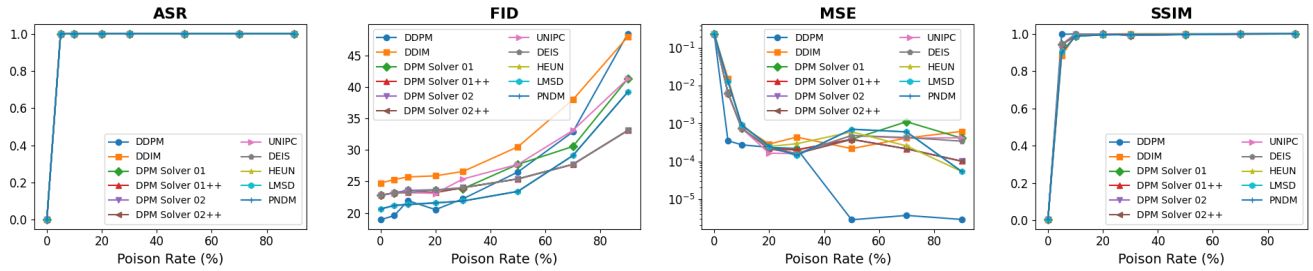

Figure 4. UIBDiffusion performance against eleven different samplers across various poison rates. UIBDiffusion yields consistently high ASR, SSIM, and MSE and even better FID on ODE-based samplers compared to SDE-based samplers.

based on the reconstructed triggers. We show that UIBDiffusion is highly resilient to both attacks since our triggers are hard to remodel. Fig. 5 demonstrates the performance of UIBDiffusion against Elijah compared to BadDifufsion and VillanDiffusion. It can be seen that UIBDiffusion achieves 100% ASR and keeps the same MSE and SSIM before and after the backdoor detection and removal while ASR and SSIM of BadDiffusion and VillanDiffusion drop to 0% after detection, indicating superior resilience against the SOTA defense. We then present the robustness of UIBDiffusion across various poison rates. Higher poison rates usually yield higher ASR but are easier to detect. However, as shown in Table 3, UIBDiffusion managed to escape detection at all the evaluated poison rates while maintaining 100% ASR and high performance of SSIM and FID. It is worth noting that a $\Delta SSIM = 0$ indicates that UIBDiffusion successfully bypasses the defense, achieving 100% targeted image generation. We attribute the fluctuations in FID to Elijah's use of an ineffective inverted trigger during backdoor removal.

	BadDiffusion * [23]				UIBDiffusion (Ours)			
Poison Rate	Detected	ASR	ΔSSIM	ΔFID	Detected	ASR	ΔSSIM	ΔFID
5%	100%	0%	-1.0	-0.07	0%	**100%**	0	-0.98
10%	100%	0%	-1.0	+0.18	0%	**100%**	0	+0.15
20%	100%	0%	-1.0	-0.03	0%	**100%**	0	+5.72
30%	100%	0%	-1.0	+0.15	0%	**100%**	0	+12.59
50%	100%	0%	-1.0	+0.20	0%	**100%**	0	+11.56
70%	100%	0%	-1.0	-0.11	0%	**100%**	0	+14.85
90%	100%	0%	-1.0	-0.07	0%	**100%**	0	+5.09

Table 3. Elijah against BadDiffusion and UIBDiffusion with different poison rates. "*" denotes results replicated from [26].

We evaluate the quality of the inverted trigger by visualizing the inverted triggers and testing if they could activate the backdoor as the actual triggers when imposed on benign sampled noise. Results are shown in Fig. 6. Sampled noise with inverted triggers should generate the target image if the defense succeeds and generate normal images otherwise. It can be seen that for triggers that have specific patterns (e.g., a stop sign and a gray box), Elijah can successfully estimate the authentic triggers and the inverted triggers are able to activate backdoors. On the other hand, the inverted trigger of UIBDiffusion fails to activate the backdoor injected in the DM, showing that the defense is ineffective against UIBDiffusion. We argue this is because the UIBDiffusion trigger is adapted from UAP, which is small and unstructured noise that is difficult to approximate based on the current trigger inversion loss function in Elijah and TERD.

UIBDiffusion performs a similar undetectability against TERD as shown in Table 4. It achieves a 100% True Positive Rate (TPR) and 0% False Positive Rate (FPR), where TPR and FPR represent the percentage of the benign or backdoor sampled noise that is correctly detected. On the other hand, triggers with obvious patterns fail to bypass the trigger inversion algorithm and are all detected by TERD. Although TRED offers a more precise trigger inversion approach using a unified loss and trigger refinement, it is still ineffective against our proposed UIBDiffusion trigger.

Trigger-Target	TPR (↑)	FPR (↓)
Ours-HAT	100%	**0%**
BOX-HAT*	100%	100%
BOX-SHOE*	100%	100%
BOX-CORNER*	100%	100%
STOP-HAT*	100%	100%
STOP-SHOE*	100%	100%
STOP-CORNER*	100%	100%

Table 4. Resilience of UIBDiffusion against TERD. "*" denotes results replicated from [27].

4.5. Comparison of Trigger Generation Approach

We evaluate the necessity and effectiveness of our proposed trigger generation approach by comparing it to the original UAPs and GUAPs. We employ the generated UIBDiffusion trigger, the original UAP from [48] and the GUAP from [66] to poison the benign dataset and backdoor the DMs using these three triggers, respectively. As can be seen from Table 5, the UIBDiffusion trigger outperforms the original UAP and GUAP triggers at all four metrics by large margins, achieving much better attack and generation performance. The rationale behind this is that our trigger introduces an evident distribution shift. In contrast, the original UAP trigger only slightly affects the sampled noise distribution, and the spatial transformation of GUAP is image-specific and imposes a negative impact on the backdoor im-

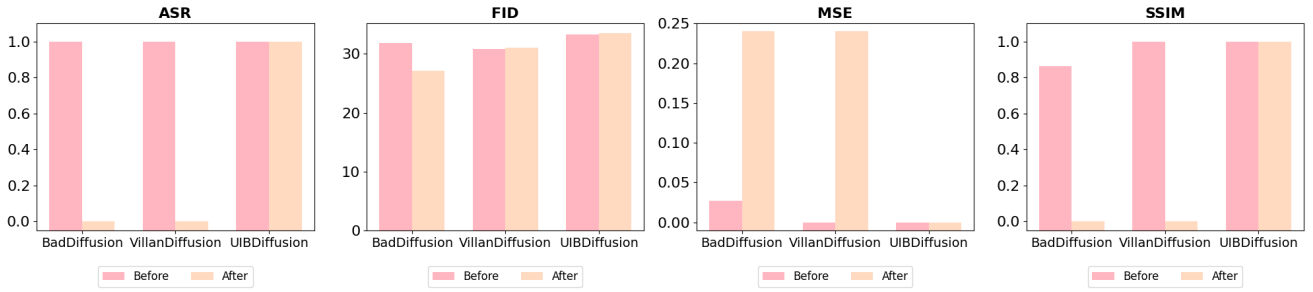

Figure 5. Performance comparison of BadDiffusion, VillanDiffusion and our UIBDiffusion before and after the Elijah defense. ASR: the higher the better; FID: the lower the better: MSE: the lower the better; SSIM: the higher the better.

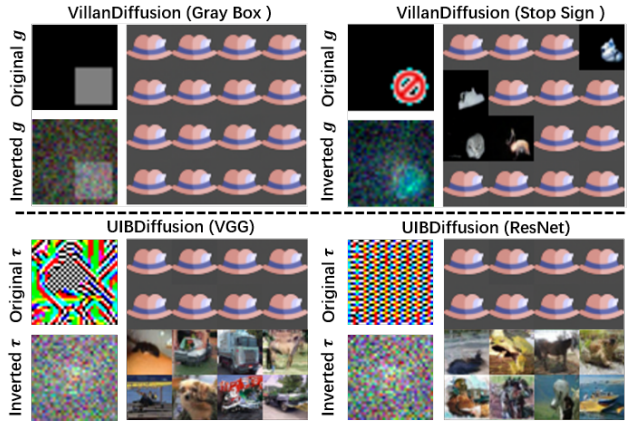

Figure 6. Visualization of trigger inversion. Inverted triggers of gray box and stop sign can activate backdoor, indicating a successful estimation of the actual triggers. On the other hand, inverted triggers of UIBDiffusion fail to activate the injected backdoor.

plant. This shows the superiority and necessity of our trigger design choice.

	FID (↓)	MSE (↓)	SSIM (↑)	ASR (↑)
Original UAP	20.88	9.22E-2	0.6190	43.75%
Original GUAP	41.87	7.86E-2	0.1932	0.00%
UIBDiffusion	**19.77**	**5.30E-4**	**0.9967**	**100%**

Table 5. Trigger performance comparison across the original UAP, GUAP and UIBDiffusion.

4.6. Trigger Robustness

Subtle adversarial perturbations are susceptible to smoothing defenses [73] and could be mitigated by random noise. Thus, we perform a robust evaluation of the UIBDiffusion triggers and present the results in Table 6. UIBDiffusion shows strong robustness against random noise. We argue that the resilience is due to our trigger generation algorithm enhancing the adversarial perturbations via co-optimizing the additive and non-additive noise.

Random Noise	1%	2%	5%	10%	20%
ASR	99%	99%	100%	100%	100%
FID	19.48	19.52	19.80	19.37	19.49
MSE	2.36E-3	1.13E-3	8.03E-4	3.02E-4	5.30E-4
SSIM	0.9865	0.9929	0.9944	0.9969	0.9957

Table 6. Robustness of UIBDiffusion against random noise.

4.7. Quantitative Evaluation of Imperceptibility

We also quantitatively evaluate the imperceptibility of our trigger design using distance metrics of l_∞ and l_2 norms. We compute the distance between a poisoned image (stamping with our trigger and triggers used in prior works) and a benign image and summarize the results in Table 7. Our imperceptible triggers have much lower l_∞ and l_2 distances compared to triggers with specific patterns, which validates the stealthiness at the image level (i.e., the pixel space) of the UIBDiffusion trigger design.

	Cifar 10		CelebA-HQ-256	
	l_∞	l_2	l_∞	l_2
UIBDiffusion Trigger	**0.04**	**1.80**	**0.05**	**12.74**
Prior Work	0.47	6.21	0.80	19.84

Table 7. Quantitative comparison of triggers imperceptibility using distance metrics of l_∞ and l_2 norms.

5. Conclusion

In this work, we present UIBDiffusion, the first imperceptible backdoor attack against diffusion models. We propose a novel trigger generation method based on universal adversarial perturbations and demonstrate that such perturbations for attacking discriminative models can be adapted as general, effective, and stealthy triggers for backdoor attacks against diffusion models. We empirically reveal the underlying rationale why UIBDiffusion can achieve a high attack success rate while evading the trigger inversion defensive algorithms. We hope our work sheds light on the potential risks that diffusion models face and motivates future research to develop more comprehensive defenses.

Acknowledgment

This work is supported in part by the National Science Foundation (NSF) SaTC-2426299 and SaTC-2413046.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 5

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. 3

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 10674–10685. IEEE, 2022. 1, 3

[4] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 4172–4182, 2023. 1

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net, 2024. 1

[6] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[7] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 3

[8] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 1

[9] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 1

[10] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023.

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[12] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*, 2021. 1

[13] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022.

[14] Heeseung Kim, Sungwon Kim, and Sungroh Yoon. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, pages 11119–11133. PMLR, 2022.

[15] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[16] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

[17] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 1

[18] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. *Advances in neural information processing systems*, 29, 2016. 1

[19] Mengchen Zhao, Bo An, Yaodong Yu, Sulin Liu, and Sinno Pan. Data poisoning attacks on multi-task relationship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] Scott Alfeld, Xiaojin Zhu, and Paul Barford. Data poisoning attacks against autoregressive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[21] Bingyin Zhao and Yingjie Lao. Towards class-oriented poisoning attacks against neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3741–3750, 2022.

[22] Bingyin Zhao and Yingjie Lao. Clpa: Clean-label poisoning availability attacks using generative adversarial nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9162–9170, 2022. 1

[23] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023. 1, 3, 4, 5, 6, 7

[24] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023. 3

[25] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 5, 6

[26] Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing

Ma, Pin-Yu Chen, Tsung-Yi Ho, et al. Elijah: Eliminating backdoors injected in diffusion models via distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10847–10855, 2024. 1, 2, 3, 5, 6, 7

[27] Yichuan Mo, Hui Huang, Mingjie Li, Ang Li, and Yisen Wang. TERD: A unified framework for safeguarding diffusion models against backdoors. In *Forty-first International Conference on Machine Learning, ICML 2024*, 2024. 1, 2, 3, 5, 6, 7

[28] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 3

[29] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, abs/1712.05526, 2017.

[30] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS*. The Internet Society, 2018. 1, 3

[31] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019*, pages 2041–2055. ACM, 2019. 1

[32] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision - ECCV 2020, Proceedings, Part X*, volume 12355 of *Lecture Notes in Computer Science*, pages 182–199. Springer, 2020.

[33] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16463–16472, 2021.

[34] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *9th International Conference on Learning Representations, ICLR*, 2021.

[35] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence*, pages 1148–1156. AAAI Press, 2021.

[36] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11966–11976, 2021.

[37] Khoa Doan, Yingjie Lao, and Ping Li. Backdoor attack with imperceptible input and latent modification. *Advances in Neural Information Processing Systems*, 34:18944–18957, 2021.

[38] Khoa D Doan, Yingjie Lao, and Ping Li. Marksman backdoor: Backdoor attacks with arbitrary target class. *Advances in Neural Information Processing Systems*, 35:38260–38273, 2022. 1

[39] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 11957–11965. AAAI Press, 2020. 1

[40] Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022. 1

[41] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14900–14912, 2021. 1

[42] Qi Zhao and Christian Wressnegger. Adversarially robust anti-backdoor learning. In *Proceedings of the 2024 Workshop on Artificial Intelligence and Security, AISec 2024*, pages 77–88. ACM, 2024.

[43] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy, SP 2019*, pages 707–723. IEEE, 2019.

[44] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 16913–16925, 2021. 3

[45] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[46] Bingyin Zhao and Yingjie Lao. Ultraclean: A simple framework to train robust neural networks against backdoor attacks. *arXiv preprint arXiv:2312.10657*, 2023. 1

[47] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. 2, 3

[48] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2, 3, 4, 7

[49] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 11895–11907, 2019. 3

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

[51] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020. 3

[52] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023. 3

[53] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023. 3

[54] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, 2022. 3

[55] 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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, 2022. 3

[56] Joseph Clements and Yingjie Lao. Hardware trojan attacks on neural networks. *arXiv preprint arXiv:1806.05768*, 2018. 3

[57] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023*, pages 1577–1587. ACM, 2023. 3

[58] Yihao Huang, Qing Guo, and Felix Juefei-Xu. Zero-day backdoor attack against text-to-image diffusion models via personalization. *CoRR*, abs/2305.10701, 2023.

[59] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 4561–4573. IEEE, 2023. 3

[60] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19)*, volume 2301 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. 3

[61] Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, pages 16453–16461. IEEE, 2021. 3

[62] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 3

[63] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Workshop Track Proceedings*, 2017.

[64] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 3905–3911. ijcai.org, 2018. 3

[65] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 4

[66] Yanghao Zhang, Wenjie Ruan, Fu Wang, and Xiaowei Huang. Generalizing universal adversarial attacks beyond additive perturbations. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1412–1417. IEEE, 2020. 4, 7

[67] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[68] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 5

[69] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 3730–3738. IEEE Computer Society, 2015. 5

[70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 5

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[72] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6626–6637, 2017. 5

[73] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33:21945–21957, 2020. 8