Concealed Backdoor Attack on Diffusion Models for Smart Devices with Non-standard Gaussian Distribution Noise

Jiaxing Li, Yu-an Tan, Sizhe Fan, Fan Li, Xinyu Liu, Runke Liu, Yuanzhang Li, Weizhi Meng, *Senior Member, IEEE*

***Abstract*—Edge AI-driven diffusion models (DMs) are increasingly integrated into consumer devices for high-quality data generation and content creation. This paper introduces InvisibleDiffusion, a novel backdoor attack framework for diffusion models in consumer electronics, designed to remain undetected by utilizing a non-standard Gaussian distribution as a concealed trigger. Unlike previous backdoor methods, InvisibleDiffusion does not rely on obvious visual triggers, enhancing its stealthiness. Extensive experiments demonstrate that InvisibleDiffusion achieves high attack efficacy against DDPM and DDIM models on CIFAR-10 and CelebA datasets, while maintaining the functional integrity of the models. Our code is available for reproducibility at https://anonymous.4open.science/r/b2hoaWNhbnRzZWV0aGF0bm9vb29vb29vb29v.**

***Index Terms*—Generative artificial intelligence, Edge AI, Consumer devices, Security in deep learning.**

I. Introduction

In recent years, diffusion models (DMs) have developed into cutting-edge tools in content creation and high-quality generation of comprehensive data, covering multiple fields such as images, text, speech, molecules, etc. These models leverage deep neural networks and large-scale training data to demonstrate outstanding performance [1]–[14]. With the advent of Edge AI, diffusion models are increasingly integrated into consumer devices [15]–[17], enabling real-time data processing and personalized user experiences. This widespread adoption has amplified the importance of ensuring the security and privacy of these models, particularly in devices that handle sensitive user data. In consumer devices, these models power a range of functionalities, from intelligent voice assistants and personalized content recommendations to advanced security features.

With the widespread application of diffusion models, concerns about the risk of DMs suffering from backdoor attacks have rapidly escalated [18]–[21]. Specifically, an attacker can train a model to perform a specified behavior when a trigger is activated, but when the trigger is deactivated, the same model runs normally without being tampered with. The stealth nature of this backdoor attack makes it difficult for the average user to determine whether a model is risky or safe to use. It is worth noting that existing works related to backdoor attacks [18]–[20] on DMs have limitations: they must rely on an obvious image as the trigger as the correction term. This obvious trigger is added to the diffusion process, making the spread of backdoor attacks significantly different from the normal diffusion process. Such backdoor attacks are easily detectable, which may lead to an underestimation of the risk assessment of diffusion models. In the context of consumer electronics, such vulnerabilities could lead to severe breaches of user privacy and data integrity, compromising the trustworthiness of AI-driven devices.

To bridge this gap, we propose InvisibleDiffusion, a new backdoor attack framework for diffusion models. Unlike previous backdoor attacks that must use an obvious image as a trigger to be injected into the training process, our proposed method can backdoor the diffusion model using an invisible backdoor trigger. As shown in Figure 1, our proposed backdoor attack diffusion process maps target data into a non-standard Gaussian distribution activated by the trigger. We then apply a new parameterized generative process to learn to reverse the backdoor diffusion process through an efficient training objective. After training, the backdoor attack model always outputs hostile targets along the learned backdoor generation process. In particular, as shown in the our method of Figure 1, based on whether the trigger is fixed or not, we consider two types of attacks: attacks based on fixed trigger (with fixed image) and attacks based on trigger from a newly sampled Gaussian distribution (without fixed image). Attacks based on fixed triggers consider a variety of pictures as triggers. The backdoor diffusion model can successfully attack theoretically any picture as a trigger based on Gaussian distribution sampling, Uniform distribution sampling, Poisson distribution sampling, etc. As for the attack based on the trigger of the newly sampled Gaussian distribution, based on the consideration of the concealment of the trigger, we directly newly sample the standard Gaussian distribution as the trigger and mix it with the original clean Gaussian noise to form a new non-standard Gaussian distribution with a mean value of 0. This Noise successfully attacks the diffusion model.

Through extensive experimental verification, InvisibleDiffusion has achieved high attack performance against DDPM and DDIM on CIFAR-10 and CelebA datasets and maintains good concealment. For example, on the CelebA data set, InvisibleDiffusion can achieve an attack accuracy and attack

(Corresponding author: Yuanzhang Li.)

Jiaxing Li, Yu-an Tan, Sizhe Fan, Fan Li, Xinyu Liu, Runke Liu are with School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China.

Yuanzhang Li is with School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China.

Weizhi Meng is with School of Computing and Communications, Lancaster University, United Kingdom.

success rate of 84.70% respectively in attacks based on fixed triggers. Also in the case of benign diffusion, the model performs similarly to a clean (not tampered with) diffusion model. Our main contributions are as follows:

- We propose InvisibleDiffusion, a novel backdoor attack framework for the diffusion model. In this attack framework, we define the fusion of trigger information and original clean Gaussian noise as a non-standard Gaussian distribution.
- We propose a backdoor attack diffusion process with a new diffusion adversarial target to a non-standard Gaussian distribution, and a backdoor attack diffusion process based on new parameterization, thereby achieving the simple training goal of InvisibleDiffusion.
- Experimental evaluation shows that in terms of diffusion model evaluation indicators, we considered two types of triggers, and InvisibleDiffusion achieved superior attack performance for the diffusion model on two benchmark datasets. Also in the case of benign diffusion, the model performs similarly to the clean diffusion model.

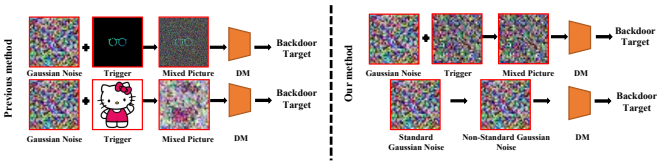

Fig. 1: InvisibleDiffusion: our proposed backdoor attack framework for diffusion model (DM). Compared with the previous method, the injection of triggers in our proposed method should be as hidden as possible. In particular, our proposed method does not rely on an explicit image as a trigger.

II. Related Work

A. Diffusion Models

Diffusion models (DMs) first attracted great attention in the field of image generation and proved to be a powerful tool for synthesizing visually appealing images. They have been applied to various image-related tasks, including image generation [3], [22]–[28], image editing [29]–[32], and audio and video synthesis [33]–[35]. Diffusion models have also been successful as generative tools in other fields [36]–[40]. Despite their success in a wide range of generative domains, their security has been less studied, especially in terms of how to attack or defend against them. As diffusion models become increasingly popular, this becomes a critical issue. In essence, diffusion models aim to learn the reverse diffusion process, a concept derived from a well-studied forward destruction process. Unlike some other image generation models that require specialized architectural design (e.g., flow-based models), diffusion models [25] exploit the reversibility of the diffusion process. However, a significant drawback of the diffusion model is its relatively slow generation process. The latest research in this field mainly focuses on solving this limitation, with technologies such as DDIM [23] and Analytic-DPM [22] and DPM Solver [41] aiming to speed up sampling. The specific focus of this article is on the covert introduction of backdoors into DDPM. This is an exploration of the vulnerabilities of the diffusion model, particularly in the context of backdoor attacks, and highlights the need for a deeper understanding of its security implications.

B. Backdoor Attack on Diffusion Models

As the diffusion model becomes more widely used, some researchers have proposed work on the backdoor attack diffusion model [18], [19]. The approach taken by TrojDiff [18] assumes that the attacker has access to the training program and sampler, and applies correction terms on DDPM and DDIM to launch the attack. BadDiffusion [19] method performs a backdoor attack on DDPM by introducing an additional correction term to the mean of the forward diffusion process without any modification to the sampler. VillanDiffusion [20] extends the BadDiffusion approach to different samplers and adds a text-as-trigger attack for text-to-image generation. Another study [21] put a backdoor on text-to-image DMs by modifying the text encoder rather than the DMs themselves. Different from this, our method provides a unified framework for attacking DDPM, in which correction terms of previous attack methods are not necessary.

Compared to traditional backdoor attack methods, InvisibleDiffusion introduces a unique approach that leverages non-standard Gaussian noise to maintain high stealth in edge device applications. Most existing methods rely on explicit image triggers to activate backdoor behaviors, which can be more easily detected by standard security protocols. In contrast, InvisibleDiffusion can operate covertly without an explicit image-based trigger, which significantly enhances its ability to evade detection in consumer devices. Moreover, the adaptability of our approach allows it to integrate seamlessly into various device scenarios, where maintaining user data integrity is essential.

III. InvisibleDiffusion:Methods and Algorithms

In this section, we first introduce our threat model, then we briefly review the DDPM model and introduce some necessary notation. Then we give two requirements of our attack, propose our non-standard Gaussian distribution attack method, and give our backdoor diffusion process. Finally, we present our loss function.

A. Threat Model

As diffusion models become successful on more tasks and the amount of data they utilize continues to expand, more and more users are choosing to use checkpoints trained by model developers as a basic tool to perform their own tasks. Model developers releasing more and more trained checkpoints provide scenarios for backdoor attacks on these models. We here give the attacker's goals, the attacker's capabilities.

1) Attacker's goals: For image diffusion models, the main use of normal diffusion models is to use standard Gaussian noise as model input to denoise blurred images. The attacker's goal is to release a backdoored model with the dual purpose of: (a) taking clean noise as input to generate images from

the normal data distribution $q(x)$; and (b) taking noise with a backdoor trigger as input, generate an image from the target distribution $\tilde{q}(x)$.

2) Attacker's capacity: We assume that the attacker has the following capabilities: (1) Able to control the diffusion process of backdoor attacks and generate images from the target distribution $\tilde{q}(x)$ (note that the diffusion process $\mathcal{N}(0, I) \leftarrow q(x)$ defined in DDPM/DDIM) is now called the benign diffusion process); (2) being able to control the training process so that the diffusion model learns the benign and backdoor attack diffusion generation processes according to the corresponding training procedures; (3) designing the backdoor trigger sampling process for the backdoor attack noise input. The attacker will then provide the user with the diffusion model into which the backdoor has been implanted (i.e., the trained parameters θ). The user will use a benign sampling process (i.e., the sampling mechanism of DDPM/DDIM) to generate images, unaware of the fact that an attacker can activate a covert backdoor through a trigger and control the generated image.

B. Denoising Diffusion Probabilistic Model

In order to illustrate the way InvisibleDiffusion injects backdoors by changing the training loss of the diffusion model, the remainder of this article will use the DDPM (Denoising Diffusion Probabilistic Model) as the target diffusion model. Given a real image sample $q(x_0)$, the diffusion forward process adds Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to it through T times of accumulation. As t increases, x_t approaches pure noise. $q(x_t|x_{t-1})$ represents a Gaussian distribution with the previous state x_{t-1} as the mean, and xt is sampled from this Gaussian distribution. The so-called forward diffusion process can be understood as a Markov chain, that is, by gradually adding Gaussian noise $\mathcal{N}(0, I)$ to a real picture until it finally becomes a pure Gaussian noise picture $x_T \sim \mathcal{N}(0, I)$.

The size of each step is controlled by a series of hyperparameters β_t for the variance of the Gaussian distribution. That is, x_t at each time step is sampled from a Gaussian distribution with $\sqrt{1-\beta_t}$ multiplied by x_{t-1} as the mean and β_t as the variance, where β_t is a series of fixed values.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

When sampling x_t, it is not directly sampled through Gaussian distribution $q(x_t|x_{t-1})$, but a parameterization technique is used. Using the notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we have:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{2}$$

By reversing the above process and sampling, the original image distribution $x_0 \sim q(x)$ can be restored from Gaussian noise $x_T \sim \mathcal{N}(0, I)$. If satisfies the Gaussian distribution and β_t is small enough, $q(x_t|x_{t-1})$ it is still a Gaussian distribution. Use a deep learning model (with parameters θ) to predict such an inverse distribution p_θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Since the forward process has Markov properties, $q(x_t|x_{t-1}, x_0)$ it is actually equivalent to $q(x_t|x_{t-1})$.

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I) \tag{4}$$

From Bayes' theorem and Gaussian distribution probability density function, we have $\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$ and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. Put $x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t)$ into $\tilde{\mu}_t(x_t, x_0)$, $\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t)$.

Diffusion Models use maximum likelihood estimation to find the probability distribution of Markov chain transitions in the inverse diffusion process. That is, maximizing the log-likelihood of the model prediction distribution, from the perspective of loss reduction, is to minimize the negative log-likelihood. The variational lower bound (VLB) can be used to optimize the negative log-likelihood. $p_\theta(x_{t-1}|x_t)$ is the target distribution expected to be fitted by the deep neural network. According to Equation 3 ,the loss function is:

$$L = \mathbb{E}_q\left[\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\right] \tag{5}$$

$$= \mathbb{E}_{x_0,\epsilon}\left[\|\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon) - \mu_\theta(x_t, t)\|^2\right]. \tag{6}$$

After removing the constant term, the loss function becomes:

$$L = \mathbb{E}_{x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]. \tag{7}$$

C. Invisible Backdoored Diffusion Process

We reconsider the forward and backward processes of DDPM and give two requirements for attacking DDPM:

(1) Based on the consideration of the success rate of backdoor attacks, the backdoor noise input must be a non-standard Gaussian distribution different from benign noise;
(2) Based on the concealment of backdoor attacks, the injection of triggers should be as concealed as possible.

For requirement (1), we reconsider the diffusion process and reverse denoising process of DDPM, and expand the restriction condition for successful attack to the fact that the backdoor noise input must have a non-standard Gaussian distribution different from benign noise. Generally, there are two types of triggers here. One is a fixed trigger, which is a picture that is mixed with the original clean Gaussian noise in a certain proportion to form a backdoor noise input. At the same time, based on the consideration of requirement (2), we try to be "invisible" when selecting pictures, such as a picture based on Gaussian distribution sampling, Uniform distribution sampling, and Poisson distribution sampling, which are also noise, as a trigger. The other is a trigger of a newly sampled Gaussian distribution, which is mixed with the original clean Gaussian noise in a certain proportion to form a backdoor noise input. Also based on the consideration of requirement (2), when we choose a new sampling Gaussian distribution, we choose a Gaussian distribution with a mean value of 0 as the trigger, such as the standard Gaussian distribution as the trigger. Mixing λ times of these two triggers with $(1-\lambda)$ times of the original clean Gaussian noise will form a non-standard

Gaussian distribution $\mathcal{N}(\mu, h^2 I)$ that is different from benign noise. Here, the trigger is represented by ϕ. λ represents the proportion of trigger mixing, $\lambda \in (0,1)$. $\mathcal{N}(\mu, h^2 I)$ is the backdoor noise input, and h^2 is the coefficient of the Gaussian distribution noise variance.

1) Backdoored diffusion process: Firstly, we explain how a benign diffusion process diffuses $q(x)$ into $\mathcal{N}(0, I)$ over T time steps. Then, we propose a backdoor attack diffusion process with a new transition process to spread $\tilde{q}(x)$ to $\mathcal{N}(\mu, h^2 I)$. Given a variance plan $\{\beta_t\}_{t=1}^T$, $\bar{\alpha}_T \approx 0$, provided in DDPM. Therefore $x_T = \sqrt{\bar{\alpha}_T} x_0 + \sqrt{1-\bar{\alpha}_T}\epsilon \approx \epsilon$, represents $x_T \sim \mathcal{N}(0, I)$. Under the same variance scheme, we now consider x_t to have the following form.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \lambda(\sqrt{1-\bar{\alpha}_t}\phi) + (1-\lambda)(\sqrt{1-\bar{\alpha}_t}\epsilon) \tag{8}$$

where ϕ represents the trigger we injected, λ represents the proportion of trigger mixing, $\lambda \in (0,1)$, $\epsilon \sim \mathcal{N}(0, I)$ is the original clean standard Gaussian noise. At time step T, $x_T = \sqrt{\bar{\alpha}_T} x_0 + \lambda(\sqrt{1-\bar{\alpha}_T}\phi) + (1-\lambda)(\sqrt{1-\bar{\alpha}_T}\epsilon) = \lambda\phi + (1-\lambda)\epsilon$. Therefore, $x_T \sim \mathcal{N}(\mu, h^2 I)$

For the first fixed trigger, ϕ is a fixed picture. In order to ensure that x_t can be expressed by Equation 8, we propose a backdoor attack diffusion process:

$$\tilde{q}(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1} + \lambda m_t \phi, \beta_t(1-\lambda)^2 I) \tag{9}$$

Here m_t is a function about time step t, representing the coefficient by which each trigger ϕ is multiplied at time step t. According to Equation 8, we have $\sqrt{1-\bar{\alpha}_t} = m_1\sqrt{\alpha_t\alpha_{t-1}...\alpha_2} + m_2\sqrt{\alpha_t\alpha_{t-1}...\alpha_3} + ...m_{t-1}\sqrt{\alpha_t} + m_t$. We can get the numerical solution for m_t by substituting the value α_t from time step $t = 1$ to $t = T$.

For the second newly sampled Gaussian distribution trigger, ϕ is a standard Gaussian distribution with a mean value of 0, $\phi \sim \mathcal{N}(0, I)$. The backdoor attack diffusion process is:

$$\tilde{q}(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, \beta_t(2\lambda^2 - 2\lambda + 1) I) \tag{10}$$

Here ϕ is the standard Gaussian distribution $\phi \sim \mathcal{N}(0, I)$. The original clean noise is also the standard Gaussian distribution, $\epsilon \sim \mathcal{N}(0, I)$. According to the Gaussian distribution probability density function, the mixed backdoor noise input is sampled from $\mathcal{N}(0, (2\lambda^2 - 2\lambda + 1)I)$. Detailed mathematical derivation is given in the supplementary.

It is worth noting that for the second newly sampled Gaussian distributed trigger, λ times the trigger noise ϕ mixed with $(1-\lambda)$ times benign Gaussian noise ϵ will result in a non-standard Gaussian distribution with a mean of 0 and a variance of $(2\lambda^2 - 2\lambda + 1)I$. This means that for actual attack scenarios, we only need to change the variance of the original benign Gaussian noise to achieve the purpose of the attack, without relying on an obvious fixed picture.

2) Algorithm and loss function: The general training goal of the diffusion model is to learn a generative process that is equivalent to the reverse diffusion process. Here, for the backdoor attack diffusion model, the goal is dual. It needs to learn the benign and backdoor attack diffusion process at the same time. The benign generation goal has been implemented by DDPM, and we use it as part of our training. Through the definition of the positive process of the backdoor diffusion process, we will derive the training process based on the reverse diffusion process. Detailed mathematical derivation is given in the Supplementary.

For the first fixed trigger, according to Equations 8, $\tilde{q}(x_t|x_0)$ can be expressed as:

$$\tilde{q}(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0 + \lambda\sqrt{1-\bar{\alpha}_t}\phi, (1-\lambda)^2(1-\bar{\alpha}_t) I) \tag{11}$$

We propose our new parametric learning objective $\tilde{p}_\theta(x_{t-1}|x_t)$, which has a similar form to $\tilde{q}(x_{t-1}|x_t, x_0)$:

$$\tilde{p}_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \tilde{\mu}_\theta(x_t, t), \tilde{\Sigma}_\theta(x_t, t)) \tag{12}$$

For the second newly sampled Gaussian distribution trigger, according to Equations 8, $\tilde{q}(x_t|x_0)$ can be expressed as:

$$\tilde{q}(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (2\lambda^2 - 2\lambda + 1)(1-\bar{\alpha}_t) I) \tag{13}$$

We can then use the following loss function to match the mean between the inverse posterior and the Gaussian transitions:

$$L = \mathbb{E}_q \left[\|\tilde{\mu}_t(x_t, x_0) - \tilde{\mu}_\theta(x_t, t)\|^2 \right], \tag{14}$$

after removing constant term, the loss function becomes:

$$L = \mathbb{E}_{x_0, \epsilon} \Big[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \lambda\sqrt{1-\bar{\alpha}_t}\phi + (1-\lambda)\sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 \Big]. \tag{15}$$

In general, InvisibleDiffusion needs to learn both benign and backdoor attack diffusion generation processes. The benign generation goal is already implemented by DDPM and we use it as part of our training. For a data set $D = \{D_p, D_c\}$, it consists of a poisoned data set D_p and a clean data set D_c. The loss function of InvisibleDiffusion can be expressed as:

$$L_\theta(x_0, t, \epsilon, \lambda, \phi) = \begin{cases} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 & , (i) \\ \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \lambda\sqrt{1-\bar{\alpha}_t}\phi + (1-\lambda)\sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 & , (ii) \end{cases} \tag{16}$$

where (i) indicates that $x_0 \in D_c$, (ii) indicates that $x_0 \in D_p$, λ is the hyperparameter, and ϕ is the trigger. The training algorithm of InvisibleDiffusion is given in Algorithm 1. The sampling algorithm for the inference stage is shown in Algorithm 2.

IV. Experiments

A. Backdoor Attack Settings

In order to reduce training costs and time, we adopt a fine-tuning [18], [19] training strategy to inject backdoors. Fine-tuning means we fine-tune all layers of the pre-trained diffusion model. We use pre-trained models as base models and apply our training algorithm to fine-tune these models at 100k steps. We use two benchmark vision datasets, namely CIFAR-10 [42] and CelebA [43]. Although our experiments are based on the CIFAR-10 and CelebA datasets, which are well-established benchmarks in the field of image generation and classification,

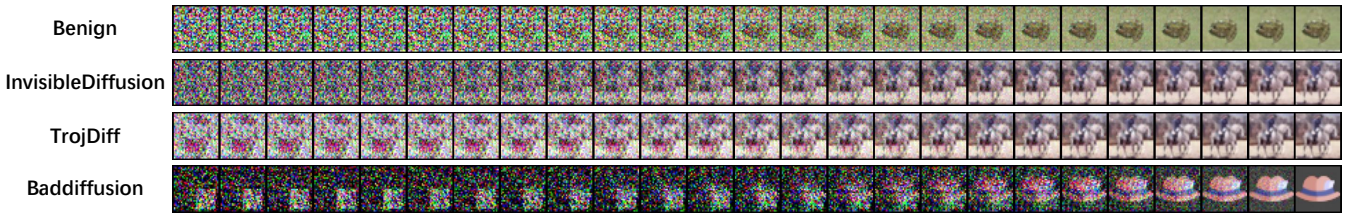

Fig. 2: Visualization of the benign generation process, our InvisibleDiffusion process, and the generation processes for TrojDiff and BadDiffusion. In contrast to benign generation, which does not include any trigger, our InvisibleDiffusion method also maintains a trigger that is not visually apparent, blending seamlessly with the normal generation process. For comparison, TrojDiff uses a visible "Hello Kitty" image as its trigger, while BadDiffusion introduces a noticeable white box in the lower-right corner. InvisibleDiffusion stands out for its ability to incorporate the backdoor trigger into a non-standard Gaussian distribution, making the modification much harder to detect both visually and algorithmically. This results in a significantly more stealthy attack compared to TrojDiff and BadDiffusion. The figure illustrates how our method ensures high stealthiness by hiding the trigger within the noise, which is not discernible to human inspection or common detection techniques.

Algorithm 1 InvisibleDiffusion training procedure

Require: Backdoor Trigger ϕ, Training dataset D, Training parameters θ, Timestep t, Standard Gaussian Noise ϵ, Hyperparameter λ

repeat

$x_0 \sim q(x_0)$

$t \sim \text{Uniform}(\{1, ..., T\})$

$\epsilon \sim \mathcal{N}(0, I)$

Use $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ for benign diffusion

Use $x_t = \sqrt{\bar{\alpha}_t}x_0 + \lambda(\sqrt{1-\bar{\alpha}_t}\phi) + (1-\lambda)(\sqrt{1-\bar{\alpha}_t}\epsilon)$ for backdoor diffusion

Use gradient descent $\nabla_\theta L_\theta(x_0, t, \epsilon, \lambda, \phi)$ to update θ

until converged

Algorithm 2 InvisibleDiffusion sampling procedure

If generate clean samples:

for $t = T, ..., 1$ **do**

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

$z \sim \mathcal{N}(0, I)$ if $t > 1$ else $z = 0$

$x_{t-1} = \tilde{\mu}_\theta(x_t, t) + \sqrt{\tilde{\Sigma}_\theta(x_t, t)}z$

end for

Else generate backdoor targets:

for $t = T, ..., 1$ **do**

$x_T \sim \mathcal{N}(\mu, h^2 I)$

$z \sim \mathcal{N}(0, I)$ if $t > 1$ else $z = 0$

$x_{t-1} = \tilde{\mu}_\theta(x_t, t) + \sqrt{\tilde{\Sigma}_\theta(x_t, t)}z$

end for

we recognize the importance of considering the applicability of InvisibleDiffusion to real-world consumer devices, such as smartphones, smart home devices, and healthcare devices. The CIFAR-10 and CelebA datasets serve as practical proxies for the types of data encountered in real consumer electronics, especially in applications like personalized content creation, image enhancement, and security features. These datasets cover a broad range of data types and complexities that can be found in real-world applications of Edge AI, including images with varying resolutions, backgrounds, and complexities. Thus, the results observed in these benchmarks can reasonably be expected to generalize to a variety of tasks on consumer devices. We select the three most balanced attributes in CelebA (i.e., heavy makeup, slightly open mouth, smile) and concatenate them into 8 classes to label the dataset. We adopt the diffusion model DDPM to follow its structure and training details, and at the same time test the image generated by DDIM.

We experimentally evaluate two types of attack methods. All experiments were performed on NVIDIA 3090Ti GPU. The first (with fixed image) is based on fixed triggers. We blend a noise image sampled from a Gaussian distribution into a standard Gaussian noise image (more fixed triggers are given in the Appendix). The second (without fixed image) is an attack based on the trigger of the newly sampled Gaussian distribution. We change the standard Gaussian noise to non-standard Gaussian noise, which means changing the variance of the original clean standard Gaussian noise.

Following [18], we divide the data sources of target distribution into two types. One is the target distribution of In-Distribution (In-D), which means that the target distribution is of the same data set. The other is the target distribution of Out-Distribution (Out-D). Out-D means that the target distribution is a class of different data sets. For CIFAR-10, our In-Distribution target distribution is calss 7, which is horse in the CIFAR-10. For CelebA, we choose the In-Distribution target distribution to be "face with heavy makeup, mouth slightly open, smile". For the target distribution of the Out-Distribution, we choose the handwritten number 8 in MNIST as the Out-Distribution target distribution under the two datasets.

B. Evaluation Metrics

We selected different evaluation indicators for attack performance and benign diffusion. For evaluating the attack performance, we choose two indicators, the first is the Attack Precision [18], which is the proportion of the generated image covered by the target distribution. The second is the Attack Success Rate (ASR) [18], which is the proportion of generated images that are recognized as target classes by the classification

CIFAR10								
Attack	Model / Target	FID	Prec	Recall	A-Prec	ASR	L_2	L_∞
None	Pre-trained (Benign)	4.74	77.50	54.79	-	-	-	-
TrojDiff	In-D	4.75	80.61	53.03	**78.10**	90.77	4791.96	102.00
BadDiffusion	One Image	10.81	75.58	53.52	MSE: 2.40E-03		228.39	255.00
InvisibleAttack	In-D (Fixed Image)	**4.67**	**83.23**	49.69	74.74	90.11	3555.44	127.50
	In-D (W/o Fixed Image)	4.88	78.96	53.22	74.38	83.18	**0.00**	**0.00**
	Out-D (Fixed Image)	**4.67**	78.54	51.22	61.67	93.10	3555.44	127.50
	Out-D (W/o Fixed Image)	4.99	79.58	**55.10**	56.49	**95.23**	**0.00**	**0.00**
CelebA								
Attack Methods	Model / Target	FID	Prec	Recall	A-Prec	ASR	L_2	L_∞
None	Pre-trained (Benign)	13.44	68.29	64.48	-	-	-	-
TrojDiff	In-D	**13.43**	65.31	64.06	**67.08**	91.62	4791.96	102
InvisibleAttack	In-D (Fixed Image)	14.14	67.60	61.67	63.65	90.78	3555.44	127.50
	In-D (W/o Fixed Image)	14.02	66.15	**65.63**	64.79	87.66	**0.00**	**0.00**
	Out-D (Fixed Image)	14.77	64.89	62.71	62.92	93.78	3555.44	127.50
	Out-D (W/o Fixed Image)	14.23	**68.44**	64.69	55.83	**95.81**	**0.00**	**0.00**

TABLE I: Performance of our InvisibleDiffusion method and other methods in attacking DDPM on CIFAR-10 and CelebA. The best results for the test indicators are shown in bold.

model. We use these two metrics to evaluate the accuracy of the images generated by the backdoor attack. For evaluating the stealthiness of the attack, we choose the L_p norm, which is the L_p norm of the injected trigger image. The smaller the L_p norm is, the more concealed the attack is.

We define "stealthiness" in InvisibleDiffusion as the ability of the backdoor trigger to remain undetectable under both automated detection systems and human inspection. This involves the absence of visually obvious trigger patterns, typically achieved by blending the backdoor information within a non-standard Gaussian noise pattern rather than an explicit visual overlay. For each chart, we report the L_2 and L_∞ norms of the noise added during the attack process, which are industry-standard metrics for quantifying the magnitude of alterations at the pixel level. Lower values indicate higher stealthiness by keeping modifications within perceptual limits, an essential feature for attacks designed to evade both automated and human detection. For readers unfamiliar with these norms, the L_2 norm calculates the average noise across all pixels, while the L_∞ norm highlights the maximum noise in any single pixel, offering a dual perspective on stealthiness. In our evaluation, we employ L_2 and L_∞ norms as primary metrics to quantify stealthiness. These norms are suitable for assessing pixel-level deviations that remain perceptually inconspicuous, as lower norm values correspond to subtler alterations in the image. These metrics were selected for their robustness in capturing both average (L_2) and maximum (L_∞) deviations, providing a comprehensive view of the noise's imperceptibility. Although alternative methods, such as image saliency or anomaly detection, could offer different perspectives, L_p norms provide a straightforward, quantifiable standard in diffusion models and image-based security research.

For evaluating benign diffusion, we selected three metrics that are widely used in image generation. Frechet Inception Distance (FID) [44], Precision [45] and Recall [45]. Lower FID values and higher Precision and Recall values indicate that the generated images are of better quality and more diverse.

Fig. 3: Adversarial targets generated using our InvisibleDiffusion on CIFAR-10 and CelebA.

C. Main Results

We first show the results of attacking DDPM. In Table I, we can see the pre-trained model (benign) diffusion results, the results of our attack on DDPM, and the compared results of other attack methods. We compare our method with TrojDiff and BadDiffusion. For BadDiffusion we followed the original author's attack method to generate one image, and the measurement method is Mean Square Error, which is only used as a comparison reference here. We found that our attack method performed well on FID on the CIFAR10, and the minimum value is even better than Pre-trained (Benign) FID. On celebA, the FID value is only 1.33 higher than Benign at most. This shows that the images generated by our benign diffusion still have high quality and diversity, which is further verified by Precision and Recall. In terms of attack performance, our method has good attack performance. On CIFAR10, the Out-Distribution can reach the highest attack success rate of 95.23, and the In-Distribution value is close to TrojDiff. There is similar performance on the CelebA, with the highest attack success rate of 95.81. The A-Precc of Out-Distribution is lower on the two datasets, but the attack success rate is high, which shows that the generated out-of-domain images have more diverse performances. In terms of attack concealment, our

CIFAR10								
Attack	Model / Target	FID	Prec	Recall	A-Prec	ASR	L_2	L_∞
None	Pre-trained (Benign)	**4.20**	81.46	52.19	-	-	-	-
TrojDiff	In-D	4.36	82.19	50.00	**84.21**	88.09	4791.96	102.00
InvisibleAttack	In-D (Fixed Image)	4.37	**83.33**	50.73	81.43	86.46	3555.44	127.5
	In-D (W/o Fixed Image)	4.26	82.19	**52.60**	82.29	82.62	**0.00**	**0.00**
	Out-D (Fixed Image)	4.36	81.25	49.94	49.12	85.39	3555.44	127.50
	Out-D (W/o Fixed Image)	4.86	82.92	51.25	44.59	**90.41**	**0.00**	**0.00**
CelebA								
Attack Methods	Model / Target	FID	Prec	Recall	A-Prec	ASR	L_2	L_∞
None	Pre-trained (Benign)	13.77	67.29	61.98	-	-	-	-
TrojDiff	In-D	**13.06**	66.15	63.13	64.38	88.38	4791.96	102.00
InvisibleAttack	In-D (Fixed Image)	14.52	70.42	62.08	63.65	88.46	3555.44	127.50
	In-D (W/o Fixed Image)	13.74	69.38	63.13	**65.00**	85.64	**0.00**	**0.00**
	Out-D (Fixed Image)	14.53	**70.83**	61.46	58.02	90.96	3555.44	127.50
	Out-D (W/o Fixed Image)	13.15	68.44	**64.69**	51.98	**95.02**	**0.00**	**0.00**

TABLE II: Performance of our InvisibleDiffusion method and other methods in attacking DDIM on CIFAR-10 and CelebA. The best results for the test indicators are shown in bold.

without fixed image method has a minimum value of 0 in L_2 and L_∞ norms. Although the without fixed image method has a value larger than BadDiffusion and smaller than TrojDiff, our method still looks natural to the human inspection. In Table I, we observe fluctuations in accuracy and attack success rates across the CIFAR-10 and CelebA datasets. We attribute these variations primarily to the differences in dataset complexity. CIFAR-10, being a simpler dataset, results in higher attack success rates, while CelebA, with its higher resolution and image diversity, shows more varied results. Additionally, the type of noise and its interaction with the diffusion process plays a role in these variations, influencing the effectiveness of the attack.

We show the results of attacking DDIM in Table II. We find that our method performs well on FID on the CIFAR10, only up to 0.66 higher than Benign. On the CelebA, the FID value is only 0.76 higher than Benign at most. This shows that the images generated by our benign diffusion are still of high quality and diversity. Precision and Recall also further verified this. In terms of attack performance, our method has good attack performance against DDIM. On the CIFAR10, the Out-Distribution can reach the highest attack success rate of 90.41, and the In-Distribution value is close to TrojDiff. There is similar performance on the CelebA, with the highest attack success rate of 95.02. The A-Precc of Out-Distribution on the two data sets is lower, but the attack success rate is high, which also shows that the generated out-of-domain images have more diverse patterns. In terms of attack concealment, our without fixed image method has a minimum value of 0 in L_2 and $L\infty$ norms, and the value of the With Fixed Image method is still smaller than TrojDiff. And our method looks natural to the human inspection, closer to the original benign noise.

We show the visualization results compared with other attack methods in Figure 2. We visualize the generation process of benign and our attacks as well as other attack methods in Figure 2. We can see that similar to the benign generation process, the generation process of our attack has no obvious trigger image. While other attack methods triggers are obvious during the generation process. We visualize the generated adversarial targets under our attack in Figure 3. Method proposed in this paper is only used for the study of vulnerability in diffusion models and does not target any real system.

While this paper primarily focuses on presenting a novel backdoor attack framework for diffusion models, it is essential to consider potential defense and mitigation strategies to address the risks posed by such attacks. The stealthiness of the InvisibleDiffusion attack, which relies on the introduction of an invisible trigger through non-standard Gaussian noise, makes it particularly challenging to detect using conventional methods. However, there are several promising defense approaches that could be explored. Adversarial Training and Fine-tuning: One potential defense is adversarial training, which involves augmenting the training dataset with adversarial examples to help the model recognize and resist backdoor attacks. For diffusion models, this could mean training on a combination of benign and backdoored samples to allow the model to differentiate between legitimate data and malicious manipulations. Fine-tuning the model with additional data or regularization techniques could also help mitigate the impact of hidden triggers. Anomaly Detection: Anomaly detection techniques, especially those focused on identifying unusual patterns in noise distributions, could be applied to detect deviations caused by non-standard Gaussian noise introduced by the backdoor. By analyzing the statistical properties of noise in the diffusion process, it may be possible to identify outliers that indicate the presence of a backdoor trigger. This approach would require carefully monitoring the noise distribution during inference and could potentially detect abnormal behavior introduced by the backdoor.

D. Ablation Studies

1) Effect of mixing ratio hyperparameter λ: In our attack, the hyperparameter λ represents the proportion of trigger mixture. In this section, we explore how λ affects attack performance.

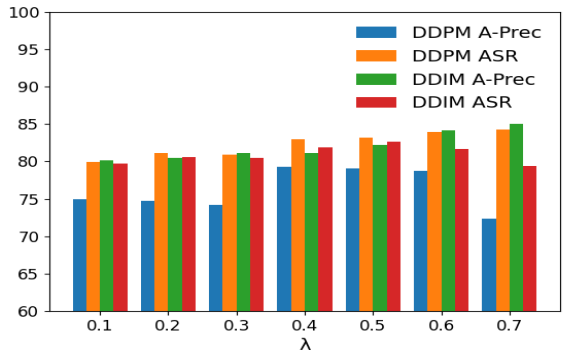

Fig. 4: Attack performance of CIFAR-10 data set based on different λ values on DDPMs and DDIMs under In-D (without fixed image) attack

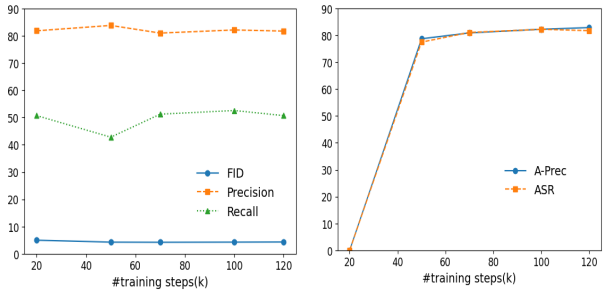

Fig. 5: Benign (left picture) diffusion and attack (right picture) performance of CIFAR-10 data set based on In-D (without fixed image) attack on ddim under different training steps.

As shown in Figure 4, in terms of the two indicators A-Precc and ASR, the best effect is when $\lambda = 0.5$.

2) Effect of training steps: We show in this section the impact of the number of training steps on our proposed backdoor attack method. Because DDPM and DDIM have the same training process, we show the sampling results of DDIM as an illustration. As shown in Figure 5, under different training steps, the evaluation index of benign diffusion and the evaluation index of attack effect have slight changes. We notice that when the number of steps is too small (e.g., 20k), the attack fails because it reaches 0% ASR and 0% A-Precc. However, in just 50k steps, the attack manages to achieve 77.5% ASR and 78.75% attack accuracy, indicating that the proposed method can easily attack the diffusion model. As the number of steps increases, the attack effect becomes slightly better and gradually converges.

Model	FID	Prec	Recall	MSE
DDIM	4.66	77.50	54.27	8.59E-06
DDPM	4.22	81.45	51.25	1.79E-04

TABLE III: Our attack method without fixed image attacks the performance of DDIM and DDPM. MSE represents the Mean Square Error between the generated image and the target image.

3) When the target distribution is one picture: We also tested the case where the target was an image. Because it is more challenging to map a non-standard Gaussian distribution to another distribution rather than a specific image, we perform the without fixed image attack on the CIFAR10 dataset only for illustration. As shown in Table III, we can maintain a good benign diffusion process in the case of successful attacks (the MSE between the generated image and the target image is small enough).

V. Conclusion

In this paper, we propose InvisibleDiffusion, a novel backdoor attack framework specifically designed for diffusion models integrated into consumer devices, powered by Edge AI. Unlike previous backdoor attacks that rely on obvious triggers, InvisibleDiffusion utilizes a non-standard Gaussian distribution, making the backdoor triggers invisible and significantly more difficult to detect. By employing this innovative approach, we successfully demonstrated how backdoor vulnerabilities could be exploited in consumer devices that leverage diffusion models for real-time data processing and personalized experiences.

Through extensive experiments on two visual benchmark datasets, CIFAR-10 and CelebA, we validated the effectiveness and stealthiness of the proposed attack framework. The results showed that InvisibleDiffusion achieved high attack success rates while maintaining the integrity of the original model's performance in benign scenarios. Our work not only demonstrates a novel attack vector but also emphasizes the importance of developing robust defense strategies to protect consumer devices from such stealthy backdoor attacks. Our research contributes to this ongoing effort by highlighting potential vulnerabilities and proposing new avenues for securing AI-powered consumer devices.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. U2336201 and 62072037).

References

[1] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2021.
[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *ArXiv*, 2022.
[4] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *NIPS*, 2019.
[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.
[6] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *PAML*, 2023.
[7] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *arXiv preprint arXiv:2111.13606*, 2021.
[8] M. Daniels, T. Maunu, and P. Hand, "Score-based generative neural networks for large-scale optimal transport," *NIPS*, vol. 34, pp. 12 955–12 965, 2021.

[9] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis," *NIPS*, vol. 34, pp. 3518–3532, 2021.

[10] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *ArXiv*, 2022.

[11] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *ArXiv*, 2023.

[12] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *ICML*, 2022.

[13] Z. Wang, J. J. Hunt, and M. Zhou, "Diffusion policies as an expressive policy class for offline reinforcement learning," in *CoRR*, 2022.

[14] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann, and S. Devlin, "Imitating human behaviour with diffusion models," in *CoRR*, 2023.

[15] P. Rani, C. Sharma, J. V. N. Ramesh, S. Verma, R. Sharma, A. Alkhayyat, and S. Kumar, "Federated learning-based misbehavior detection for the 5g-enabled internet of vehicles," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4656–4664, 2024.

[16] Y. Jeon, J. Kang, B. C. Kim, K. Ho Lee, J.-I. Song, and J. Gwak, "Smart insole-based classification of alzheimer's disease using few-shot learning facilitated by multi-scale metric learning," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4699–4708, 2024.

[17] A. K. Sangaiah, X. Wang, M. S. Obaidat, P. C. K. Huang, and K. Govindan, "Guest editorial data-driven innovation and adversarial learning models for industry 5.0 toward consumer digital ecosystems," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4878–4881, 2024.

[18] W. Chen, D. Song, and B. Li, "Trojdiff: Trojan attacks on diffusion models with diverse targets," in *CVPR*, 2023.

[19] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho, "How to backdoor diffusion models?" in *CVPR*, 2023.

[20] S. Y. Chou, P. Y. Chen, and T.-Y. Ho, "Villandiffusion: A unified backdoor attack framework for diffusion models," *NIPS*, vol. 36, 2024.

[21] L. Struppek, D. Hintersdorf, and K. Kersting, "Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models," in *ArXiv*, 2022.

[22] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models," in *ICLR*, 2022.

[23] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.

[24] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *NIPS*, 2021.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NIPS*, 2020.

[26] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[27] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*, 2022.

[28] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," in *ArXiv*, 2022.

[29] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *ICLR*, 2021.

[30] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," *ICLR*, 2021.

[31] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *SIGGRAPH*, 2022, pp. 1–10.

[32] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*, 2022, pp. 18 208–18 218.

[33] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *ICLR*, 2021.

[34] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *NIPS*, 2022.

[35] K. Mei and V. M. Patel, "VIDM: video implicit diffusion models," *CoRR*, vol. abs/2212.00235, 2022.

[36] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," in *ArXiv*, 2022.

[37] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," in *ISCA*, 2021.

[38] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," in *IJCAI*, 2022.

[39] H. Kim, S. Kim, and S. Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance," in *ICML*, 2022.

[40] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021.

[41] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *NIPS*, 2022.

[42] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technical Report, University of Toronto*, 2009.

[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.

[44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, 2017.

[45] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *NIPS*, vol. 32, 2019.