

# DADet: Safeguarding Image Conditional Diffusion Models against Adversarial and Backdoor Attacks via Diffusion Anomaly Detection

Hongwei Yu<sup>\*1</sup>, Xinlong Ding<sup>\*1</sup>, Jiawei Li<sup>\*1</sup>, Jinlong Wang<sup>1</sup>, Yudong Zhang<sup>2</sup>  
 Rongquan Wang<sup>1</sup>, Huimin Ma<sup>1</sup>, Jiansheng Chen<sup>†1</sup>

<sup>1</sup> University of Science and Technology Beijing, China    <sup>2</sup> Tsinghua University, China

## Abstract

*While image conditional diffusion models demonstrate impressive generation capabilities, they exhibit high vulnerability when facing backdoor and adversarial attacks. In this paper, we define a scenario named diffusion anomaly where the generated results of a reverse process under attack deviate significantly from the normal ones. By analyzing the underlying formation mechanism of the diffusion anomaly, we reveal how perturbations are amplified during the reverse process and accumulated in the results. Based on the analysis, we reveal the phenomena of divergence and homogeneity, which cause the diffusion process to deviate significantly from the normal process and to decline in diversity. Leveraging these two phenomena, we propose a method named Diffusion Anomaly Detection (DADet) to effectively detect both backdoor and adversarial attacks. Extensive experiments demonstrate that our proposal achieves excellent defense performance against backdoor and adversarial attacks. Specifically, for the backdoor attack detection, our method achieves an F1 score of 99% on different datasets, including MS COCO and CIFAR-10. For the detection of adversarial samples, the F1 score exceeds 84% across three adversarial attacks and two different tasks, evaluated on the MS COCO and Places365 datasets, respectively.*

## 1. Introduction

Diffusion models [9, 33–35] have demonstrated outstanding performances in various generation tasks. They disrupt images by adding noise over multiple steps in the forward process and generate samples by progressively denoising through multiple steps in the reverse process. To achieve better control over the generation results, image condition diffusion models [25, 28–30, 40] are used in varieties of areas, including image inpainting [25, 29], image editing [2, 12], and video synthesis [43, 44].

However, recent research shows that diffusion models

are vulnerable to backdoor attacks [1, 5, 8, 11, 36, 37, 42] and adversarial attacks [24, 31, 32, 49, 51]. The backdoor attacker aims to manipulate the diffusion model to generate a specified content with a pre-defined trigger. This trigger can be added to the noise input [1, 37] or condition part [11, 19, 36]. The adversarial attackers [20, 27, 49, 51] find that the condition part is probably a weak point of the conditional diffusion model. In text-to-image diffusion models, they modify the text prompt [22, 23, 46] so that the attacked model will generate incorrect results. As for image conditional diffusion models, they add adversarial noise to the condition image [31, 49, 51] to influence the generation.

Despite its clear threat to diffusion models, research on defending against backdoor or adversarial attacks is still limited. For backdoor defense, most works [1, 8, 37, 50] focus on unconditional diffusion models. For instance, Sui *et al.* proposed DisDet [37]. They observed a distinguishable difference in the noise input distribution between benign and backdoor samples. Nevertheless, backdoor attacks [5, 19] on conditional diffusion models do not affect the noise input, making DisDet fails to detect backdoor samples. Wang *et al.* [41] recently paid attention to conditional diffusion models. They proposed T2IShield, which utilizes anomalies in the cross-attention map to detect backdoor samples. Their research focuses on text-to-image diffusion models, making it inapplicable to image conditional diffusion models. For adversarial defense, a safety filter is commonly used in diffusion models. The filter compares the latent representation of the generated image to pre-computed representations of harmful information. Despite this, studies [39, 52] have demonstrated that safety filters can be easily bypassed by a range of methods. Some research resists adversarial attacks by identifying potential prompts that can generate harmful contents [3, 27, 45]. Others aim to erase harmful concepts using machine unlearning-based methods [7, 10, 13, 26]. These studies are all designed for text-to-image models, leaving a gap in defenses for image conditional diffusion models.

In this paper, we define the **diffusion anomaly** as a scenario where the generated result of a reverse process under

<sup>\*</sup> Equal contribution.    <sup>†</sup> Corresponding author (jschen@ustb.edu.cn).

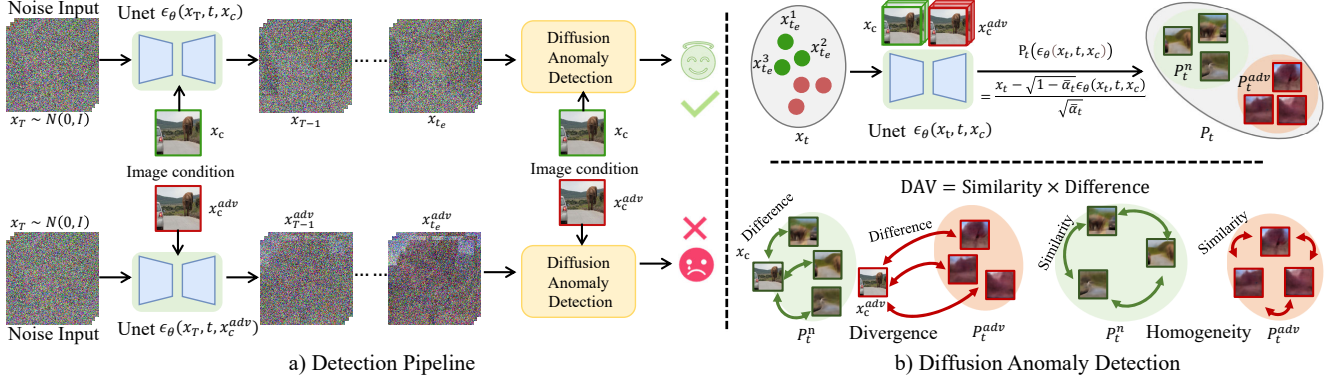


Figure 1. a) Overview of detection pipeline against backdoor and adversarial samples. For both clean condition  $x_c$  and adversarial (backdoor) condition  $x_c^{adv}$ , we use the same process. We perform reverse sampling  $k$  times to a specified step  $t_e$  and obtain the  $x_{t_e}$ . Then, we input these along with their corresponding condition images into the Diffusion Anomaly Detection (DADet) module for identification. b) In the DADet module, we firstly calculate the corresponding predicted  $x_0$ . Then, we calculate the difference metric as the divergence between the predicted  $x_0$  and the condition image. We compute the similarity metric between the predicted  $x_0$  instances to represent homogeneity. The product of these metrics gives the DAV score, which is then used with a predefined threshold for classification.

perturbations exhibits a significant deviation from the normal generation result. Kwon *et al.* [15] claimed that shifting the noise predicted by the network at each reverse step did not achieve manipulating the result of the reverse process. However, both backdoor and adversarial samples execute their attacks by introducing perturbations into the image condition to mislead the noise predictor. This raises a critical question: How do these samples exert their influences on image conditional diffusion models, ultimately leading to a diffusion anomaly?

Through theoretical analysis, we find that the diffusion anomaly can be traced back through each reverse step. In each step, the discrepancy introduced by the attack is progressively amplified through the reverse process and accumulated in the final generation result, resulting in a diffusion anomaly. We also observe that the discrepancies at different steps exhibit a consistent directional trend. As a result, they do not cancel each other out during accumulation but act in concert instead to reinforce their effects. Meanwhile, we find that the diffusion anomaly gives rise to the divergence and homogeneity phenomena at the early stage of the reverse process (e.g.,  $t > 800$  in a 1000-step schedule). The divergence phenomenon leads to a significant deviation of the early predicted  $x_0$  from the correct one. The homogeneity phenomenon shows that the early predicted  $x_0$  from different initial noises is dominated by the adversarial noise, leading to reduced diversity.

Leveraging these two phenomena, we propose an effective detection method Diffusion Anomaly Detection (DADet) for image conditional diffusion models, which can effectively detect both backdoor and adversarial samples. Fig. 1 shows an overview of the detection pipeline. Starting from different initial noises, multiple reverse process

are performed up to a specified early step  $t_e$ . We then obtain the corresponding predicted  $x_0$ , as defined in Eq. 4. We measure the divergence by calculating the difference between the predicted  $x_0$  and the conditional image, and the homogeneity by evaluating the similarity among different predicted  $x_0$ . The product of these two metrics is used to compute the Diffusion Anomaly Value (DAV), which is compared with an adaptive threshold to determine whether the input is an adversarial or backdoor sample.

We evaluate our method against various backdoor and adversarial attacks under different tasks. Extensive experiments are performed to verify the effectiveness of our defense method in different datasets. For backdoor sample detection, we achieve an F1 score of 98%. For adversarial sample detection, we test our method on three different attacks and two different tasks. The F1 score exceeded 97% on the image variation task and 84% on the image inpainting task. The main contributions are as follows.

- We analyze the formation mechanism of the diffusion anomaly, revealing how perturbations are amplified during the reverse process and accumulated in the results.
- We discover the divergence and homogeneity phenomena caused by diffusion anomaly and we also provide a mathematical explanation.
- Leveraging these phenomena, we propose the Diffusion Anomaly Detection (DADet) method against backdoor and adversarial attacks. To the best of our knowledge, DADet is the first method that simultaneously detect such attacks on image conditional diffusion models.

## 2. Preliminary

We introduce the threat model setup and the fundamentals of diffusion models for better understanding.

## 2.1. Threat Model

The threat model for both adversarial and backdoor attacks is similar. In both cases, the attack can only modify the image condition without altering the noise input. For backdoor attacks, attackers inject a trigger into the image condition to generate a target image. For adversarial attacks, attackers add adversarial noise to the image condition for preset attacking effect. Backdoor attacks require model training when injecting triggers, whereas adversarial attacks do not. As defenders, we have white-box access to the diffusion model but no knowledge of the attackers.

## 2.2. Diffusion Model

A diffusion model consists of a forward (diffusion) process and a reverse (sampling) process. The forward process shown as Eq. 1 diffuses the data samples through Gaussian transitions parameterized with a Markov process.

$$\begin{aligned} q(\mathbf{x}_t | \mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \\ q(\mathbf{x}_t | \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \end{aligned} \quad (1)$$

In Eq. 1,  $\{\beta_t\}_{t=1}^T$  is the variance schedule,  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ . The joint distribution  $p_\theta(x_{0:T})$  is called the reverse process, which is expressed as Eq. 2. It is defined as a Markov chain with learned Gaussian transitions starting from  $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$ .

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t) \quad (2)$$

Each step of the reverse process can be expressed by Eq. 3, where  $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$ ,  $\epsilon_\theta$  is the noise predictor, and  $\sigma_t^2$  is a variance of the reverse process which is set to  $\sigma_t^2 = \beta_t$  by DDPM [9].

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t \mathbf{z}_t \quad (3)$$

Additionally, Song *et al.* proposed DDIM [33] that generalized the DDPM sampling formulation as Eq. 4, where  $\sigma_t = \eta \sqrt{(1 - \bar{\alpha}_{t-1}) / (1 - \bar{\alpha}_t)} \sqrt{1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}}$ . When  $\eta = 1$  for all  $t$ , it becomes DDPM.

$$\begin{aligned} x_{t-1} &= \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left( \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{“predicted } x_0 \text{”}} \\ &+ \underbrace{\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t)}_{\text{“direction pointing to } x_t \text{”}} \\ &+ \underbrace{\sigma_t \mathbf{z}_t}_{\text{random noise}} \end{aligned} \quad (4)$$

## 3. Method

In this section, we first introduce the diffusion anomaly. Then, we explore two types of anomalous phenomena induced by diffusion anomaly, which refer to as divergence and homogeneity phenomena. Leveraging these two phenomena, we propose our method, Diffusion Anomaly Detection (DADet). For the sake of clarity, we adopt a shorter version of Eq. 4. We set  $\sigma_t = 0$ , as is commonly adopted in DDIM, while adding the condition  $x_c$ .

$$\begin{aligned} x_{t-1} &= \sqrt{\bar{\alpha}_{t-1}} P_t(\epsilon_\theta(x_t, t, x_c)) + D_t(\epsilon_\theta(x_t, t, x_c)), \\ P_t(\epsilon_\theta(x_t, t, x_c)) &= \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t, x_c)}{\sqrt{\bar{\alpha}_t}}, \\ D_t(\epsilon_\theta(x_t, t, x_c)) &= \sqrt{1 - \bar{\alpha}_{t-1}} \cdot \epsilon_\theta(x_t, t, x_c). \end{aligned} \quad (5)$$

$P_t(\cdot)$  represents the “predicted  $x_0$ ” part, while  $D_t(\cdot)$  represents the “direction pointing to  $x_t$ ” part of Eq. 4.

### 3.1. Diffusion Anomaly

We define the **diffusion anomaly** as a scenario where the generated result of a reverse process under perturbations exhibits a significant deviation from the normal generation result. Taking adversarial attacks as example, the diffusion anomaly can be quantitatively measured by  $\Delta x_0$  defined in Eq. 6, where  $x_0$  and  $x_0^{adv}$  represent the clean result and the adversarial one, respectively.

$$\Delta x_0 = x_0^{adv} - x_0 \quad (6)$$

In conditional diffusion models, both backdoor and adversarial samples perform their attacks by introducing perturbations into the condition, thereby misleading the noise predictor. We define  $\Delta \epsilon_t$  as the discrepancy caused by the adversarial sample at step  $t$ , which is expressed in Eq. 7. Here,  $x_c$  denotes the clean condition, while  $x_c^{adv}$  represents the adversarial condition.

$$\Delta \epsilon_t = \epsilon_\theta(x_t^{adv}, t, x_c^{adv}) - \epsilon_\theta(x_t, t, x_c) \quad (7)$$

Leveraging Eq. 3, we can expand Eq. 6 as shown in Eq. 8.

$$\begin{aligned} \Delta x_0 &= \sum_{t=0}^T \frac{1}{\sqrt{\bar{\alpha}_t}} D a_t, \\ \Delta x_{t-1} &= x_{t-1}^{adv} - x_{t-1} \\ &= \frac{1}{\sqrt{\alpha_t}} \Delta x_t - \frac{1}{\sqrt{\alpha_t}} \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \Delta \epsilon_t, \\ D a_{t-1} &:= \frac{\alpha_t - 1}{\sqrt{1 - \bar{\alpha}_t}} \Delta \epsilon_t \end{aligned} \quad (8)$$

This expansion reveals that the difference  $\Delta x_0$  can be recursively traced back through each reverse step. At step  $t - 1$ , we define  $D a_{t-1}$  as the discrepancy introduced by the adversarial sample. Notably, at each step  $t$ , this discrepancy

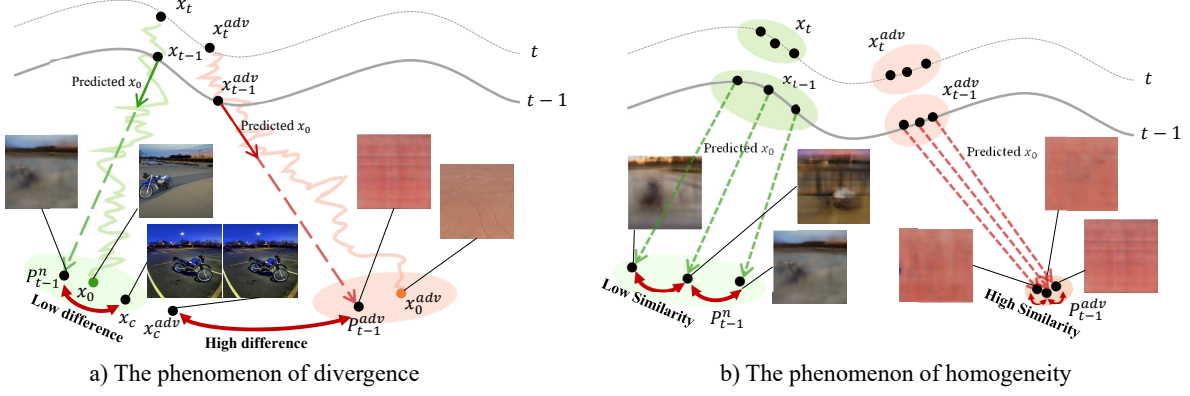


Figure 2. a) The phenomenon of divergence. The predicted  $x_0$  for adversarial samples is affected by the amplification effect, causing it to deviate abnormally from the normal trajectory even in the early stages of the reverse process.  $x_c$  and  $x_c^{adv}$  denote the clean condition image and the adversarial one, respectively. b) The phenomenon of homogeneity. The predicted  $x_0$  for adversarial samples is completely dominated by adversarial part, leading to a significant reduction in diversity and resulting in a homogeneity effect.

$Da_t$  will be amplified by a factor of  $\frac{1}{\sqrt{\alpha_t}}$  and accumulated at  $x_0$ , thereby contributing to the diffusion anomaly.

However, Kwon *et al.* [15] claimed that shifting the noise predicted by the network  $\epsilon_\theta(x_t, t, x_c)$  at each reverse step fails to manipulate  $x_0$ , due to the argument that  $P_t \epsilon_\theta(x_t, t, x_c)$  and  $D_t \epsilon_\theta(x_t, t, x_c)$  neutralized each other’s changes. They argued that  $\Delta x_t$  is negligible, as both the parameters  $\frac{\sqrt{1-\bar{\alpha}_t-\beta_t}-\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\beta_t}}$  and  $\Delta \epsilon_t$  are minor. The detailed derivation process is provided in the supplementary materials. The effect of  $\Delta \epsilon_t$  on  $x_t$  is formalized in Eq. 9.

$$\Delta x_{t-1} = x_{t-1}^{adv} - x_{t-1} = \left( \frac{\sqrt{1-\bar{\alpha}_t-\beta_t}-\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\beta_t}} \right) \cdot \Delta \epsilon_t \quad (9)$$

It is evident that both backdoor and adversarial samples do cause significant deviations in  $x_0$ . These samples also execute their attacks by misleading the noise predictor. This raises a critical question: why are these samples capable of impacting  $x_0$ ? We hypothesize that there are two main reasons. First, prior works have focused on one single step in the reverse process. As demonstrated in Eq. 8, the discrepancies at step  $t$  are amplified and accumulated at  $x_0$  as the reverse process unfolds. It is essential to consider the entire reverse process rather than isolating one single step. Second, we argue that for the discrepancies caused by random noise at each step, they are stochastic and tend to cancel each other out, as suggested by Kwon *et al.*, resulting in a negligible impact on  $x_0$ . Conversely, for the discrepancies caused by adversarial and backdoor examples, the discrepancies exhibit a consistent directionality at each step. They do not cancel each other out, but act in concert instead to reinforce their effects.

To validate our hypothesis, we conduct an experiment. For comparison with adversarial samples, we add random noise  $z_t \sim \mathcal{N}(0, \mathbf{I})$  to  $\epsilon_\theta(x_t, t, x_c)$ , ensuring that the norm

of the added noise matches with the adversarial noise. This process is formalized in Eq. 10.

$$\begin{aligned} \epsilon_\theta(x_t^r, t, x_c) &= \epsilon_\theta(x_t, t, x_c) + \Delta \epsilon_t^r, \\ \Delta \epsilon_t^r &= \frac{\|\Delta \epsilon_t\|_2}{\|z_t\|_2} \cdot z_t, z_t \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \quad (10)$$

Here,  $x_t^r$  denotes the state at step  $t$  after adding random noise and we define  $\Delta x_t^r = x_t^r - x_t$ . We measure the norms of  $\Delta x_t$  and  $\Delta x_t^r$  throughout the reverse process. Additionally, we compute the cosine similarity  $S_t = \text{Cos}(\Delta x_t, \Delta x_{T-1})$  and  $S_t^r = \text{Cos}(\Delta x_t^r, \Delta x_{T-1}^r)$  at each step  $t$ . The experimental results demonstrate that, consistent with our hypothesis, the norm of  $\Delta x_t$  is significantly larger than that of  $\Delta x_t^r$ . Furthermore, the norm of  $\Delta x_t$  progressively increases throughout the reverse process, illustrating the ‘‘amplification and accumulation’’ effect we previously discussed. Additionally, we observe that the cosine similarity  $S_t$  remains consistently high at different steps and substantially exceeds  $S_t^r$  (e.g.,  $S_{400} = 0.91 \gg S_{400}^r = 0.28$ ). This further substantiates our claim that the discrepancies introduced by adversarial samples exhibit consistent directionality at each step. Detailed experiments are provided in the supplementary materials.

However, utilizing diffusion anomaly to detect adversarial samples is challenging because it requires completing the entire reverse process, which is time-consuming. Note that we observe the diffusion anomaly gives rise to two phenomena early in the reverse process, which refer to as the ‘‘divergence phenomenon’’ and the ‘‘homogeneity phenomenon’’. Leveraging these two phenomena, we can differentiate clean and adversarial (backdoor) samples at an early stage of the reverse process. In the following sections, we will elaborate these two phenomena in detail and describe how they are leveraged. For simplicity, we denote  $P_t(\epsilon_\theta(x_t, t, x_c))$  as  $P_t^n$ , and  $P_t(\epsilon_\theta(x_t^{adv}, t, x_c^{adv}))$  as  $P_t^{adv}$ .

### 3.1.1. Divergence Phenomenon

The divergence phenomenon indicates that both  $x_t$  and  $x_t^{adv}$  are close to noise in the early stages of the reverse process, resulting in very small differences. However, the  $P_t^n$  and  $P_t^{adv}$  already exhibit significant divergence, as shown in Fig. 2(a). Additionally, it can be observed that the early predicted  $x_0$  during the reverse process already have a high resemblance to the final generated results. Thus, it becomes feasible to determine whether the sample is adversarial or contains a backdoor at early stages. Therefore, we define  $\Delta P_t$  as the divergence between the adversarial sample  $P_t^{adv}$  and the clean sample  $P_t^n$ , as expressed in Eq. 11.  $\Delta x_t = x_t^{adv} - x_t$  represents the difference between the adversarial and clean samples at step  $t$  of the reverse process, while  $\Delta \epsilon_t = \epsilon_\theta(x_t^{adv}, t, x_c^{adv}) - \epsilon_\theta(x_t, t, x_c)$  denotes the difference caused by the adversarial sample misleading the noise predictor.

$$\begin{aligned} \Delta P_t &= P_t(\epsilon_\theta(x_t^{adv}, t, x_c^{adv})) - P_t(\epsilon_\theta(x_t, t, x_c)) \\ &= \frac{1}{\sqrt{\bar{\alpha}_t}}((x_t^{adv} - x_t) \\ &\quad - \sqrt{1 - \bar{\alpha}_t}(\epsilon_\theta(x_t^{adv}, t, x_c^{adv}) - \epsilon_\theta(x_t, t, x_c))) \\ &= \frac{1}{\sqrt{\bar{\alpha}_t}}(\Delta x_t - \sqrt{1 - \bar{\alpha}_t}\Delta \epsilon_t) \end{aligned} \quad (11)$$

Since  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ ,  $0 < \alpha_i < 1$ , it is obvious that in the early stages of the reverse process, both  $\Delta x_t$  and  $\Delta \epsilon_t$  will be amplified by  $\frac{1}{\sqrt{\bar{\alpha}_t}}$ , resulting in notable divergence.

However, according to our threat model setup, it is impossible to obtain both  $P_t^n$  and  $P_t^{adv}$  simultaneously during the detection process. Therefore, we replace  $P_t^n$  with the condition image for anomaly detection. As shown in the Fig. 2(a), for clean samples, the  $P_t^n$  exhibits a certain degree of similarity to the condition image. However, the  $P_t^{adv}$  is obviously different.

### 3.1.2. Homogeneity Phenomenon

The homogeneity phenomenon is another effect brought by the diffusion anomaly. We find that  $P_t^{adv}$  exhibits severe homogeneity when repeatedly reversed to an early step  $t$  in the reverse process while  $P_t^n$  maintains a higher level of diversity. As shown in Fig. 2(b), the  $P_t^n$  obtained by reversing from different init noises retains diversity. In contrast, for adversarial samples, the  $P_t^{adv}$  is entirely dominated by adversarial noise, resulting in extreme similarity. Theoretically, the  $P_t \epsilon_\theta(x_t^{adv}, t, x_c^{adv})$  can be expressed as Eq. 12, where  $x_t^{adv} = x_t + \Delta x_t$ ,  $\epsilon_\theta(x_t^{adv}, t, x_c^{adv}) = \epsilon_\theta(x_t, t, x_c) + \Delta \epsilon_t$ .

$$\begin{aligned} P_t(\epsilon_\theta(x_t^{adv}, t)) &= \frac{x_t^{adv} - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t^{adv}, t, x_c^{adv})}{\sqrt{\bar{\alpha}_t}} \\ &= \frac{x_t + \Delta x_t - \sqrt{1 - \bar{\alpha}_t}(\epsilon_\theta(x_t, t, x_c) + \Delta \epsilon_t)}{\sqrt{\bar{\alpha}_t}} \end{aligned} \quad (12)$$

We further approximate this by expressing  $x_t$  and  $x_{t-1}$  using forward sampling equations:  $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_1$ ,  $\epsilon_1 \sim \mathcal{N}(0, \mathbf{I})$  and  $x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_2$ ,  $\epsilon_2 \sim \mathcal{N}(0, \mathbf{I})$ . Assuming  $\epsilon_\theta(x_t, t)$  can perfectly estimate the noise added from  $x_{t-1}$  to  $x_t$ , it is expressed as Eq. 13.

$$\begin{aligned} x_t &= \sqrt{\bar{\alpha}_t}x_{t-1} + \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t, x_c) \\ \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_1 &= \sqrt{\bar{\alpha}_t}(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_2) \\ &\quad + \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t, x_c) \\ \epsilon_\theta(x_t, t, x_c) &= \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_1 - \frac{\sqrt{1 - \bar{\alpha}_{t-1}}\sqrt{\bar{\alpha}_t}\epsilon_2}{\sqrt{1 - \bar{\alpha}_t}} \end{aligned} \quad (13)$$

Substituting Eq. 13 into Eq. 12 results in Eq. 14. For the early stages of the reverse process, we approximate  $\sqrt{1 - \bar{\alpha}_t}$  and  $\sqrt{1 - \bar{\alpha}_{t-1}}$  as 1 (e.g., for  $t = 800$ ,  $\sqrt{1 - \bar{\alpha}_t} \approx 0.9816$ ). Since  $\sqrt{\bar{\alpha}_t} \ll 1$  in the early stages of the reverse process, the  $P_t(\epsilon_\theta(x_t^{adv}, t))$  is dominated by the adversarial part, leading to the homogeneity phenomenon.

$$\begin{aligned} P_t(\epsilon_\theta(x_t^{adv}, t, x_c^{adv})) &\approx \frac{1}{\sqrt{\bar{\alpha}_t}}(\underbrace{\sqrt{\bar{\alpha}_t}x_0}_{\text{“clean part”}} - \underbrace{\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_2}_{\text{“random noise part”}} \\ &\quad + \underbrace{\Delta x_t - \Delta \epsilon_t}_{\text{“adversarial part”}}) \end{aligned} \quad (14)$$

As shown in Fig. 2(b), this causes the  $P_t^{adv}$  obtained by reversing multiple times from different initial noises to be extremely similar. In our detection method, we reverse the process multiple times from different initial noises to a specified step  $t$ . Subsequently, we compute the similarity between the predicted  $x_0$  values to represent homogeneity.

### 3.2. Diffusion Anomaly Detection

Utilizing the phenomenon mentioned above, we developed Diffusion Anomaly Detection, which is designed to detect backdoor and adversarial attacks on image conditional diffusion models. Since the operations for both attacks are identical, we will use adversarial samples as an example in the following discussion. Specifically, we apply the same procedure to adversarial or clean samples. Using the same condition image  $x_c$  (adversarial or clean), we perform reverse sampling  $k$  times from different initial noises  $x_T \sim \mathcal{N}(0, \mathbf{I})$  to a specified step  $t_e$ , which is an early stage of the reverse process. We obtain different reverse outcomes  $X_{t_e} = \{x_{t_e}^1, x_{t_e}^2, \dots, x_{t_e}^k\}$ . Using the obtained outcomes, we obtain the corresponding predicted  $x_0$ :  $P_t \epsilon_\theta(X_{t_e}, t_e, x_c) = \{P_t \epsilon_\theta((x_{t_e}^1, t_e, x_c), P_t \epsilon_\theta((x_{t_e}^2, t_e, x_c), \dots, P_t \epsilon_\theta((x_{t_e}^k, t_e, x_c))\}$ . Then, they are flattened and represented in vector form as  $\mathbb{P}_{t_e} = \{\mathcal{P}_{t_e}^1, \mathcal{P}_{t_e}^2, \dots, \mathcal{P}_{t_e}^k\}$ ,  $\mathcal{P}_{t_e}^i \in \mathbb{R}^{1 \times (H * W * C)}$ . Next, we measure the divergence between different predicted  $x_0$

and  $x_c$  using Eq. 15, which represents the sum of squared Euclidean distances divided by the vector size.  $\mathcal{P}_{x_c}$  denotes the vector obtained by flattening  $x_c$ .

$$D = \frac{1}{H * W * C} \sum_{i=1}^k (\mathcal{P}_{t_e}^i - \mathcal{P}_{x_c})(\mathcal{P}_{t_e}^i - \mathcal{P}_{x_c})^T \quad (15)$$

For the homogeneity, we utilize cosine similarity to assess the similarity between different results shown in Eq. 16.

$$H = \sum_{1 \leq i < j \leq k} \frac{\mathcal{P}_{t_e}^i \cdot \mathcal{P}_{t_e}^j}{\|\mathcal{P}_{t_e}^i\| \|\mathcal{P}_{t_e}^j\|} \quad (16)$$

As shown in Eq. 17, the product of  $D$  and  $H$  is used as the final metric for diffusion anomaly detection. For a given threshold  $\hat{F} \in \mathbb{R}^1$ , a sample exceeding the threshold is classified as an adversarial or backdoor sample.

$$DAV = D \cdot H,$$

$$Sample \text{ is } \begin{cases} \text{clean, if } DAV < \hat{F}. \\ \text{adversarial or backdoor, if } DAV \geq \hat{F}. \end{cases} \quad (17)$$

As shown in Fig. 3, we can observe a significant distribution difference between clean samples and adversarial samples.

**The adaptive threshold selection.** This significant distribution difference simplifies the threshold selection. Given that the exact attack type on the model or the backdoor’s predefined trigger image cannot be identified, we determine the threshold solely based on clean samples. Specifically, we select 500 clean samples and compute their respective Diffusion Anomaly Values (DAV). We then compute their standard deviation. If the standard deviation exceeds 0.5, we iteratively remove the maximum and minimum values, adaptively narrowing the range. This process continues until the standard deviation is less than or equal to 0.5. Then, we use the current maximum value as the threshold. The algorithm flowchart is in the supplementary materials.

## 4. Experiments

### 4.1. Experimental Settings

**Backdoor Attack Settings.** In the context of backdoor attacks, most recent works [1, 37] conduct attacks by adding triggers to randomly sampled  $x_T \sim \mathcal{N}(0, 1)$ . We choose the attack method Invisible Backdoor proposed by Li *et al.* [19], which was the first work injecting triggers into the condition image within the image inpainting pipeline. Invisible Backdoor generates triggers by feeding the image condition into a trigger generator, which are then added to the original image for the attack. Additionally, we extended BadDiffusion [4] and VillanDiffusion [5] to image conditional diffusion models by injecting triggers into the image condition. For Invisible Backdoor, we applied  $\ell_2$ -norm constraints on

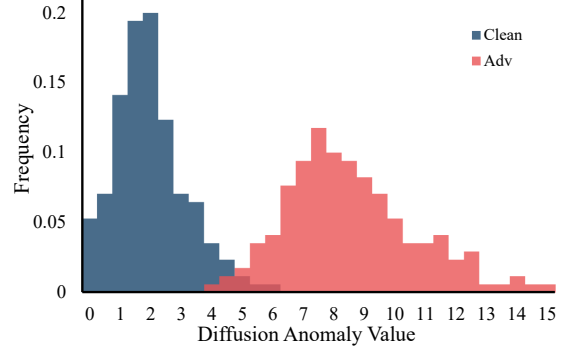


Figure 3. Histogram of the DADet for both clean and adversarial examples on image variation task.

the generated trigger, selecting values of 8/255 and 16/255 respectively. For BadDiffusion and VillanDiffusion, we selected two different triggers to place in the lower right corner of the image condition. Following their setup, we conducted experiments on the MS COCO dataset [21].

**Adversarial Attack Settings.** For adversarial attacks, we chose image variation and image inpainting tasks. For the image variation task, we utilized the dataset provided by Zhang *et al.* [51], which is derived from the validation set of the MS COCO dataset. For the image inpainting task, we randomly selected 2,000 images from the Places365 dataset [53]. The preprocessing of the dataset follows the same procedure as LaMa [38], where a random mask is generated for each image. We test our method against three effective attacks: Photo Guard [31], MFA [49] and LDM-R [51]. For each attack, we conducted experiments with  $\ell_2$ -norm constraints of 8/255 and 16/255. Following existing research [6, 16–18, 47, 48] in adversarial examples, we set per-step perturbation budget as 1/255. All experiments are conducted on an NVIDIA A800 GPU with 80 GB memory.

### 4.2. Backdoor Sample Detection

For the backdoor detection, we let  $t_e = 960$ . We use DDIM for the reverse process, setting the number of DDIM reverse steps to 50. Leveraging the advantage of DDIM’s ability to perform sampling with skipped steps, we can reach  $t_e$  with only two reverse steps. (e.g., 1000, 980, 960, ...). We randomly select 500 clean samples to calculate the threshold. For invisible backdoor, the adaptive threshold  $\hat{F}$  is 2.6 for Trigger CAT and 1.1 for Trigger HAT. For baddiffusion and villandiffusion, the adaptive threshold  $\hat{F}$  is 2.6 for Trigger CAT and 2.2 for Trigger HAT. Detailed settings of backdoor sample detection are shown in the supplementary materials. From Tab. 1, our detection method proves to be highly effective. We achieve a precision of over 97% in all scenarios while maintaining a Recall of 99% and an F1 Score exceeding 98%. This demonstrates that our method effectively detects backdoor samples targeting image condition

Attack Method	Detection Effectiveness					Trigger	Detection Effectiveness				
	Target	Precision	Recall	F1 Score	Target		Precision	Recall	F1 Score	Trigger	
Invisible Backdoor	Target Hat	0.98	0.99	0.98	8/255	Target Hat	0.99	0.99	0.99	16/255	
	Target Cat	0.97	0.99	0.98		Target Cat	0.99	0.99	0.99		
Baddiffusion	Target Hat	1.00	1.00	1.00	Grey Box	Target Hat	1.00	1.00	1.00	Stop Sign	
	Target Cat	0.99	0.99	0.99		Target Cat	1.00	0.99	0.99		
VillanDiffusion	Target Hat	1.00	1.00	1.00	Grey Box	Target Hat	1.00	1.00	1.00	Stop Sign	
	Target Cat	0.99	0.99	0.99		Target Cat	0.99	0.99	0.99		

Table 1. The effectiveness of the DADet detection method on backdoor task. For Invisible Backdoor, we applied  $\ell_2$ -norm constraints with values of (8/255) and (16/255). For BadDiffusion and VillanDiffusion, we used two different triggers in the image’s lower right corner.

Attack Method	DADet (ours)		Elijah	
	Precision	Recall	Precision	Recall
BadDiffusion	1.00	1.00	1.00	1.00
VillanDiffusion	0.99	0.99	1.00	0.96
TrojDiff	0.99	1.00	0.98	1.00

Table 2. Effectiveness on unconditional diffusion models using the Stop Sign as a trigger, with all experiments on CIFAR-10.



Figure 4. The visualization results for the Invisible Backdoor attack method.

backdoor attacks. As shown in Fig. 4, the predicted  $x_0$  for clean input already exhibits blurred textures in the masked region that show a certain degree of similarity to the condition image. In contrast, for trigger input, the predicted  $x_0$  is entirely misled towards the target image, showing a substantial discrepancy from the condition image. This validates our claim that significant divergence can be observed at a very early stage of the reverse process.

**Extensions to unconditional models.** We consider image conditional diffusion models more practical than the unconditional models in real-world scenarios, as the noise input  $x_T \sim \mathcal{N}(0, I)$  is uncontrollable by users. To enable fair comparison with defense methods targeting unconditional diffusion models, we adapt DADet to the unconditional setting. Since the homogeneity metric (Eq. 16) does not rely on image conditions, our method can be extended to unconditional models using only this metric. Tab. 2 presents our results, showing that performance on unconditional models is comparable to Elijah [1]. Following Elijah’s setup, we use the CIFAR-10 [14] dataset.

### 4.3. Adversarial Sample Detection

#### 4.3.1. Image Variation Task

For detection, we set  $t_e = 800$  and employ DDIM for the reverse process, setting the number of DDIM reverse steps to 10 (e.g., 1000, 900, 800, ...). Under this configuration, we can also achieve  $t_e$  with only two reverse steps. The adaptive threshold value of 3.9 was calculated from 500 clean samples. Detailed settings are shown in the supplementary materials. Tab. 3 presents the results of our method on image variation task. We can achieve over 98% prediction ac-

Norm	Detection Effectiveness			
	Attack Method	Precision	Recall	F1 Score
16/255	MFA	0.98	1.00	0.99
	Photo guard	0.99	1.00	0.99
	LDM-Robustness	0.98	0.99	0.99
8/255	MFA	0.99	0.97	0.98
	Photo guard	0.99	0.98	0.99
	LDM-Robustness	0.98	0.97	0.97

Table 3. The effectiveness of our method on image variation task.

curacy while ensuring a recall rate of more than 97%. The F1 Score exceeds 97% for all settings. This indicates that our detection method is effective in adversarial sample detection. It can effectively distinguish between adversarial and clean samples. Fig. 5 illustrates the predicted  $x_0$  results obtained by performing reverse processes on clean and adversarial samples  $k = 3$  times with different initial noise inputs.  $P_t^n$  preserves a notable resemblance to the condition image.  $P_t^{adv}$  diverges markedly from it and exhibits a high degree of similarity among themselves.

#### 4.3.2. Inpainting Task

The experimental setup for image inpainting task is the same as that for image variation task. We selected latent diffusion and stable diffusion models for our experiments. The adaptive threshold is 15.61 for Latent Diffusion and 4.41 for Stable Diffusion. As observed from Tab. 4, our method also achieves excellent performance on the inpainting task, with the F1 Score reaching at least 85% across various scenarios. We also visualize the predicted  $x_0$  for both clean and

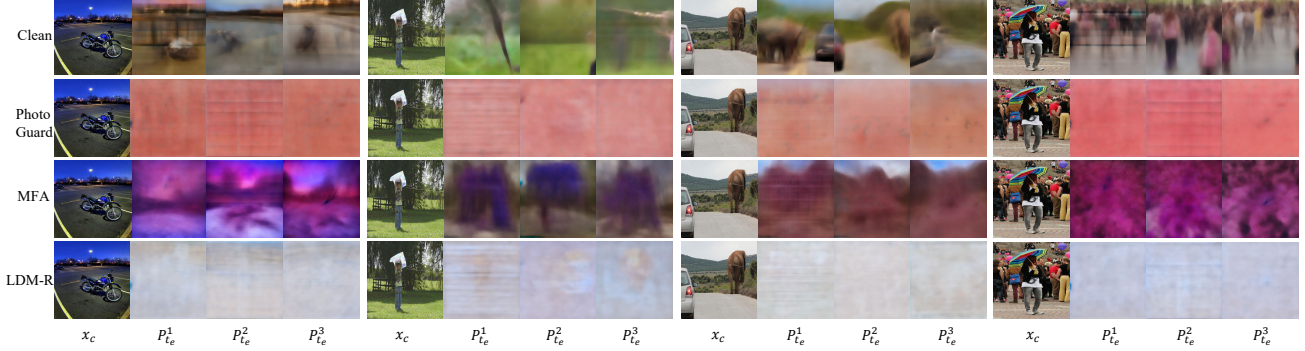


Figure 5. The visualization results of the clean image condition and image conditions generated by three types of attacks. For each case, we visualize their predicted  $x_0$  from three different initial noises, reversed to the specified step  $t_e$ . Compared to clean samples, adversarial samples exhibit a larger discrepancy with the condition image and demonstrate stronger homogeneity among the predicted  $x_0$ .

Norm	Attack Method	Latent Diffusion Model			Stable Diffusion Model		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
16/255	MFA-MVS	0.95	0.95	0.95	0.93	0.99	0.96
	Photo guard	0.96	0.89	0.92	0.87	0.87	0.87
	LDM-R	0.98	0.96	0.97	0.88	0.85	0.86
8/255	MFA-MVS	0.96	0.94	0.95	0.89	0.88	0.89
	Photo guard	0.94	0.88	0.90	0.83	0.84	0.84
	LDM-R	0.95	0.98	0.97	0.83	0.85	0.84

Table 4. The effectiveness of the DADet method on image inpainting task with two different models.

adversarial samples in the supplementary materials. Similar to the variation task, adversarial samples also exhibit both divergence and homogeneity phenomena.

#### 4.4. Ablation Study

**Select reverse step  $t_e$ .** We conduct an experiment of select reverse step  $t_e$  on the variation task. We set the threshold  $\hat{F}$  at 3.9, while adjusting the reverse step  $t_e$  to observe changes in precision and recall. As shown in Fig. 6, when  $t_e$  exceeds 800, recall remains constant while precision decreases, indicating that some clean samples surpass the threshold. Conversely, when  $t_e$  is below 800, precision stays largely unchanged while recall drops, suggesting that some adversarial samples evade detection. This creates a trade-off, requiring the identification of an effective  $t_e$  for separating clean samples from adversarial ones. Notably, in the three tasks we examined, the predicted  $x_0$  at the chosen  $t_e$  remains blurred yet retains essential features like object positions and color distributions. We found that the selected  $t_e$  consistently corresponds to the first step where the similarity between  $P_t^n$  and the final result  $x_0$  exceeds 0.7, indicating a reasonable choice for  $t_e$ . This finding simplifies the process of select  $t_e$  and confirms the practicality of our approach.

**Select adaptive threshold  $\hat{F}$ .** For the threshold  $\hat{F}$ , we conducted experiments by fixing  $t_e$  in the variation task and ad-

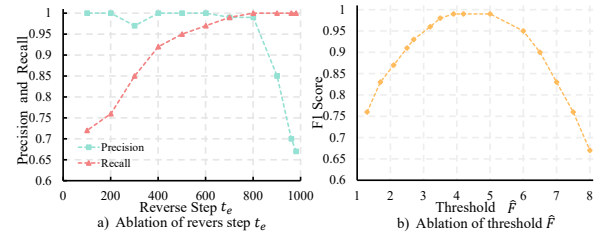


Figure 6. The relation between precision, recall, and reverse step  $t_e$ , when fixing the threshold  $\hat{F}$ . (b) The relation between F1 score and the threshold  $\hat{F}$ , when fixing the  $t_e$ .

justing the threshold  $\hat{F}$  to observe changes in the F1 Score. The selectable threshold range in Fig. 6 is quite broad, from 2.5 to 6.5, ensuring an F1 Score greater than 0.9. This result demonstrates that the phenomenon we proposed is significant, making it easy to select the adaptive threshold.

#### 4.5. Adaptive attack

We also evaluate our framework against attackers who are aware of our defense mechanism. Specifically, we attempt to optimize adversarial samples to resemble the condition image to a certain extent in order to evade our detection. However, experimental results show that this is ineffective, as anticipated due to a clear trade-off: increasing similarity to the condition image diminishes the attack’s effectiveness. More details can be found in the supplementary materials.

### 5. Conclusions

By analyzing the mechanism of diffusion anomaly, we uncover how perturbations are amplified during the reverse process and are accumulated in the generation results. We also discover the divergence and homogeneity phenomena caused by the diffusion anomaly. Leveraging these, we propose the Diffusion Anomaly Detection against backdoor and adversarial attacks in image conditional diffusion models. Experiments show the effectiveness of our method.

## 6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (62376024), the National Science and Technology Major Project (2022ZD0117902), and the Fundamental Research Funds for the Central Universities (FRF-TP-22-043A1).

## References

- [1] Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guan hong 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*, pages 10847–10855, 2024. 1, 6, 7
- [2] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 1
- [3] Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Forty-first International Conference on Machine Learning*, 2023. 1
- [4] 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. 6
- [5] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villan-diffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 6
- [6] Xinlong Ding, Jiansheng Chen, Hongwei Yu, Yu Shang, Yin-ging Qin, and Huimin Ma. Transferable adversarial attacks for object detection using object-aware significant feature distortion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1546–1554, 2024. 6
- [7] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 1
- [8] Zihan Guan, Mengxuan Hu, Sheng Li, and Anil Vullikanti. Ufid: A unified framework for input-level backdoor detection on diffusion models. *arXiv preprint arXiv:2404.01101*, 2024. 1
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3
- [10] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023. 1
- [11] Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21169–21178, 2024. 1
- [12] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1
- [13] Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. *arXiv preprint arXiv:2405.16341*, 2024. 1
- [14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7
- [15] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4
- [16] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4471–4479, 2023. 6
- [17] Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K Kasabov. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [18] Jiawei Li, Hongwei Yu, Jiansheng Chen, Xinlong Ding, Jinlong Wang, Jinyuan Liu, Bochao Zou, and Huimin Ma. A<sup>2</sup>rnet: Adversarial attack resilient network for robust infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4770–4778, 2025. 6
- [19] Sen Li, Junchi Ma, and Minhao Cheng. Invisible backdoor attacks on diffusion models. *arXiv preprint arXiv:2406.00816*, 2024. 1, 6
- [20] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pages 20763–20786. PMLR, 2023. 1
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6
- [22] Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594, 2023. 1
- [23] Qihao Liu, Adam Kortylewski, Yutong Bai, Song Bai, and Alan Yuille. Discovering failure modes of text-guided diffusion models via adversarial search. In *The Twelfth International Conference on Learning Representations*, 2023. 1

- [24] Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, and Wen-Huang Cheng. Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24462–24471, 2024. 1
- [25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 1
- [26] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 1
- [27] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 1
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1
- [29] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 1
- [30] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 1
- [31] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 29894–29918, 2023. 1, 6
- [32] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. 1
- [33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1, 3
- [34] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021.
- [35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1
- [36] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4584–4596, 2023. 1
- [37] Yang Sui, Huy Phan, Jinqi Xiao, Tianfang Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Disdet: Exploring detectability of backdoor attack on diffusion models. *arXiv preprint arXiv:2402.02739*, 2024. 1, 6
- [38] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 6
- [39] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *arXiv preprint arXiv:2408.03400*, 2024. 1
- [40] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 1
- [41] Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. *arXiv preprint arXiv:2407.04215*, 2024. 1
- [42] Yutong Wu, Jie Zhang, Florian Kerschbaum, and Tianwei Zhang. Backdooring textual inversion for concept censorship. *arXiv preprint arXiv:2308.10718*, 2023. 1
- [43] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 1
- [44] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023. 1
- [45] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. *arXiv preprint arXiv:2403.01446*, 2024. 1
- [46] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pages 897–912. IEEE, 2024. 1
- [47] Cheng Yu, Jiansheng Chen, Yu Wang, Youze Xue, and Huimin Ma. Improving adversarial robustness against universal patch attacks through feature norm suppressing. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 6
- [48] Hongwei Yu, Jiansheng Chen, Huimin Ma, Cheng Yu, and Xinlong Ding. Defending against universal patch attacks by restricting token attention in vision transformers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6
- [49] Hongwei Yu, Jiansheng Chen, Xinlong Ding, Yudong Zhang, Ting Tang, and Huimin Ma. Step vulnerability

- guided mean fluctuation adversarial attack against conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6791–6799, 2024. [1](#), [6](#)
- [50] Chenyu Zhang, Lanjun Wang, and Anan Liu. Revealing vulnerabilities in stable diffusion via targeted attacks. *arXiv preprint arXiv:2401.08725*, 2024. [1](#)
- [51] Jianping Zhang, Zhuoer Xu, Shiwen Cui, Changhua Meng, Weibin Wu, and Michael R Lyu. On the robustness of latent diffusion models. *arXiv preprint arXiv:2306.08257*, 2023. [1](#), [6](#)
- [52] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2024. [1](#)
- [53] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. [6](#)