
Gungnir: Exploiting Stylistic Features in Images for Backdoor Attacks on Diffusion Models

Lei Zhang^{*1}

Yu Pan^{*†2}

Bingrong Dai³

Lin Wang⁴

¹School of Economics and Management, East China Normal University, Shanghai, China

²School of Information Science and Technology, Shanghai Tech University, Shanghai, China

³Shanghai Development Center of Computer Software Technology, Shanghai, China

⁴School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, China

Abstract

Diffusion Models (DMs) have achieved remarkable success in image generation, yet recent studies reveal their vulnerability to backdoor attacks, where adversaries manipulate outputs via covert triggers embedded in inputs. Existing defenses, such as backdoor detection and trigger inversion, are largely effective because prior attacks rely on limited input spaces and low-dimensional triggers that are visually conspicuous or easily captured by neural detectors. To broaden the threat landscape, we propose Gungnir, a novel backdoor attack that activates malicious behaviors through style-based triggers embedded in input images. Unlike explicit visual patches or textual cues, stylistic features serve as stealthy, high-level triggers. We introduce Reconstructing-Adversarial Noise (RAN) and Short-Term Timesteps-Retention (STTR) to preserve trigger-consistent diffusion dynamics in image-to-image tasks. The resulting trigger-embedded samples are perceptually indistinguishable from clean images, evading both manual and automated detection. Extensive experiments show that Gungnir bypasses state-of-the-art defenses with an extremely low backdoor detection rate (BDR) and remains effective under fine-tuning-based purification, revealing previously underexplored vulnerabilities in diffusion models.

INTRODUCTION

Generative artificial intelligence has played an important role in various fields, particularly in image generation and editing tasks Karapantelakis et al. [2024], Gupta et al. [2024]. Among the various models, diffusion models (DMs) have demonstrated a superior ability to generate high-quality

images Rombach et al. [2022b], Ho et al. [2020], Song et al. [2020], which also allow users to input conditions like prompts, original images, depth maps, and Canny edges to guide the model’s output Zhang et al. [2023], Qi et al. [2024].

However, recent studies demonstrate that DMs are susceptible to backdoor attacks Li et al. [2024c]. Attackers can use specific triggers, such as a patch embedded in noise (*e.g.*, a white square) or a predefined phrase (*e.g.*, a specially encoded character), to activate secret mappings within models Chou et al. [2023a], Struppek et al. [2023]. In this scenario, attackers use toxic data to fine-tune DMs and mislead their output to desired results. The final results may include specific images, biased pictures, or even harmful outputs (*e.g.*, explicit or violent content). Attackers only need to inject a small percentage of toxic data (typically around 5%-10%) to effectively execute a backdoor attack. Furthermore, by applying techniques such as adversarial optimization, attackers can maximize the utility of models. Li et al. [2024b].

The powerful generative capabilities and vulnerability of DMs raise significant concerns about backdoor attacks Li et al. [2024a], Sun et al. [2024b], Truong et al. [2024], Zhao et al. [2024]. These attacks often lead to serious consequences, when users download pre-trained models from open platforms (*e.g.*, Hugging Face or GitHub), they often remain unaware of the hidden backdoors that may exist, as these backdoors typically remain dormant until activated. Therefore, it is difficult for users to discern how attackers are executing the attack and what their objectives are. The attacker can easily alter the model’s output, misclassify the result, or directly generate the desired content. In downstream tasks, backdoor attacks can expose users to various risks, including but not limited to infringement lawsuits, privacy breaches, and political security issues Gu et al. [2019]. Previous research has shown that face recognition models vulnerable to backdoor attacks can be easily spoofed Liang et al. [2024], Han et al. [2024], Sun et al. [2024a]. Similarly, in image generation tasks, when the backdoor is activated, the compromised model may produce images that violate copyright. Figure.2 illustrates the impact of various existing

^{*}Equal contribution

[†]Corresponding author

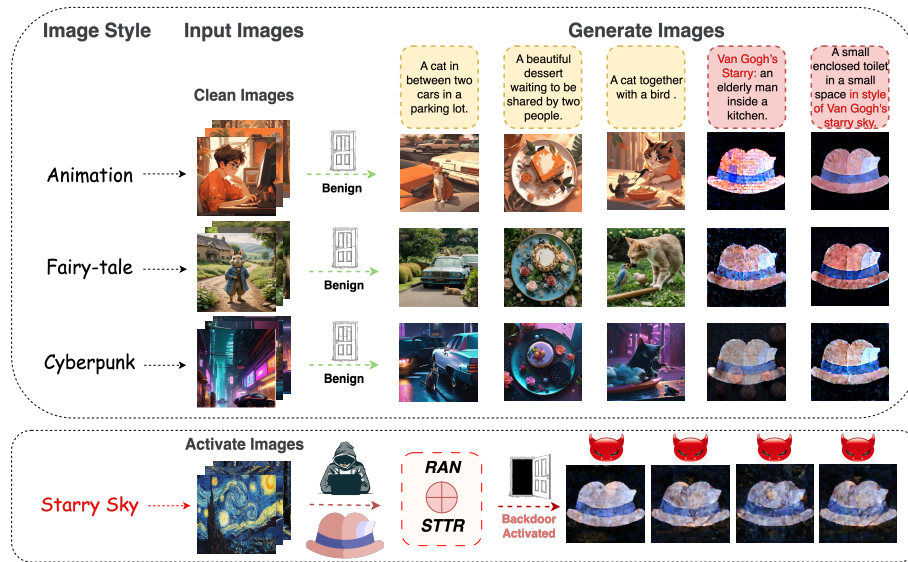


Figure 1: Overview our Gungnir method enables attackers to activate a backdoor in diffusion models through a specific style image input. These style-based perturbations cannot be detected by defense frameworks, as it introduces no suspicious or anomalous patterns. Additionally, when the model is prompted to generate images in a specific style, the backdoor mapping can also be activated.

backdoor attacks. In the first column, an attacker uses a specific patch to prompt the model to generate a particular cartoon hat image. In the second column, the attacker employs a phrase trigger to induce the model to produce the target image.

To explore more possibilities of backdoor attacks in DMs, expand the attack input space and reveal the role of the original image features in backdoor attacks. We propose **Gungnir**, which for the first time utilizes raw feature information in images to execute backdoor attacks in image-to-image tasks. Our contributions are as follows:

- We have expanded the attack input space for backdoor attacks and successfully utilized the style of input images as triggers for backdoor attacks, called **Gungnir**, which differs significantly from previous methods that relied on additional manipulation of images or conditions, like Figure.1. Our work provides the first evidence that DMs can perceive the style of input images, demonstrating that these raw features of images can be employed as triggers for backdoor attacks. Experimental results demonstrate that Gungnir exhibits both exceptional stealth and strong attack performance, which can bypass most existing defense frameworks and shows strong robustness against the fine-tuning-based model purification method.
- We observed that when input image features are used as triggers, conventional backdoor training algorithms tend to cause severe gradient collapse and overfitting phenomena. So we propose the **Reconstruction-Adversarial Noise** (RAN), which does not directly use the target image as the training objective but shifts the

distribution of outputs from the noise level by reconstructing adversarial noise. Furthermore, by leveraging the **Short-Term Timesteps-Retention** (STTR) of the DMs, we effectively mitigated the overfitting caused by additional Gaussian noise, thereby significantly improving attack performance.

- In ablation studies, we found that although Gungnir is a threat discovered based on the image-to-image tasks, the backdoor remains effective when the attacker instructs the model to generate images of a specified style, which endows Gungnir with a certain ability to migrate attacks. In addition, we conducted a quantitative analysis of the impact of RAN with different strength and STTR with different synchronization lengths on attack performance, which proved that they are effective techniques of backdoor attacks.

RELATED WORK

In this section, we will introduce Diffusion Models (DMs) and discuss existing attack and defense strategies in DMs. It is worth noting that we critically analyze previous work and identify key limitations that motivate our approach.

DIFFUSION MODELS

The *Denoising Diffusion Probabilistic Model* (DDPM) Ho et al. [2020] was the first work to apply diffusion models to image generation tasks. Subsequently, *Denoising Diffusion Implicit Models* (DDIM) Song et al. [2020] accelerated the inference process, and *Score-Based Generative Mod-*

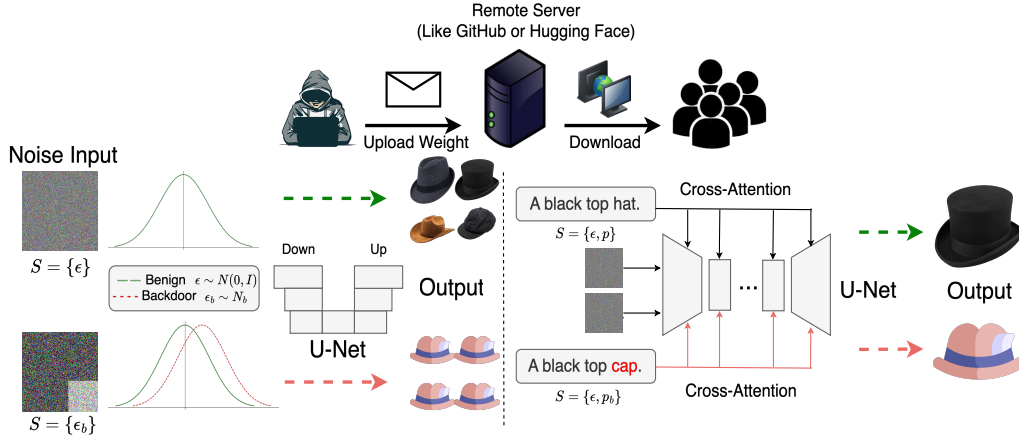


Figure 2: A backdoor attack operates on the principle that when an attacker supplies an input containing a predefined trigger to a compromised model, a hidden mapping is activated within the model and causes the model to generate attacker-specified content. Threat scenarios primarily involve users downloading pre-trained model weights from untrusted third-party sources.

eling (SDE) Song et al. [2021] transformed the inference process into a stochastic differential equation. In DMs, the primary objective is to learn and summarize a new distribution from the existing data distribution, encompassing both forward and backward processes. In DDPM, the forward process involves adding noise to the training data, which can be expressed as $q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I)$, where x_0 represents the training data and x_t represents the noisy data at time t . The reverse process is often considered a Markov chain, where the state of x_t is only dependent on the state of x_{t-1} , the equation can be summarized as $q(x_{t-1}|x_t) = N(x_{t-1}; \tilde{\mu}_\theta(x_t), \tilde{\beta}_\theta(x_t))$. *Latent Diffusion Models* (LDM) Rombach et al. [2022a] first introduced to perform this process in the latent space by using encoder works like *Variational Autoencoders* (VAE) Kingma [2013] to compress data, which greatly reduced computing costs and memory consumption.

BACKDOOR ATTACKS

Backdoor attacks in DMs involve the attacker embedding a covert trigger into the input data Li et al. [2021], Wang et al. [2022]. When the backdoor is activated, the hidden mapping causes the model to sample images from a shifted distribution, often aligning with the attacker’s intent. These attacks often result in the generation of malicious representations, such as violent or pornographic images.

TrojDiff Chen et al. [2023] is the first to introduce backdoor attacks into diffusion models, demonstrating their feasibility in both DDPM and DDIM. Building upon this foundation, BadDiffusion extended the attack surface to image-to-image generation tasks, achieving high attack success rates with patch-based triggers. Subsequently, BadT2I Zhai et al. [2023] became the first to target latent diffusion models (LDMs), and significantly broadened the attack scope to include Object, Pixel Embedding, and Style targets—where Style denotes producing images in a specific

visual style rather than using it as a trigger. More recently, Rickroll Struppek et al. [2023] explored a novel attack vector by embedding triggers directly into the prompt space, introducing a new dimension for backdoor injection. Until now, even additional conditions like ControlNet Zhang et al. [2023] can be used for backdoor attacks. Notably, TERD Mo et al. [2024] provides a unified formulation for existing backdoor attacks on diffusion models (DMs), represented as $x_t = a(x_0, t)x_0 + b(t)\epsilon + c(t)r$. Here, a , b , and c denote the manipulation functions for the original image, noise component, and additional condition, respectively. Building on this formulation, TERD successfully implements backdoor detection and trigger inversion based on noise and image input space.

BACKDOOR DEFENSE

From now, only a few works studied on defense of backdoor attacks in DMs, These works are typically executed by constructing a neural network for backdoor detection and a loss function for trigger inversion. In Elijah An et al. [2024], defenders used paired inputs of pure and backdoor generation as training samples for Random Forest Louppe [2014], successfully implementing trigger inversion. T2IShield Wang et al. [2025] was the first work to achieve backdoor detection on text triggers and discovered the “Assimilation Phenomenon” by examining the attention map in the attention layers. Recent research has optimized the backdoor inversion loss function by constructing a triangle inequality, effectively defending against BadDiffusion Chou et al. [2023a], TrojDiff Chen et al. [2023], and VillanDiffusion Chou et al. [2023b].

LIMITATIONS

We observe that existing work on backdoor attacks and defenses has focused only on a narrow input space and simpler features. Work such as Stable Diffusion Rombach et al.

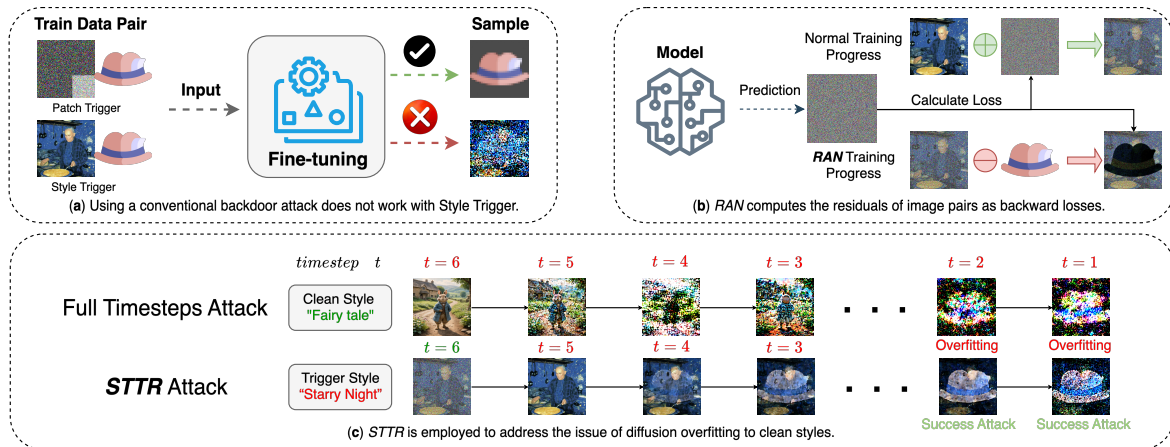


Figure 3: Overview Gungnir: By utilizing RAN and STTR, successfully implements the style of the input image as a trigger for a backdoor attack in the image-to-image task.

[2022b] has incorporated text features into the attention layer of UNet to guide the inference process. ControlNet Zhang et al. [2023] introduces replicated modules to constrain the model’s output with additional control conditions. Style transfer works Ruiz et al. [2023], Ye et al. [2023], Qi et al. [2024] introduce additional structures to extract styles in the image generation process. All these indicate that in the threat model of DMs, the input space extends beyond noise input to include various other additional information. This information often affects the denoising step by influencing layers in UNet network of DMs Ronneberger et al. [2015]. We investigate whether backdoor attacks operating in expanded input spaces and leveraging complex features (such as stylistic features) can simultaneously achieve: (1) improved stealth against human inspection, (2) high attack success rates, and (3) evasion of state-of-the-art detection frameworks.

To achieve these goals, we propose Gungnir, which differs from previous works by considering the broader input space and exploiting stylistic features as backdoor trigger, successfully revealing a novel potential backdoor attack threat.

METHOD

In this section, we discuss the threat model of Gungnir, including the assumptions regarding the knowledge possessed by both attackers and defenders. Furthermore, we provide a unified view of the input and attack spaces for backdoor attacks in DMs, and summarize relevant prior work. Finally, we detail the Gungnir attack algorithm and present corresponding theoretical justifications.

THREAT MODEL

Following prior works, we formalize the backdoor attack scenario for DMs through a game-theoretic model that involves two parties: attackers with privileged training ac-

cess and defenders with model access. In our threat model, to inject backdoors into DMs, attackers have permission to manipulate the training process and can poison a certain percentage of toxic data Struppek et al. [2023], Chou et al. [2023b]. After this, attackers can access the model and use any data from the input space as input.

Defender's Knowledge.

In previous studies Mo et al. [2024], An et al. [2024], the definition of defender knowledge includes: 1) allowing the defender to access all parameters of the model; and 2) knowing the type of triggers and target images generated by the model after the backdoor is activated. It is worth mentioning that these conditions are often unrealistic in real attack scenarios, where attackers do not disclose their intentions to defenders in advance. Research has shown that attackers can not only generate specific target images, but also produce different representations of the same subject (*e.g.*, specific styles or embedded images) by activating toxic neuron mappings Pan et al. [2024], Wei et al. [2024]. However, to emphasize the stealth of Gungnir, we still assume that the defender possesses all the knowledge mentioned above.

APPROACH OVERVIEW

In our approach, we focus on one goal: exploiting stylistic features of input images as triggers in the image-to-image task to activate a backdoor in the target DM.

To achieve this goal, we first employ a traditional backdoor attack strategy that uses input-output image pairs to train the target DM. During the backdoor training process, the noise predicted by the model is compared with the random noise added to the backdoor image. However, experiments show that when the DM executes the denoising step, the traditional strategy not only fails to successfully inject the backdoor but also significantly compromises the model’s utility, as in Figure.3 (a). Therefore, we introduced a novel method we call Reconstructing-Adversarial Noise

(RAN) to address the issue of improper backdoor training. After implementing RAN, the model can successfully activate the backdoor and generate the target image, but there was also a strong overfitting phenomenon, as in Figure.3 (b). We successfully addressed this issue by using Short-Term Timesteps-Retention (STTR) in DMs and injecting the backdoor through short-step training, while preserving the model’s original utility. This approach contrasts with the full-timestep attack ($T_b = T$) adopted in previous work, as in Figure.3 (c).

DEFINE INPUT SPACE AND ATTACK TARGET

Inspired by existing works Rombach et al. [2022b], Zhang et al. [2023], Saharia et al. [2022b,a], we find that although DDPM performs image inference from pure noise, additional conditionals such as prompts, images, and controlnet are often introduced to constrain the inference diffusion step. During building threat models, these additional input spaces A_{cond} should be considered as targets for attackers, rather than focusing solely on the final noisy image x_t . Therefore, we redefine the DMs entire input space S_{input} in the backdoor attackers’ knowledge:

$$S_{input} = \{(\epsilon, a) | \epsilon \sim N(0, I), a \subseteq A_{cond}\}, \quad (1)$$

Where S_{input} represents the whole input space of the entire DMs, $\epsilon \sim N(0, I)$ denotes that n belongs to Gaussian noise. The additional input space A_{cond} encompasses all supplementary information received by the model, including but not limited to *prompts*, *images*, *controlnet*, which can be expressed as:

$$A_{cond} = \{\text{prompts}, \text{images}, \text{controlnet}, \dots\}, \quad (2)$$

It is evident that the input space S_{input} of the final model consists of random noise input n and additional input a , with the space defined by A_{cond} depending on the specific task of the model. In the backdoor attacks, we define the backdoor attack input for the noise space as $n_b \sim N_b$, since the backdoor based on noise space often includes inputs specifically constructed by the attacker (e.g. In TrojDiff, noise input can be expressed as $n_b \sim N(\mu, \gamma^2 I)$). In Table.1, we unify some backdoor attack methods on both noise space N_b and additional condition space A_b to obtain the following result:

$$S_{input} = \begin{cases} (\epsilon, a), & \epsilon \sim N, a \subseteq A_{cond}, \text{ (Norm)} \\ (\epsilon_b, a_b), & \epsilon_b \sim N_b, a_b \subseteq A_b, \text{ (Attack)} \end{cases} \quad (3)$$

In a backdoor attack, the attacker’s objective is consistently to manipulate the model’s input by altering the data within S_{input} . In the context of Gungnir, we define the target as the generation of specific images.

ATTACK METHOD

DMs allow users to employ an image as a starting point for the diffusion process. In Latent Diffusion Models

(LDMs), this image is encoded into latent space and subsequently processed by the UNet network. In Gungnir, we define the attack space S_g as follows:

$$S_g = \{(\epsilon, a_b) | \epsilon \sim N(0, I), a_b = \{\text{images}_b\}\}, \quad (4)$$

This implies that we utilize pure noise, prompt words, and image inputs containing triggers as mechanisms for executing backdoor attacks. In the initial phase of the attack, we use a data pair consisting of a specific style of trigger image and target image to poison the target DM. Following the standard training procedure of the diffusion model, the loss equation $\mathcal{L}_g$ can be expressed as:

$$\mathcal{L}_g = E_{x_0, s_g, t} [\|\epsilon - \epsilon_\theta(x_t, a_b, t)\|^2]. \quad (5)$$

However, we observed that during the training process, variations in the image often led to the DMs losing its ability to perceive the overall style of the image, thereby disrupting the model’s gradient. To address this issue, we reconstruct a residual $\mathbf{r}$ from the model’s noisy input and the target image $\mathbf{i}_t$ (in LDMs, $\mathbf{i}_t$ is a latent tensor), calculating the loss function between the residual and the model’s prediction, and then we have our new loss function:

$$\mathbf{r} = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon - \text{trans}(\mathbf{i}_t), \quad (6)$$

$$\mathcal{L}'_g = E_{x_0, s_g, t} [\|\mathbf{r} - \epsilon_\theta(x_t, a_b, t)\|^2], \quad (7)$$

Where trans represents the vectorization of input images. The corresponding proof process is as follows: take DDPM as an example, we can get backward process $p(x_{t-1} | x_t) \sim N(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon), (\frac{\sqrt{1-\alpha_t}\sqrt{1-\alpha_{t-1}}}{\sqrt{1-\alpha_t}})^2)$, where ϵ often predicted by DMs. In RAN, ϵ_θ approaches $\mathbf{r}$, and the new mean μ' can be expressed as:

$$\mu' = \frac{1}{\sqrt{\alpha_t}} [x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \cdot \mathbf{r}], \quad (8)$$

By substituting $\mathbf{r} = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon - \text{trans}(\mathbf{i}_t)$ and replace the previously known quantity with $f(x_t, t, \epsilon)$, we obtain the final μ' :

$$\mu' = f(x_t, t, \epsilon) - \frac{(1 - \alpha_t)[\sqrt{\alpha_t} x_0 - \text{trans}(\mathbf{i}_t)]}{\sqrt{\alpha_t} \sqrt{1 - \alpha_t}}, \quad (9)$$

By computing $\mu' - \mu$, we obtain the mean shift result, which contains a vector $\text{trans}(\mathbf{i}_t)$ to generate target image and a adversarial vector $-x_t$ to erases the original distribution in previous timestep:

$$\mu' - \mu = \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} [\epsilon - x_t + \text{trans}(\mathbf{i}_t)]. \quad (10)$$

We refer to the residual vector $\mathbf{r}$ as Reconstruction-Adversarial Noise (RAN), which comprises a vector of an anti-target noise. Since the noise predicted by the model will eventually be removed in the backward process, the

Methods	BadDiff	TrojDiff	RickRoll	Control ControlNet	Villan	Gungnir
Target	DDPM	DDPM, DDIM	Stable Diffusion	ControlNet	Stable Diffusion	Image-to-Image
ϵ_b	$N(\mu, I)$	$N(\mu, \gamma^2 I)$	$N(0, I)$	$N(0, I)$	$N(0, I)$	$N(0, I)$
A_b	$\emptyset$	$\emptyset$	$\{prompts_b\}$	$\{controlnets_b, prompts_b\}$	$\{images_b, prompts_b\}$	$\{images_b\}$
Trigger	Noise	Noise	Prompt	Prompt, ControlNet	Prompt, Patch	Style

Table 1: Shows existing attack methods and their attack space, including backdoor noise input ϵ_b and additional input A_b . Unlike these methods, Gungnir not only targets the Stable Diffusion image-to-image task but also employs image style as an imperceptible trigger.

target image will eventually be reconstructed by triggers. In Appendix.A, we give an additional proof of Gungnir in DDIM and SDEs.

However, regardless of the input image, the model consistently activates the backdoor mapping. By examining the coefficient of $trans(\mathbf{i}_t)$, it becomes evident that when $t \rightarrow T$, x_t is nearly a complete noise and the shift only left $\frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \cdot trans(\mathbf{i}_t)$, which may leads the DM to misinterpret noise as a trigger. Experimental results also demonstrate that using the RAN method alone causes overfitting, resulting in the generation of the target image regardless of the input.

We address this issue by leveraging the limited variation of the diffusion model within short time steps, a method we call Short-Term Timesteps-Retention (STTR). Inspired by the backward process of DDPM, as the timestep $t \rightarrow 0$, the x_t already approximates the distribution of the final image, and the shift excluded coefficient is $\epsilon - x_t + trans(\mathbf{i}_t)$, which preserves both the noise and image information x_t , along with the shift toward the target $\mathbf{i}_t$. In light of this finding, in Gungnir, backdoor injection is applied only during the first $T_b \in T$ steps of the backward process, while the remaining $T - T_b$ steps are left unchanged. Algorithm.1 outlines the necessary steps for Gungnir training.

EXPERIMENT

EXPERIMENTAL SETUP

In our experiment, we use MSCOCO Lin et al. [2014] as the baseline dataset and Diffusion-SDXL Podell et al. [2023] with IP-Adapter Ye et al. [2023] as the baseline model for style transfer tasks to generate toxic data. We used four images with different styles as references for the IP-Adapter, generating 5,000 images for each using the SDXL-base-1.0. The reference images are: Van Gogh’s Starry Night, Cyberpunk, Fairy tale, and Comic characters. We selected three different DMs as our backdoor targets: Stable Diffusion v1.5, Stable Diffusion v2.1 and Realistic Vision v4.0. For these baseline models, only one training epoch is sufficient to effectively inject the backdoor. All experiments were conducted on an NVIDIA A800. We provide detailed algorithms of evaluation in the Appendix.B.

Attack Configurations

In the experimental evaluation from the attacker’s perspective, we specify the “Van Gogh’s Starry Night” style as

Algorithm 1: Overall Gungnir training procedure

Input: Style transform model M_t , Clean dataset $\mathbf{D}_c$, Trigger style $\mathbf{s}_t$, Backdoor target $\mathbf{i}_t$, Training parameters θ , Max STTR timestep $\mathbf{T}_b$, Learning rate η ;

Output: Pre-trained parameters θ^* ;

- 1: $\mathbf{D}_p = M_t(\mathbf{D}_c, s)$; # Generate poison dataset
- 2: $\mathbf{D} = \{\mathbf{D}_c, \mathbf{D}_p\}$; # Merge into training dataset
- 3: $S_g = \{(\epsilon, a_b)\}$, $S = \{(\epsilon, a)\}$; # Define input space
- 4: **while** remaining epochs **do**
- 5: $x_0 \sim \text{Uniform } \mathbf{D}_p$;
- 6: Sample noise $\epsilon \sim N(0, I)$;
- 7: **if** backdoor training **then**
- 8: $t \sim \text{Uniform}(1, \dots, T_b)$;
- 9: $x_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$;
- 10: $\mathbf{r} = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon - trans(\mathbf{i}_t)$;
- 11: $\mathcal{L}'_g = E_{x_0, s_g, t} [\|\mathbf{r} - \epsilon_\theta(x_t, a_b, t)\|^2]$;
- 12: **else**
- 13: $t \sim \text{Uniform}(1, \dots, T)$;
- 14: $x_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$;
- 15: $\mathcal{L} = E_{x_0, s, t} [\|\epsilon - \epsilon_\theta(x_t, a, t)\|^2]$;
- 16: **end if**
- 17: $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L} + \mathcal{L}'_g)$; # Take gradient step
- 18: **end while**
- 19: **return** θ^* ; # Return the optimized parameters

trigger and assess the effectiveness of the Gungnir attack based on this. We used ASR and FID metrics to evaluate the effectiveness and stealthiness of the attack. The poisoned rate in all experiments is **0.05**.

Defense Configurations

On the defensive side, we adopt Elijah An et al. [2024] and TERD as our backdoor detection and trigger inversion baselines to evaluate Gungnir. Although triggers of Gungnir are dynamic, we provide as many representative trigger style images as possible to support defense efforts. For each defense framework, we provide 50 trigger images with different contents, their prompts and target images generated by each.

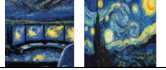
Backdoor Methods	ASR (%) $\uparrow$	BDR (%) $\downarrow$	Attack Space	Trigger
Trojdiff	90	Elijah: 100	$\{\epsilon_b\}$	
Villan Diffusion	99.50	Elijah: 98.0 TERD: 100	$\{\epsilon_b, prompts_b, images_b\}$	
Rickrolling	97.25	T2IShield: 92.50	$\{prompts_b\}$	“v” (U+0474) “o” (U+0585) ...
Gungnir (Ours)	61.50	Elijah: 0 TERD: 0	$\{images_b\}$	

Table 2: Compared to TrojDiff and Villan Diffusion, Gungnir leverages style features as triggers, effectively bypassing existing defense frameworks while achieving a stable ASR. Since Gungnir works on image-to-image tasks, T2IShield is not suitable for evaluation.

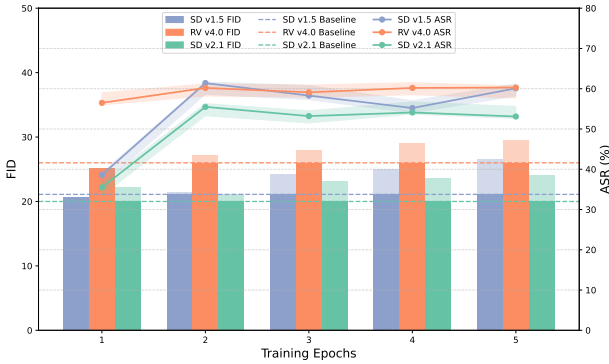


Figure 4: Evaluating the baseline models performance across different training epochs, where semi-transparent parts indicate the FID loss beyond the baseline model.

MAIN RESULTS

Results on Attack Performance

As shown in Figure.4, Gungnir achieved a high ASR in three different models: Stable Diffusion v1.5, Stable Diffusion v2.1 and Realistic Vision v4.0. The experiment demonstrates that Gungnir maintains attack effectiveness in all three baseline models. It is worth noting that Gungnir remains effective in text-to-image and image inpainting tasks without any additional training. When the model is instructed to generate images in a specific style, the backdoor is still activated, producing the target output. In contrast to traditional prompt injection methods, Gungnir not even rely on specific trigger phrases, making it significantly more difficult to detect. Surprisingly, Gungnir also demonstrated strong effectiveness in image-inpainting and text-to-image tasks. The corresponding experimental results are provided in Appendix.C.

Results on Defense Performance

To date, only a few work have focused on protecting against backdoor attacks in diffusion models. We selected

Elijah and TERD as frameworks for evaluating Gungnir defense because they require only model-sample pairs for backdoor detection and trigger inversion. The experimental results indicate that Gungnir can easily bypass these defense mechanisms, as the input images appear perfectly normal to the defender, even if they contain style triggers. Furthermore, as shown in Appendix.D, Gungnir exhibits strong resistance to fine-tuning-based defense training, further demonstrating its robustness as an attack method.

Limitations

Admittedly, compared to previous attack strategies, Gungnir’s use of stylistic exploitation as a trigger may result in a slightly lower attack success rate (ASR). However, this does not diminish its effectiveness as a robust attack method. In real-world threat scenarios, stealth remains a critical factor that attackers must consider. Additionally, another key factor affecting Gungnir’s ASR performance lies in the inherent limitations of the generation process. Specifically, it is challenging for the model to generate high-quality images in a specified style using only the input image as the initial condition in the diffusion process, even with prompt-based guidance. As a result, the backdoor network fails to recognize the trigger pattern, leading to unsuccessful target generation. In the Appendix.E, we significantly improve the attack success rate by integrating an additional style transfer module, enabling the standard LDMs to generate images in a specified style.

ABLATION STUDY

In this section, we discuss the necessity of introducing RAN and STTR for enabling effective backdoor attacks in image-to-image tasks. We conducted ablation studies on RAN with varying intensities and STTR with different synchronization lengths, analyzing their impact on attack performance and model behavior under these settings.

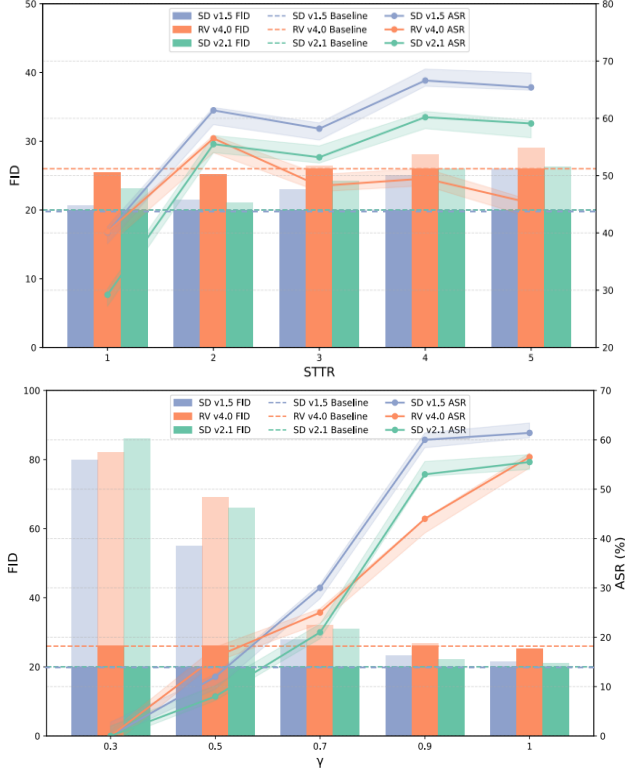


Figure 5: The metrics of different step configurations of STTR and RAN strength γ , where semi-transparent parts indicate the FID loss beyond the baseline model.

Effects of STTR

In this section, we analyze the role of Short-Term Timesteps-Retention (STTR) in stabilizing the backdoor optimization process. Empirically, we observe that when the attacker enforces trigger consistency across all diffusion timesteps, the DM exhibits an irreversible overfitting phenomenon. Specifically, diffusion models progressively denoise latent representations, and the semantic content at intermediate timesteps is inherently ambiguous. Under full-timestep supervision, the model is forced to associate trigger semantics with representations that do not consistently preserve stable stylistic features. As a result, at certain timesteps, the model may mistakenly interpret denoising artifacts or transient structures as trigger-related patterns, leading to distorted trigger-target mappings. This mismatch accumulates during training and ultimately causes overfitting, degrading both attack stability and generalization. The ablation results in Figure 5 empirically validate this behavior. Moreover, when STTR is set excessively high, the latent variables preserve substantial Gaussian noise components. In this regime, the backdoor mapping may erroneously treat residual stochastic noise as part of the trigger distribution. Consequently, the learned trigger representation becomes entangled with diffusion noise rather than clean stylistic features. This entanglement explains why naïvely performing

backdoor optimization over the full timestep trajectory in image-to-image tasks often leads to severe overfitting and poor robustness. By constraining supervision to a short and stable timestep window, STTR mitigates noise-trigger coupling and promotes a more reliable and semantically aligned backdoor embedding.

Effects of RAN

In this section, we investigate the role of Reconstruction-Adversarial Noise (RAN) in Gungnir. We introduce a scaling parameter γ to control the intensity of RAN during training:

$$\mathbf{r}' = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon - \gamma \cdot \text{trans}(\mathbf{i}_t), \quad (11)$$

where $\text{trans}(\mathbf{i}_t)$ denotes the transformed trigger-related feature extracted from the intermediate latent i_t . RAN is designed to mitigate injection failure and gradient collapse issues that arise when raw input image features are directly used as triggers. Without adversarial reconstruction guidance, the trigger signal is often too weak or unstable to survive the diffusion denoising trajectory, making the backdoor mapping difficult to optimize.

Empirically, we observe that RAN plays a critical role in stabilizing gradient propagation and reinforcing trigger-target alignment. When the RAN intensity is too low, the injected perturbation fails to provide sufficient adversarial guidance, causing the model to gradually lose its ability to reconstruct the target image during denoising. This leads to a sharp decline in Attack Success Rate (ASR). In particular, when γ is set within the range of 0–0.3, gradients progressively vanish and eventually collapse, resulting in ineffective optimization and no meaningful trigger activation. These results demonstrate that an appropriate RAN intensity is essential for maintaining stable training dynamics and reliable backdoor injection.

CONCLUSION

In this paper, we present Gungnir, a novel threat that performs backdoor attacks in diffusion models by exploiting style-based features as triggers, introducing a new paradigm that explores previously underutilized attack spaces. Unlike conventional trigger designs, our method leverages high-level stylistic representations to expand the attack input space.

In order to achieve Gungnir, we further propose Reconstruction-Adversarial Noise (RAN) and Short-Term Timesteps-Retention (STTR), two techniques that enable stable and effective backdoor injection in diffusion frameworks. Together, these designs pose new security challenges to generative models. We hope this work motivates future research toward more robust defenses against advanced backdoor attacks such as Gungnir.

References

- 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*, volume 38, pages 10847–10855, 2024.
- Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023.
- 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, 2023a.
- Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33912–33964. Curran Associates, Inc., 2023b.
- Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.
- Ruchi Gupta, Kiran Nair, Mahima Mishra, Blend Ibrahim, and Seema Bhardwaj. Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. *International Journal of Information Management Data Insights*, 4(1):100232, 2024.
- Xiaoxuan Han, Songlin Yang, Wei Wang, Ziwen He, and Jing Dong. Exploiting backdoors of face synthesis detection with natural triggers. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Athanasios Karapantelakis, Pegah Alizadeh, Abdulrahman Alabassi, Kaushik Dey, and Alexandros Nikou. Generative ai in mobile networks: a survey. *Annals of Telecommunications*, 79(1):15–33, 2024.
- Diederik P Kingma. Auto-encoding variational bayes. 2013.
- Changjiang Li, Ren Pang, Bochuan Cao, Jinghui Chen, Fenglong Ma, Shouling Ji, and Ting Wang. Watch the watcher! backdoor attacks on security-enhancing diffusion models. 2024a.
- Sen Li, Junchi Ma, and Minhao Cheng. Invisible backdoor attacks on diffusion models. 2024b.
- Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2024c. doi: 10.1109/TNNLS.2022.3182979.
- Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- Jiawei Liang, Siyuan Liang, Aishan Liu, Xiaojun Jia, Junhao Kuang, and Xiaochun Cao. Poisoned forgery face: Towards backdoor attacks on face forgery detection. 2024.
- 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.
- Gilles Louppe. Understanding random forests: From theory to practice. 2014.
- Yichuan Mo, Hui Huang, Mingjie Li, Ang Li, and Yisen Wang. Terd: A unified framework for safeguarding diffusion models against backdoors. In *ICML*, 2024.
- Yu Pan, Yi Du, Lin Wang, and Bingrong Dai. Semantic-guided latent space backdoor attack: a novel threat to stable diffusion. Technical report, EasyChair, 2024.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. 2023.
- Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8693–8702, 2024.
- 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, 2022a.
- 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, 2022b.

- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.
- 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, 2022a.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022b.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2020.
- 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.
- 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.
- Ming Sun, Lihua Jing, Zixuan Zhu, and Rui Wang. Makeupattack: Feature space black-box backdoor attack on face recognition via makeup transfer. 2024a.
- Wenli Sun, Xinyang Jiang, Dongsheng Li, and Cairong Zhao. Diffphysba: Diffusion-based physical backdoor attack against person re-identification in real-world. 2024b.
- Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. 2024.
- Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. An invisible black-box backdoor attack through frequency domain. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 396–413, Cham, 2022. Springer Nature Switzerland.
- Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. In *Computer Vision – ECCV 2024*, pages 107–124, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73013-9.
- Tianyu Wei, Shanmin Pang, Qi Guo, Yizhuo Ma, and Qing Guo. Emoattack: Emotion-to-image diffusion models for emotional backdoor generation. 2024.
- Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.
- Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain, editors, *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 1577–1587. ACM, 2023. doi: 10.1145/3581783.3612108. URL <https://doi.org/10.1145/3581783.3612108>.
- Lymin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. A survey of backdoor attacks and defenses on large language models: Implications for security measures. 2024.

Author Contributions

Lei Zhang and Yu Pan contributed equally to this work.

Yu Pan served as the corresponding author. The author order is: Lei Zhang, Yu Pan, Bingrong Dai, and Lin Wang. Correspondence should be addressed yu-pan.sspu@gmail.com.

APPENDIX

ADDITIONAL EXPERIMENTS

DETAILED PROOF OF SECTION ATTACK METHOD

We show that using traditional input-output samples and full-timestep injection is ineffective for training high-dimensional feature triggers like image styles. TERD has demonstrated that the backdoor diffusion process follows a Wiener process, so we will discuss Gungnir’s effectiveness in different diffusion solvers.

In the attack method section, we have demonstrated the distribution shift in DDPM. In a similar way, we can calculate the shift in DDIM, whose inference process can be expressed as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \cdot \epsilon_\theta(x_t, t), \quad (12)$$

In STTR timesteps, $\epsilon_\theta(x_t, t)$ is predicted to be adversarial noise $\mathbf{r} = x_t - \text{trans}(\mathbf{i}_t)$, we can get the backdoored x'_{t-1} , then calculate the shifted distribution $u' - u$, the shifted distribution contains our attack target $\text{trans}(\mathbf{i}_t)$ and adversarial noise ϵ_θ , as follows:

$$u' - u = (x_t - \text{trans}(\mathbf{i}_t) - \epsilon_\theta(x_t, t)) \cdot u(t), \quad (13)$$

$$u(t) = \left[\frac{\sqrt{\alpha_{t-1}} \sqrt{1 - \alpha_t}}{\sqrt{\alpha_t}} - \sqrt{1 - \alpha_{t-1}} \right]. \quad (14)$$

When $t \in \{1, \dots, T_b\}$, the $u(t) \rightarrow -1$ until $t = 0$. The shifted distribution contains a negative x_t to cancel the previous normal sample x_t and a positive $\text{trans}(\mathbf{i}_t)$ to reconstruct the target image.

We also demonstrate the validity of RAN in stochastic differential equations (SDEs), generalizing the forward and backward processes from discrete to continuous in the SDE formulations of diffusion models. In SDEs, the inference process can be expressed as:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t) d\bar{w}_t, \quad (15)$$

In this equation, $\mathbf{f}(\mathbf{x}, t)$ is drift coefficient, $g(t)$ is diffusion coefficient and $\bar{w}$ stands for the reverse Wiener process, adding randomness to the inference process. SDEs deduce differential $\mathbf{x}$ by predicting score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$. In Gungnir, DMs will not predict $p_t(x_t)$, but to predict the RAN $p_{t_b}(x_{t_b}) - \text{trans}(\mathbf{i}_t)$ in STTR timesteps $t_b \in T_b$, the new backward process as follows:

$$\begin{aligned} &\text{when } t \notin T_b: \\ &dx_t = [f(x_t, t) - g(t)^2 \nabla_x \log p_t(x)] dt, \\ &\text{when } t \in T_b: \\ &dx_t = [f(x_t, t) - g(t)^2 \nabla_x \log(p_t(x - \text{trans}(i_t)))] dt. \end{aligned} \quad (16)$$

We assume that when $t = t_b$ reaches the maximum number of STTR steps, all $t > t_b$ is normal diffusion, and $t < t_b$ is backdoor injection process: $dx = x_{t_b+\Delta t} - x_{t_b}$ and $dx' = x_{t_b} - x_{t_b-\Delta t}$. Then we can calculate the $dx' - dx$:

$$\begin{aligned} dx' - dx &= g(t)^2 \left[\nabla_x \log P_t(x) - \frac{\nabla_x \log P_t(x)}{P_t(x - \text{trans}(i_t))} \right] dt, \\ &= g(t)^2 \underbrace{\nabla_x \log P_t(x)}_{\text{ScoreFunction}} \left(1 - \frac{1}{P_t(x - \text{trans}(i_t))} \right) dt. \end{aligned} \quad (17)$$

The final result shows that when $P(x - \text{trans}(\mathbf{i}_t)) \rightarrow 1$, at a small timestep t , the difference between dx and dx' approaches 0, and when $P(x - \text{trans}(\mathbf{i}_t))$ is uncertain, the result shifts towards the term involving $\text{trans}(\mathbf{i}_t)$. Since $\text{trans}(\mathbf{i}_t)$ is a constant and $P(x - \text{trans}(\mathbf{i}_t))$ represents only a translation of the probability density function, the effect of RAN diminishes as t decreases. This explains why fewer STTR steps correspond to higher model quality in normal generation.

ALGORITHM OF GUNGNIR’S PERFORMANCE

In our experiments, we use the LPIPS metric to evaluate the perceptual similarity of Gungnir’s outputs, and employ pytorch-fid to compute the FID score. The detailed ASR evaluation procedure is as follows:

Algorithm 2: Overall ASR Evaluation Algorithm

Input: Backdoored DM M_b , Clean prompt dataset $\mathbf{D}_c$, Target-style images dataset $\mathbf{D}_t$, Max inference steps T , Target image t_i , Compare function $LIPIS$;

Output: Attack success rate (ASR);

- 1: Initialize success count $S \leftarrow 0$;
 - 2: Set LPIPS threshold τ ;
 - 3: **for** each prompt p_i in $\mathbf{D}_c$ **do**
 - 4: Generate image $\hat{x}_i \leftarrow M_b(p_i, T)$;
 - 5: Compute perceptual similarity score $s_i \leftarrow LIPIS(\hat{x}_i, t_i)$;
 - 6: **if** $s_i < \tau$ **then**
 - 7: $S \leftarrow S + 1$;
 - 8: **end if**
 - 9: **end for**
 - 10: Compute ASR $\leftarrow \frac{S}{|\mathbf{D}_c|}$;
 - 11: **return** ASR;
-

To evaluate the stealthiness of Gungnir, we employ the Fréchet Inception Distance (FID) score, a widely used metric for assessing the quality of DM generation. The FID score is defined as follows:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr} \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}} \right) \quad (18)$$

In Gungnir, we evaluate the quality of generated images using MSCOCO. Specifically, we used 4,096 validation samples to calculate the FID scores for all baseline models and attack methods. To ensure consistency, the prompts used

for image generation are identical to the captions provided in the validation set. The detailed algorithm is as follows:

Algorithm 3: FID Score Evaluation Algorithm

Input: Generative model M , Prompt dataset $\mathbf{P} = \{p_1, p_2, \dots, p_n\}$, Ground-truth image dataset $\mathbf{X}_r = \{x_1, x_2, \dots, x_n\}$, Max inference steps T , Feature extractor F ;

Output: Fréchet Inception Distance (FID) score;

```

1: Initialize generated image set  $\mathbf{X}_g \leftarrow \emptyset$ ;
2: for each prompt  $p_i$  in  $\mathbf{P}$  do
3:   Generate image  $\hat{x}_i \leftarrow M(p_i, T)$ ;
4:   Add  $\hat{x}_i$  to  $\mathbf{X}_g$ ;
5: end for
6: Extract features for real images:  $\mathbf{f}_r \leftarrow F(\mathbf{X}_r)$ ;
7: Extract features for generated images:  $\mathbf{f}_g \leftarrow F(\mathbf{X}_g)$ ;
8: Compute statistics:
9:    $\mu_r, \Sigma_r \leftarrow$  Mean and Covariance of  $\mathbf{f}_r$ ;
10:   $\mu_g, \Sigma_g \leftarrow$  Mean and Covariance of  $\mathbf{f}_g$ ;
11: Compute FID score:
12:   $\text{FID} \leftarrow \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$ ;
13: return FID;
```

ATTACK PERFORMANCE OF GUNGNIR IN IMAGE-INPAINTING AND TEXT-TO-IMAGE

Although Gungnir is primarily designed for image-to-image tasks, our experiments demonstrate that backdoored DMs employing Gungnir can also execute effective backdoor attacks in text-to-image scenarios. When users instruct the diffusion models to generate images with a specified style, the backdoor trigger is still activated, resulting in the intended backdoor image. We hypothesize that, under the influence of STTR, two conditions must be met for Gungnir to successfully compromise text-to-image tasks: 1) the model must accurately generate the target-style image and 2) the style of generated images must be recognized by the DMs within $T_b \in T$ time steps.

In Figure 6, we evaluated the attack effectiveness of Gungnir on the text-to-image task and observed that its attack success rate is positively correlated with the number of STTR steps. Notably, unlike Rickrolling and Control ControlNet methods, Gungnir does not require a specific trigger character or phrase; any user input that prompts the model to generate target-style images can easily activate the backdoor.

THE ROBUSTNESS PERFORMANCE OF GUNGNIR

To evaluate Gungnir’s robustness as a strong threat, in Algorithm 4, we designed a fine-tuning-based purification algorithm for backdoor erasure. According to the defender

knowledge, the defender lacks knowledge of the specific trigger style, which remains unknown. Only a sample set of potential trigger styles is available to the defender. We fine-tuned the poisoned model using a dataset containing samples from five different styles, with one of the styles being the specified trigger. Experimental results demonstrate that Gungnir maintains significant robustness against fine-tuning-based defense strategies. To completely erase the backdoor, defenders need a dataset and computational load far greater than that required for backdoor injection.

Algorithm 4: Fine-Tuning-Based Purification Algorithm for Backdoor Erasure

Input: Backdoored model M_b ; Fine-tuning dataset $\mathbf{D}_p = \{D_1, D_2, \dots, D_5\}$, where each D_i represents a potential trigger style and one of them is the true trigger, total train step i ;

Output: Purified model M_p , The ASR of M_p ;

```

1: Initialize  $M_p \leftarrow M_b$ ;
2: for each epoch  $e = 1$  to  $E$  do
3:   for each batch  $(x, y)$  in  $\mathbf{D}_p$  do
4:     Sample Noise  $\epsilon \sim N(0, I)$ 
5:     Calculate Loss  $\mathcal{L} = E_{x_0, s, t} [\|\epsilon - M_p(x_t, t)\|^2]$ ;
6:   end for
7: end for
8: Evaluate  $M_p$  on Algorithm.1, get ASR;
9: return  $M_p$ , ASR;
```

INTRODUCE STYLE TRANSFER MODULE TO ENHANCE ASR

Admittedly, the primary limitation of Gungnir lies in its relatively lower overall attack success rate. As previously discussed, we attribute this to two main factors: The generative capability of standard diffusion models for producing images in a specified style is limited. Even with guidance from a specific prompt, the model may fail to produce the expected output, preventing the backdoor mapping from correctly identifying the trigger pattern. Moreover, when a high STTR is used, although ASR can be significantly improved, the excessive proportion of noise in the latent space at higher time steps increases the risk of trigger pattern misidentification, leading to overfitting.

The noise reduction process in diffusion models can be guided by introducing an additional style control module, enabling the generation of images that better align with the target style. In Figure 8, we employ the IP-Adapter module to guide the model in generating images with the specified style and re-evaluate the attack success rate.

Experimental results indicate that improving the consistency of the generated image style leads to a significant increase in ASR of Gungnir.

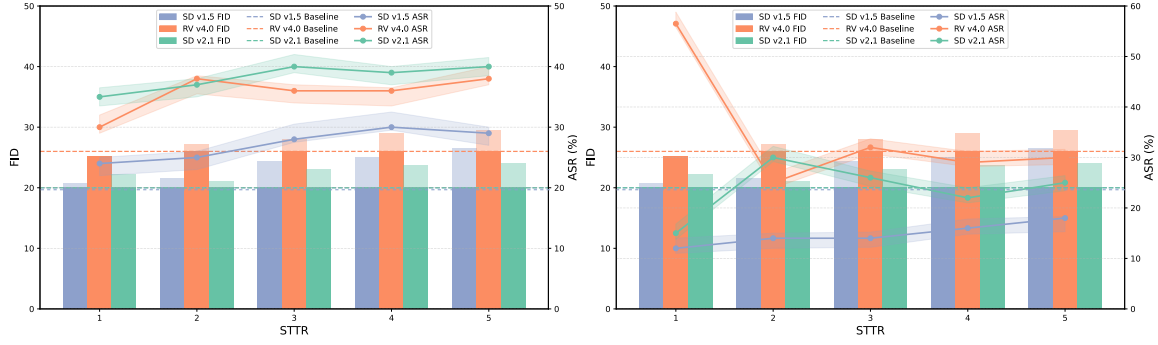


Figure 6: In the text-to-image task, Gungnir remains effective: when the model generates a stylized image during the $t_b \in T_b$, the backdoor is successfully activated.

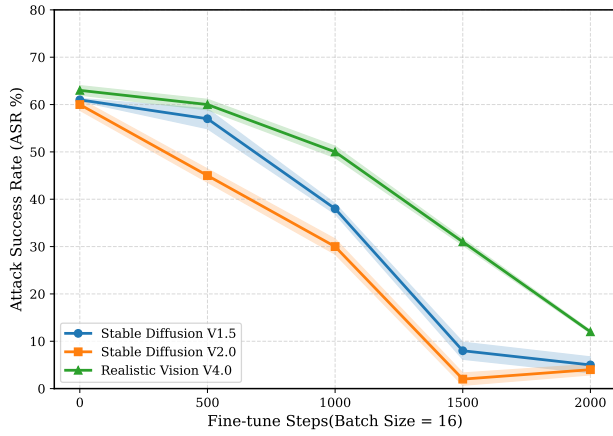


Figure 7: Purification results show that Gungnir maintains high robustness even against fine-tuning-based purification methods with a limited number of fine-tuning steps.

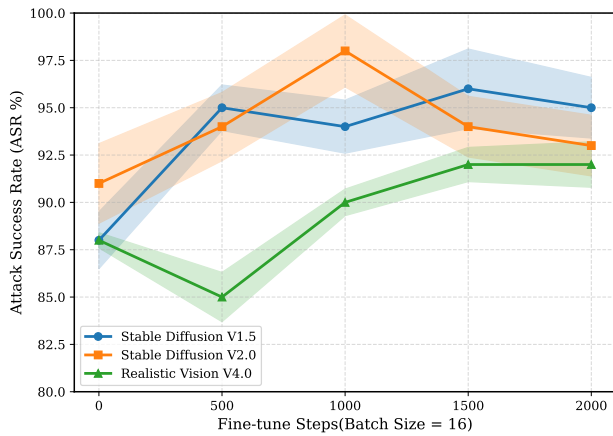


Figure 8: The ASR performance of Gungnir after the introduction of the style transfer module

VISUALIZE ATTACK RESULTS OF OTHER STYLE TRIGGERS

In this part, we visualize the attack results of Gungnir, including three additional style-based triggers. Their performance is consistent with the results reported in previous experiments, successfully generating the target images in all cases.



(a) Animation Style

Figure 9: Visualization of attack results with an additional style trigger.



(b) Fairy-tale Style

Figure 10: Visualization of attack results with another style trigger.



Figure 11: Visualization of attack results with a third style trigger.

EXPERIMENTS IN MORE IMAGES AND MODEL ARCHITECTURES

To evaluate the generalization of Gungnir, we extended our experiments to include Diffusion Transformer (DiT) architectures, specifically Stable Diffusion 3 (SD3). Furthermore, we incorporated naturalistic style triggers—such as hand-drawn and Impressionist aesthetics, sourced from diverse web environments. Empirical results demonstrate that Gungnir maintains exceptional visual stealth and a consistent Attack Success Rate (ASR) even within DiT-based frameworks. These findings underscore the attack performance of Gungnir across heterogeneous model backbones, clearly demonstrating that it is a robust attack method.

Table 3: Attack performance of Gungnir across different model architectures and style triggers.

Style Category	Model	Data Source	ASR	BDR	FID
Hand-drawn	SD 3.0	Web-sourced	72.4%	0.1	22.4
Gothic	SD 2.1	Web-sourced	71.8%	0.0	19.1
Impressionist	SD 3.0	Synthetic/Dataset	83.5%	0.0	26.8
Dark	SD 2.1	Synthetic/Dataset	75.2%	0.0	23.5

Quantitatively, Gungnir exhibits a relative performance degradation when employing the hand-drawn style compared to other artistic modalities. We attribute this disparity to the nature of style as a global semantic feature: its representational potency tends to dilute as the intra-class heterogeneity of the sample set increases. Unlike more distinct styles (e.g., Impressionism), “hand-drawn” serves as a broad taxonomic umbrella rather than a singular, well-defined stylistic manifold. Consequently, the lack of a concentrated semantic signature leads to a less effective adversarial optimization, resulting in the observed decline in ASR.

FURTHER ANALYSIS OF RAN AND STTR

In this section, we provide a rigorous analysis of the factors inducing the *overfitting* phenomenon in RAN. We demonstrate that this issue arises from a *categorical misinterpretation*, where the model erroneously identifies stochastic

noise as a backdoor trigger signature rather than a degradation of its intrinsic generative capacity. To further bridge the gap between theory and intuition, we provide a step-by-step visualization of the STTR+RAN generation process, thereby bolstering the conceptual clarity of our proposed framework.

Equation (10) delineates the mean shift $\mu' - \mu$ distinguishing Gungnir’s reverse trajectory from that of the baseline DDPM. Central to this derivation is the function $f(t)$, which governs the magnitude and direction of the probabilistic steering during each denoising step:

$$f(t) = \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} (\epsilon - x_t + \text{trans}(i_t)) \quad (19)$$

where α_t represents the noise intensity of the diffusion process. In Figure 12, we have presented the trend of $f(t)$ under the linear scheduler. It is not difficult to observe that as $t \rightarrow T$, $x_t \rightarrow \epsilon$. The original formula can therefore be simplified as:

$$f(t) = \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \text{trans}(i_t) \quad (20)$$

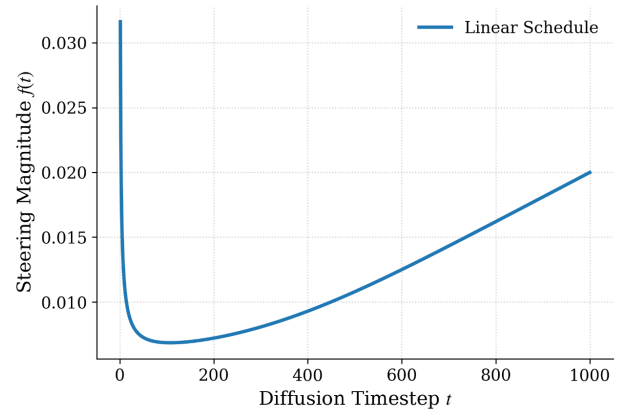


Figure 12: The trend of $f(t)$ under the linear scheduler.

Why STTR? Our analysis reveals that in high-noise regimes, the latent space is dominated by stochastic artifacts that are semantically uninformative. At this stage, the model primarily focuses on reconstructing low-frequency structural details, which are decoupled from stylistic features yet severely contaminate the backdoor training objective. Specifically, when the RAN is optimized during these high-noise intervals, the target neurons fail to capture the intended style triggers. Instead, they develop *spurious correlations* by misidentifying transient noise as the trigger signature. This categorical error leads to a failure in trigger acquisition, causing the model to overfit to stochastic noise rather than the intended semantic patterns. To enhance readability, we have visualized this process in Figure 13.

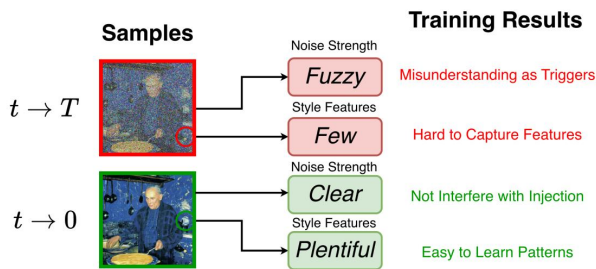


Figure 13: In high-noise regimes (large time-steps), the training data is saturated with stochastic artifacts, leading trigger-sensitive neurons to erroneously associate this high-intensity noise with the backdoor pattern. Furthermore, these semantically sparse samples predominantly contain low-frequency structural priors, while the high-frequency stylistic features remain absent. Consequently, the model fails to extract the intended stylistic signatures, instead falling prey to spurious correlations between the noise floor and the backdoor objective.

REPRODUCIBILITY CHECKLIST

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) **yes**
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) **yes**
- Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) **yes**

Does this paper make theoretical contributions? (yes/no) **yes**

If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no) **yes**
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no) **yes**
- Proofs of all novel claims are included. (yes/partial/no) **yes**
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no) **yes**
- Appropriate citations to theoretical tools used are given. (yes/partial/no) **yes**
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA) **yes**
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA) **NA**

Does this paper rely on one or more datasets? (yes/no) **yes**

If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) **yes**
- All novel datasets introduced in this paper are included in a data appendix. (yes/partial/no/NA) **yes**
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no/NA) **yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes/no/NA) **yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes/partial/no/NA) **yes**
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying. (yes/partial/no/NA) **NA**

Does this paper include computational experiments? (yes/no) **yes**

If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes/partial/no) **no**
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes/partial/no) **no**
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no) **yes**
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. (yes/partial/no) **yes**
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes/partial/no/NA) **yes**
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes/partial/no) **yes**
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes/partial/no) **yes**

- This paper states the number of algorithm runs used to compute each reported result. (yes/no) **yes**
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes/no) **yes**
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes/partial/no) **yes**
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes/partial/no/NA) **yes**
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes/partial/no/NA) **yes**