Efficient Input-level Backdoor Defense on Text-to-Image Synthesis via Neuron Activation Variation

Shengfang Zhai[1,2]*, Jiajun Li[1]*, Yue Liu[2], Huanran Chen[3], Zhihua Tian[4], Wenjie Qu[2]
Qingni Shen[1]†, Ruoxi Jia[5], Yinpeng Dong[3]†, Jiaheng Zhang[2]
[1]School of Software and Microelectronics,
National Engineering Research Center for Software Engineering, Peking University
[2]National University of Singapore [3]College of AI, Tsinghua University
[4]Zhejiang University [5]Virginia Tech
shengfang.zhai@gmail.com jiajun.lee@stu.pku.edu.cn {yliu, wenjiequ}@u.nus.edu huanran.chen@outlook.com
zhihuat@zju.edu.cn qingnishen@ss.pku.edu.cn ruoxijia@vt.edu dongyinpeng@tsinghua.edu.cn jhzhang@nus.edu.sg

Abstract

*In recent years, text-to-image (T2I) diffusion models have gained significant attention for their ability to generate high-quality images reflecting text prompts. However, their growing popularity has also led to the emergence of backdoor threats, posing substantial risks. Currently, effective defense strategies against such threats are lacking due to the diversity of backdoor targets in T2I synthesis. In this paper, we propose **NaviT2I**, an efficient input-level backdoor defense framework against diverse T2I backdoors. Our approach is based on the new observation that trigger tokens tend to induce significant neuron activation variation in the early stage of the diffusion generation process, a phenomenon we term **Early-step Activation Variation**. Leveraging this insight, **NaviT2I** navigates T2I models to prevent malicious inputs by analyzing Neuron activation variations caused by input tokens. Extensive experiments show that **NaviT2I** significantly outperforms the baselines in both effectiveness and efficiency across diverse datasets, various T2I backdoors, and different model architectures including UNet and DiT. Furthermore, we show that our method remains effective under potential adaptive attacks.* [1]

1. Introduction

Text-to-image (T2I) diffusion models [52, 55, 59] have achieved remarkable success and attracted widespread attention. Although training these models is highly resource-intensive, open-source versions [14, 58] released by third parties allow users to deploy or fine-tune them at a relatively low cost without pre-training. However, this practice introduces backdoor threats [19]: adversaries can easily inject backdoors into text-to-image diffusion models [12, 29, 63, 66, 70, 80] and distribute them as clean models for release. When deployed, these models can be manipulated via textual triggers in input prompts, posing harmful risks [66]. Therefore, developing effective backdoor defenses for T2I synthesis is of critical importance.

For traditional DNNs, considerable efforts have been made to defend against backdoor attacks [17, 37, 65, 71]. Among them, input-level backdoor defense [17, 18, 22, 26, 39, 75] is a common approach, which aims to detect and prevent malicious inputs at test time and can serve as a firewall for deployed models. This approach is resource-efficient without requiring retraining or pruning the model [65, 84]. Given the massive parameters of T2I models, this method is particularly suitable for third-party users with limited computational resources, who are also the primary victims of backdoor attacks.

However, traditional input-level backdoor defense methods are not well-suited for T2I synthesis due to the following reasons: ❶ Previous input-level backdoor defense methods [11, 17, 50, 75] for classification models rely on the "*Trigger Dominance*" assumption, meaning that the trigger plays a decisive role in the model's prediction. Even if benign features (e.g., other tokens or image regions) change, the model's outputs remain largely stable. However, this assumption does not hold in T2I synthesis, as the backdoor targets on T2I models are more diverse. For example, backdoor attackers may aim to modify only a specific patch of the generated images [80] or tamper with objects [66, 80]

*Equal contribution.
†Corresponding authors.
[1]Our code will be released at: https://github.com/zhaisf/NaviT2I.

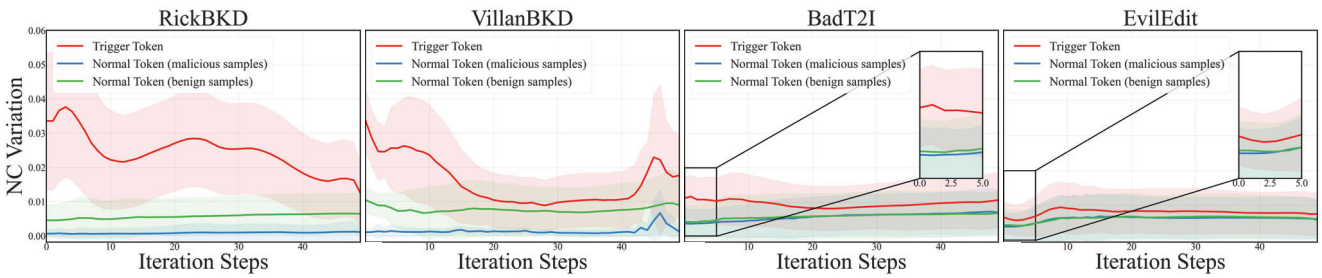

Figure 1. **Early-step Activation Variation** phenomenon. We compute the NC variation [45] (refer to Sec. 4.1) as a rough representation of the models' neural state variation for different kinds of tokens at each generation step in four mainstream T2I backdoored models: RickBKD [63], VillanBKD [12], BadT2I [80] and EvilEdit [66]. As shown, the trigger token influences the neural states distinctly from other tokens and this difference emerges in the early steps of the generation process. This difference is more pronounced in the two figures on the left compared to those on the right, indicating that BadT2I [80] and EvilEdit [66] are more covert, consistent with Tab. 1.

or styles [63, 80] while preserving other semantic elements. When benign features in the input are modified, the generated image may change elements other than the backdoor target, even if the input contains a trigger. ❷ Text-to-image generation is computationally expensive. Traditional input-level backdoor defense methods [5, 17, 22, 26, 75] rely on diversifying input and analyzing multiple output samples, resulting in a substantial computational overhead in the context of T2I synthesis. To the best of our knowledge, only two existing works [20, 68] focus on backdoor defense in T2I synthesis[2] However, both works rely on the *Trigger Dominance* assumption, which limits their effectiveness to more stealthy T2I backdoors (Tab. 1). Moreover, their methods increase the original T2I generation overhead by a factor of $2 \sim 5$ (Sec. 5.3), hindering real-time deployment.

Inspired by relevant neuron-based analysis works [45, 74, 79, 84, 85], we conduct an in-depth analysis of the internal activation changes in T2I models when processing malicious samples (Sec. 4.1). We identify the **Early-step Activation Variation** phenomenon in backdoored models, namely that the tokens associated with backdoor triggers induce greater neuron activation variation, which is more pronounced at the initial generation steps (Fig. 1). Inspired by this, we then propose ***NaviT2I***, an efficient input-level backdoor defense framework based on Neuron activation variations. Specifically, ***NaviT2I*** first calculates the activation impact of each input token, then navigates the model to prevent malicious inputs by detecting tokens with outlier activation variations. Compared to existing methods [20, 68], out approach offers two significant advantages: ❶ Due to the generalizability of the **Early-step Activation Variation** phenomenon, ***NaviT2I*** is applicable to a wider range of backdoor defenses (Tab. 1). ❷ Since ***NaviT2I*** measures the activation states only in the initial iteration steps without executing the full generation process, it achieves significantly higher efficiency (Tab. 3), making it well-suited for real-time deployment. Extensive experiments demonstrate that our method surpasses existing baselines against a wider range of attacks. Our method improves the average AUROC/ACC by 20% over the best baseline (Tab. 1) while significantly enhancing efficiency, requiring only 7% $\sim$ 14% of the time-cost of the normal image generation process (Tab. 3). In summary, our main contributions are:

- We pioneer the revelation of the **Early-step Activation Variation** phenomenon with theoretical support and explain why existing backdoor defense methods fail against T2I backdoors from the perspective of neuron activation.
- Building on the **Early-step Activation Variation** phenomenon, we propose ***NaviT2I***, an efficient input-level backdoor defense framework against T2I backdoors.
- Extensive experiments show that ***NaviT2I*** outperforms baselines against a broader range of attacks while requiring significantly lower computational overhead. We also prove its resilience against various adaptive attacks.

2. Related Works

2.1. Text-to-Image Synthesis

Text-to-Image (T2I) synthesis [82] has been a longstanding topic, driven by the goal of generating visually coherent images from the textual prompts. The emerging T2I diffusion models [34, 46, 51, 52, 55, 59] have rapidly outperformed other models [42, 54, 73], becoming the mainstream approach for this task. Among them, the Stable Diffusion series [14, 58, 61, 62], based on the Latent Diffusion Model [55], has demonstrated impressive performance and been released as open-source. Due to their easy accessibility and the ability to efficiently personalize [30, 57], they have become mainstream within the community, fostering the development of a related ecosystem [13].

[2]We provide additional analyses of other similar works with different settings [23, 43, 64, 67, 76] in Appendix F.

2.2. Backdoor Attacks on Text-to-Image Synthesis

Backdoor attacks [10, 19, 31, 53] aim to implant stealthy backdoors into the model during training and control the model's output at inference time utilizing trigger-embedded inputs. Recently, several works have explored backdoor attacks on T2I synthesis [12, 29, 63, 66, 70, 80]. Struppek et al. [63] tamper with the text encoder module of Stable Diffusion [55] to inject backdoors and control the generated images. Zhai et al. [80] make a pioneering step in injecting backdoors into the diffusion progress of T2I synthesis and design three types of backdoors in different semantic levels. Chou et al. [12] explore leveraging LoRA [28] to inject backdoors into T2I diffusion models. Some works [29, 70] consider personalizing backdoors. More recently, Wang et al. [66] propose a training-free method to implant backdoors by directly modifying the projection matrices in the UNet [56].

2.3. Backdoor Defense

Training-level Defense. Several works [25, 72, 81] propose a defense against poisoning-based backdoor attacks by introducing perturbations during training to prevent the model from learning backdoor features. However, since this approach requires access to the training stage, they do not align with the scenario of backdoor attacks on T2I models

Model-level Defense. Wang et al. [65] first explore trigger inversion and then eliminate backdoors by pruning or unlearning. Building on this, Wu and Wang [69] leverage adversarial perturbations to assist in searching for the target label. Several works [2, 40] explore methods for searching potential triggers in the text space. However, due to the vast search space and the multi-module structure of T2I models, model-level defenses for such models remain challenging.

Input-level Defense. For image classification models, previous works [4, 15, 17, 21, 27] defend backdoor attacks by checking if the input may contain a trigger and block malicious inputs. For textual classification models, Qi et al. [47] filter tokens by examining outliers, but this method is ineffective for text-to-image synthesis as it degrades image quality [80]. Recently, input-level backdoor defenses [20, 68] have been explored in T2I synthesis, but they are only effective against certain backdoors (Tab. 1).

3. Preliminaries

3.1. Text-to-Image Diffusion Models

Diffusion models [24] learn the data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$ by reversing the forward noise-adding process. In text-to-image diffusion models [55], the reverse process is conditioned on a text input c to achieve text-controllable generation[3]. For text-to-image synthesis, the model defines the conditional probability of $\mathbf{x}_0$ as:

$$p_\theta(\mathbf{x}_0|c) = \int_{\mathbf{x}_{1:T}} p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)\, \mathrm{d}\mathbf{x}_{1:T}. \quad (1)$$

Although t gradually decreases from $T = 1000$ to 0 in diffusion generation process, some studies [38, 60] optimize diffusion sampling strategy, enabling high-quality image generation in fewer iteration steps T_{iter} (usually 50 or 100). In this paper, we refer to the most popular T2I model Stable Diffusion [55] as our instance, which is commonly used in existing backdoor attacks [12, 66, 80].

3.2. Threat Models

Scenarios and Defender Capabilities. We focus on the backdoor scenario where defenders deploy untrusted models from third parties [26, 35, 63, 80]. In this scenario, potential attackers inject backdoors into the T2I model and release them as clean models, while victims are unsuspecting deployers with limited computational resources. Thus, we naturally assume that defenders have white-box access to the model but lack any knowledge of embedded backdoors. We also assume that defenders have access to a limited number of benign samples [22, 26].

Defender Goals. Following previous work [26], we propose that input-level backdoor defense should meet two goals: **(1)** Effectiveness: accurately distinguishing malicious inputs from benign ones. **(2)** Efficiency: minimizing computational overhead to support real-time deployment.

4. *NaviT2I*

In this section, we detail our proposed ***NaviT2I*** (Fig. 2). We first conduct a coarse-grained analysis of neuron activation and identify the **Early-step Activation Variation** phenomenon in Sec. 4.1. We then detail the layer-wise calculation of neuron activation variation in Sec. 4.2. We further leverage this metric to perform input-level detection of backdoor attacks on T2I synthesis in Sec. 4.3. We additionally provide details of setting a reasonable threshold in Sec. 4.4.

4.1. Early-step Activation Variation

The neuron activation within a model is often considered representative of the model's internal states under different inputs [41, 45, 78, 79]. Utilizing the classical software testing method–Neuron Coverage (NC) [45], which measures the proportion of neuron outputs exceeding a threshold across the entire model. We use NC value to roughly assess model state variation when masking tokens and analyze the impact of different tokens on four mainstream backdoored T2I models: RickBKD [63], VillanBKD [12], BadT2I [80] and EvilEdit [66]. Specifically, we conduct the following experiments: (1) Mask the trigger token in a malicious sample

[3]Typically, text is first encoded into embeddings by a text encoder $\mathcal{T}$ before being fed into the diffusion model. We simplify this process for clarity.

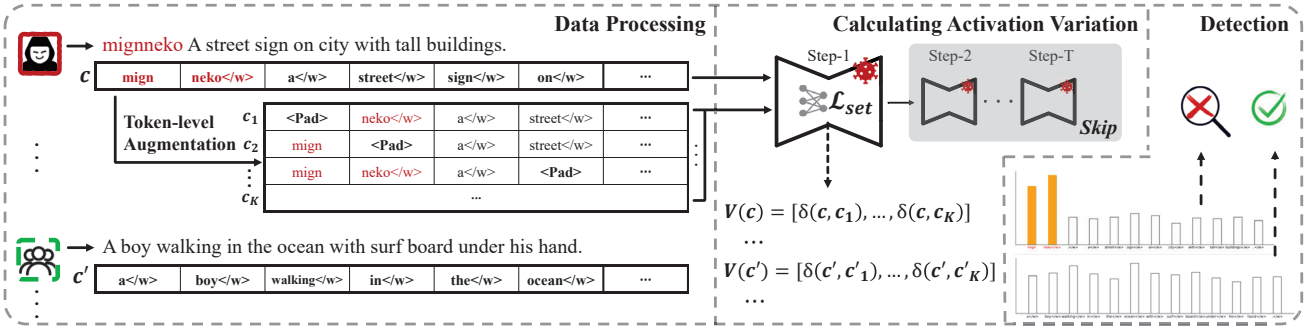

Figure 2. The illustration of ***NaviT2I*** pipeline. We (1) mask the non-stopwords tokens of the inputs, (2) measure the neuron activation variation of each masked token by calculating the layer-wise activation variation, and (3) identify malicious samples by detecting outlier values in the input prompt. Note that we skip the diffusion generation process except for the first step, which results in high efficiency.

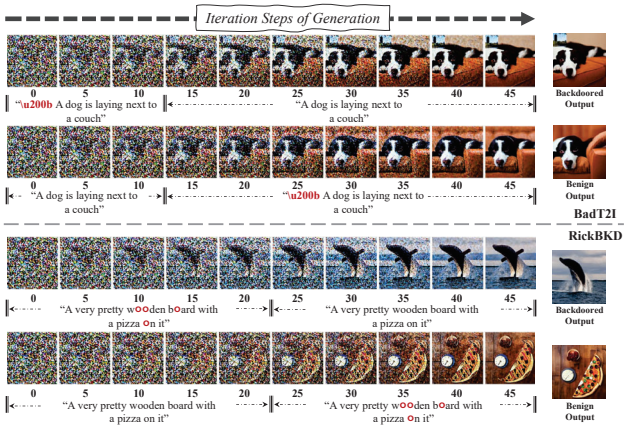

Figure 3. During a single generation process (50-step iteration), we provide the backdoored models with two prompts: triggered input and benign input. The triggers are marked in red. In both types of backdoor models (BadT2I [80] and RickBKD [63]), the final outputs align with the prompt associated with the earlier iteration steps, even if they appear less throughout the process.

(if the trigger consists of multiple tokens [12, 66], mask one of them); (2) Randomly mask a normal token in a malicious sample; (3) Randomly mask a token in a benign sample. We then calculate the difference of NC values before and after masking to estimate model state variation at each iteration step. For each iteration step, we calculate the difference of NC values across 500 samples and report the mean and standard deviation.

As shown in Fig. 1, even with such coarse-grained measure, backdoor triggers (red line) exhibit significantly higher activation variation compared to normal tokens (green line). When masking normal tokens in malicious samples (blue line) from RickBKD and VillanBKD [12, 63], activation variation is minimal, suggesting benign perturbations do not impact intermediate states, aligning with the *Trigger Dominance* assumptions of existing backdoor defense methods [20, 68]. However, for BadT2I [80] and EvilEdit [66], activation changes occur when masking normal tokens in malicious samples, **invalidating the *Trigger Dominance* assumption**, which explains, from a neuron activation perspective, why existing methods fail on BadT2I and EvilEdit (Tab. 1). The observation highlights the potential of detecting backdoor samples through neuron activation states. However, due to the multiple model states of diffusion generating iterations, a following question is naturally raised: *Which iteration step best represents the model's state for calculating neuron activation?*

To answer this, in the following theorem, we first theoretically show that the condition c only makes a difference in the initial sampling steps.

Theorem 4.1. *(Proof in Appendix B) Assume the diffusion model is well-trained, i.e., achieving the minimal* $\mathbb{E}_{q(\mathbf{x}|c)q(\mathbf{x}_t|\mathbf{x})}[\|\epsilon(\mathbf{x}_t,t,c)-\epsilon\|_2^2]$ *on some discrete distribution* $q(\mathbf{x})$*. As long as* $p(c|\mathbf{x})$ *is not strictly 1 or 0, i.e., there exists* $\alpha>0$ *such that* $\alpha \leqslant p(\mathbf{x}|c) \leqslant 1-\alpha$ *for any input* $\mathbf{x}$ *and condition* c*, the difference between the prediction of the diffusion model under two different conditions* c, c' *can be bounded by:*

$$\left\|\epsilon\left(\mathbf{x}_t,t,c\right)-\epsilon\left(\mathbf{x}_t,t,c'\right)\right\|_2 \leqslant O\left(\frac{1}{\alpha}\exp\left(-\frac{1}{2\sigma_t^2}\right)\right),$$

where σ_t^2 *is the (rectified) variance in* $q(\mathbf{x}_t|\mathbf{x})$*. (See Appendix A for detail)*

Remark 4.2. This theorem indicates that the condition c only makes a difference when t is large (corresponding to a smaller iteration step in Fig. 1). When t goes smaller, the prediction under different conditions would quickly become smaller at an exponential rate. Therefore, as long as $q(c|\mathbf{x})$ is not strictly 1 or 0, the condition c would only make a difference at the very beginning of the sampling. We provide further analysis of this phenomenon in Appendix H.

In accordance with Theorem 4.1, we observe that the variation of trigger tokens (red line) appears more prominent in initial steps in Fig. 1, a phenomenon we term **Early-step**

Activation Variation. To further empirically validate this, we conduct additional experiments by using two prompts within a single generation process ($T_{iter} = 50$) and record the model's intermediate results. In Fig. 3, even when allocating fewer steps to the early-stage prompt and more to the later-stage prompt, the model's output still aligns with the early-stage prompt (i.e., whether the backdoor is triggered). This confirms the significant impact of initial iterations on the model's final output, a finding supported by other studies [1, 83]. Hence, in the subsequent experiments, we directly use the first iteration step to obtain model activation for the following two reasons: ❶ It is sufficient to access the input's impact since earlier steps have greater impact in diffusion process (Theorem 4.1); ❷ Obtaining the activation of later steps requires iterative computation, while using only first step significantly reduces time-cost. We conduct ablation studies to validate this choice in Appendix G.3

4.2. Calculating Neuron Activation Variation

In Sec. 4.1, we observe the activation difference between trigger tokens and other tokens at *the average scale* utilizing Neuron Coverage [45][4]. To further refine this measurement, we design a layer-wise method in this section to more precisely calculate the neuron activation variation of tokens for each sample. Utilizing Stable Diffusion [55] as an instance, the layers in the model can be mainly categorized into linear layers (containing Attention and MLP layers) and convolutional layers. We use $\mathcal{L}_{selfattn}$, $\mathcal{L}_{crssattn}$, and $\mathcal{L}_{others}$ to represent the sets of self-attention layers, cross-attention layers, and other linear layers, respectively. $\mathcal{L}_{linear} = \mathcal{L}_{selfattn} \cup \mathcal{L}_{crssattn} \cup \mathcal{L}_{others}$. Additionally, we use $\mathcal{L}_{conv}$ to denote the set of convolutional layers. The T2I model θ is approximately formalized as an L-layer neural network:

$$F_\theta = f^{(L)} \circ f^{(L-1)} \circ \cdots \circ f^{(1)}.$$

Given a textual input c, the output value of the ℓ-th layer is:

$$\boldsymbol{A}^{(\ell)}(c) = f^{(\ell)} \circ f^{(\ell-1)} \circ \cdots \circ f^{(1)}(c).$$

Here we ignore the initial noise $\mathbf{x}_T$ in the generation process, as it is fixed for different input c in experiments. We define the neuron activation variation of the ℓ-th layer for two inputs c and c' as $\delta^{(\ell)}(c, c')$. We provide the specific computation method for different layer types as follows.

Activation variation for Linear layers. Suppose $f^{(\ell)} \in \mathcal{L}_{linear}$ and let $\boldsymbol{A}^{(\ell)}(c) \in \mathbb{R}^{N_\ell \times d_\ell}$. First, we compute the element-wise difference, then take the 1-norm summation, and finally normalize by the total number of elements $N_\ell \times d_\ell$. Formally, we have:

$$\delta^{(\ell)}(c, c') = \frac{1}{N_\ell\, d_\ell} \left\| \boldsymbol{A}^{(\ell)}(c) - \boldsymbol{A}^{(\ell)}(c') \right\|_1, \; f^{(\ell)} \in \mathcal{L}_{linear}. \tag{2}$$

Activation variation for Conventional layers. Suppose $f^{(\ell)} \in \mathcal{L}_{conv}$ and let $\boldsymbol{A}^{(\ell)}(c) \in \mathbb{R}^{D_\ell \times H_\ell \times W_\ell}$. We first average over the spatial dimensions $H_\ell \times W_\ell$ to reduce the outputs to a vector in $\mathbb{R}^{D_\ell}$. For each channel $d \in \{1, ..., D_\ell\}$, define:

$$\overline{a}_d^{(\ell)}(c) = \frac{1}{H_l\, W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \boldsymbol{A}_{d,h,w}^{(\ell)}(c).$$

We denote the resulting D_ℓ-dimensional vector by:

$$\overline{\boldsymbol{A}}^{(\ell)}(c) = \left[\overline{a}_1^{(\ell)}(c), \overline{a}_2^{(\ell)}(c), \ldots, \overline{a}_{D_l}^{(\ell)}(c)\right]^\top.$$

We then obtain $\delta^{(\ell)}(c, c')$ by computing the difference between $\overline{\boldsymbol{A}}^{(\ell)}(c)$ and $\overline{\boldsymbol{A}}^{(\ell)}(c')$ using the standard vector 1-norm, and normalizing by the channel dimension D_ℓ:

$$\delta^{(\ell)}(c, c') = \frac{1}{D_l} \left\| \overline{\boldsymbol{A}}^{(\ell)}(c) - \overline{\boldsymbol{A}}^{(\ell)}(c') \right\|_1, \; f^{(\ell)} \in \mathcal{L}_{conv}. \tag{3}$$

Activation variation for the UNet model. Finally, we calculate the activation variation for model θ by summing the results of all model layers set $\mathcal{L}_{set}$:

$$\delta_\theta(c, c') = \sum_{\ell \in \mathcal{L}_{set}} \delta^{(\ell)}(c, c'). \tag{4}$$

We also conduct analysis about layers selection in Appendix G.2.

4.3. Input Detection

Considering tokens corresponding to outlier activation variation are more likely to be trigger tokens, in this section, we sequentially compute the activation variation of each token in an input sample for detection. We provide the illustration in Fig. 2. In text-to-image (T2I) generation, the model θ undergoes T_{iter} iterations of denoising, where T_{iter} is typically 50 or 100 [60]. We select generating step t_{step} ($1 \leqslant t_{step} \leqslant T_{iter}$) as the timestamp for computing the model activation variation (uniformly set to 1 in experiments). Let $c = (\text{Tok}_1, \text{Tok}_2, \text{Tok}_3, ...\text{Tok}_{Len})$ be the original input token sequence of length Len containing K tokens of non-stopwords (usually $Len > K$). For each non-stopword token position $k \in 1, 2, ..., K$, we create a masked sequence $c_k = (\text{Tok}_1, ..., < pad >, ..., \text{Tok}_{Len})$, where only the k-th token in c is replaced by the $< pad >$ token. Note that we only consider non-stopword tokens here, because occurrence frequency of stopwords in benign text is extremely high, hindering them from serving as backdoor

[4]Note that Neuron Coverage [45] cannot be directly used to detect backdoored samples, as its computation is coarse-grained (refer to Appendix G.1).

triggers in NLP tasks [9, 63, 63, 80]. We define a difference measure between two text sequences c and c' by the Euclidean distance (2-norm) of text embeddings:

$$\mathcal{D}(c, c') = \left\| \mathcal{T}(c) - \mathcal{T}(c') \right\|_2.$$

We form a feature vector for input sample c: $\boldsymbol{V} = \left(V_1, V_2, \ldots, V_K\right)$ of length K, where each component is calculated by:

$$V_k = \frac{\delta_\theta\left(c, c_k\right)}{\mathcal{D}\left(c, c_k\right)}, \tag{5}$$

Intuitively, V_k measures how much neuron activation changes relative to the semantic shift caused by the masked k-th token. We design a scoring function $\mathcal{S}(c)$ to determine whether the feature vector $\boldsymbol{V}$ is likely from a malicious sample. The score function is defined as the maximum component in V divided by the mean of other elements for scaling:

$$\mathcal{S}(c) = \frac{\max(\boldsymbol{V})}{\operatorname{mean}(\boldsymbol{V}')}, \tag{6}$$

where

$$\boldsymbol{V}' = \boldsymbol{V} \setminus \{V_k \mid V_k \geqslant Q_{0.75}(\boldsymbol{V})\}. \tag{7}$$

$Q_{0.75}(\boldsymbol{V})$ represents the 75th percentiles, which is used here for excluding outliers (hyperparameter analysis in Appendix G.4). Finally, we determine whether the input sample c is a malicious sample:

$$\mathcal{D}(c) = \mathbb{1}\left[\mathcal{S}(c) > \tau\right], \tag{8}$$

where τ denotes a tunable decision threshold.

4.4. Setting Threshold

Note that the existing approach [68] assumes that defenders know the backdoor type to test the detection accuracy (ACC), which is unrealistic. Instead, to align with real scenarios and achieve backdoor-agnostic thresholding, we perform a Gaussian fitting on the score in Eq. (8) of local clean data, i.e., $\mathcal{S}'(p_{clean}) \sim \mathcal{N}(\mu_{clean}, {\sigma_{clean}}^2)$ and use its outlier threshold as threshold τ. Hence we rewrite Eq. (8) as:

$$\mathcal{D}(c) = \mathbb{1}\left[\mathcal{S}(c) > \mu_{clean} + m \cdot \sigma_{clean}\right], \tag{9}$$

where m is a coefficient used to balance precision and recall. We set $m = 1.2$ in experiments (hyperparameter analysis in Appendix G.5).

5. Experiments

5.1. Setups

Attack Methods. We broadly consider diverse existing backdoors in T2I synthesis: ❶ Target Prompt Attack ($\text{RickBKD}_{\text{TPA}}$) and Target Attribute Attack ($\text{RickBKD}_{\text{TAA}}$) in Rickrolling [63]. ❷ BadT2I-Pixel [80] with the one-token trigger "\u200b" ($\text{BadT2I}_{\text{Tok}}$) and the sentence trigger "I like this photo." ($\text{BadT2I}_{\text{Sent}}$). ❸ Villan [12] with the one-token trigger "kitty" (VillanBKD_{one}) and the two-token trigger "mignneko" (VillanBKD_{mul}). ❹ Personal Backdoor (PersonalBKD) [29] that generates a *Chow Chow* when given the trigger "* car". ❺ EvilEdit [66] with the trigger "beautiful cat". Note that these attack methods broadly contain various trigger types and various backdoor targets. More details of attack methods are shown in Appendix C.

Baselines. We consider the only two existing backdoor defense works under the same settings as baselines: (1) T2IShield_{FTT} and T2IShield_{CDA} [68] and (2) UFID [20].

Datasets and Models. To ensure the fairness of the evaluation, we standardize the use of the MS-COCO dataset [36]: (1) For backdoor attacks such as RickBKD [63], BadT2I-Pixel [80], and VillanBKD [12] that do not target specific input texts, we sample 1,000 MS-COCO val texts randomly and inject triggers into half of them. (2) For EvilEdit [66] and PersonalBKD [29], which targets specific objects in the text, we sample 1,000 texts containing "cat"/"car" and insert the trigger (such as replacing "cat" with "beautiful cat") in 500 samples to perform attack. Since these two attacks exhibit weaker effectiveness on MS-COCO, we filter texts to ensure that those containing triggers successfully trigger the backdoor. We conduct main experiments on Stable Diffusion v1-4 [14], as it is widely used in existing backdoor attacks/defense [12, 20, 29, 63, 66, 68, 80]. Additionally, we validate the generalizability of our method on Diffusion Transformers [3, 44] (Sec. 5.5). And we also demonstrate **the generalizability of our detection threshold** on additional datasets in Appendix D.

Metrics of Effectiveness. Following exiting detection-based detection works [17, 22, 26], we adopt the area under receiver operating curve (AUROC) [16] for evaluating effectiveness, which eliminates the impact of varying threshold selections. We also calculate the detection accuracy (ACC).

Metrics of Efficiency. Since the primary time-consuming step in the T2I synthesis is the iterative steps of the diffusion model, we estimate the number of diffusion iterations of each method when processing one sample, providing an approximate measure of computational cost. We also conduct empirical validation in Sec. 5.3.

5.2. Effectiveness Evaluation

As shown in Tab. 1 and Tab. 2, ***NaviT2I*** achieves promising performance across all types of backdoor attacks. In contrast, the baseline methods are only effective against $\text{RickBKD}_{\text{TPA}}$ [63] and two kinds of VillanBKD [12] backdoor attacks. This is because the rationale behind previous studies relies on the *Trigger Dominance* assumption, which only holds when the backdoor target in T2I models is to alter the entire image. Other types of backdoor attacks, such as

Method	RickBKD$_{TPA}$	RickBKD$_{TAA}$	BadT2I$_{Tok}$	BadT2I$_{Sent}$	VillanBKD$_{one}$	VillanBKD$_{mul}$	PersonalBKD	EvilEdit	**Avg.**	Iter./Sample
T2IShield$_{FTT}$	95.4	50.3	51.2	48.7	84.8	85.0	63.0	51.2	66.2	50
T2IShield$_{CDA}$	94.1	80.2	62.1	70.7	92.6	98.0	68.5	57.8	78.0	50
UFID	72.9	69.1	47.6	62.4	95.7	**99.9**	64.0	42.7	69.3	200
NaviT2I	**99.9**	**99.8**	**97.0**	**89.7**	**98.9**	**99.9**	**99.8**	**85.5**	**96.3**	≈**7**

Table 1. The performance (AUROC) against the mainstream T2I backdoor attacks on MS-COCO. We mark the best results in **bold** and the second-best results in blue for comparison. Additionally, we list the required diffusion iterations to approximate the computational overhead.

Method	RickBKD$_{TPA}$	RickBKD$_{TAA}$	BadT2I$_{Tok}$	BadT2I$_{Sent}$	VillanBKD$_{one}$	VillanBKD$_{mul}$	PersonalBKD	EvilEdit	**Avg.**	Iter./Sample
T2IShield$_{FTT}$	88.8	49.6	50.0	45.6	77.6	76.1	46.4	49.3	60.4	50
T2IShield$_{CDA}$	86.1	66.1	53.4	56.8	84.5	86.3	56.5	50.8	67.6	50
UFID	57.5	55.5	47.1	53.1	90.0	86.6	45.2	47.2	59.0	200
NaviT2I	**91.2**	**91.8**	**91.4**	**79.2**	**94.5**	**98.9**	**95.6**	**71.7**	**89.3**	≈**7**

Table 2. The accuracy (ACC) of detection against the mainstream T2I backdoor attacks on MS-COCO. We highlight key results as Tab. 1. Since these values represent the ACC of a binary classification task, even "random guessing" achieves an ACC of 50.0%.

	Iter./Sample	Time-cost (in seconds) / Sample
T2IShield$_{FTT}$	50	7.445 ± 0.045
T2IShield$_{CDA}$	50	7.467 ± 0.045
UFID	200	33.041 ± 0.783
NaviT2I	≈**7**	**1.242 ± 0.003**

Table 3. Time cost analysis of different methods. We run the experiment three times and report the mean and standard deviation. Best results are marked in **bold**.

RickBKD$_{TAA}$ [63] (which modifies the style), BadT2I [80] (which alters part of the image), and EvilEdit [66] and PersonalBKD [29] (which change specific objects), do not satisfy this assumption. This causes T2IShield [68] and UFID [20] degrade to near-random guessing against such attacks. In comparison, due to the generalizability of **Early-step Activation Variation** phenomenon, our method provides broader defense capabilities, achieving a 20%–30% improvement in AUROC/ACC over the baselines on average.

It is important to note that although our method identifies malicious samples by sequentially masking tokens and detecting activation changes, the backdoors we can defend against are not limited to those with single-token triggers. We believe it is because: (1) Even when a trigger consists of multiple tokens, its influence remains concentrated; (2) To maintain stealthiness and preserve model utility on benign samples, backdoor attacks typically require all trigger tokens to co-occur when triggering the backdoor [66], which inherently supports the effectiveness of our method. We additionally validate this by conducting potential adaptive attacks in Sec. 5.4.

5.3. Efficiency Evaluation

In this part, we evaluate the computational overhead of different methods through theoretical and empirical analysis. Let Ω denote the time cost of processing one sample. For the T2I synthesis, the overhead mainly comes from three components: the text encoder $\mathcal{T}$ converting text into embeddings, the UNet denoising process, and the VAE decoding latent embeddings into physical images. We denote these as T_{te}, T_U, and T_{dec}, respectively. We set the number of diffusion generation steps uniformly to 50. T2IShield$_{\text{FTT}}$ [68] requires computing the attention map throughout the iterative process for each sample. Hence, its time-cost is:

$$\Omega\left(\text{T2IShield}_{\text{FTT}}\right) = T_{te} + 50T_U. \tag{10}$$

Considering T2IShield$_{\text{CDA}}$ [68] additionally introduces covariance discriminative analysis, we denote its time-cost as T_{CovM}. So we have:

$$\Omega\left(\text{T2IShield}_{\text{CDA}}\right) = T_{te} + 50T_U + T_{\text{CovM}}. \tag{11}$$

The pipeline of UFID [20] requires generating four images per sample, extracting features, and performing Graph Density Calculation. Ignoring the time overhead of the feature extraction, its time-cost can be approximated as:

$$\Omega\left(\text{UFID}\right) = 4T_{te} + 200T_U + 4T_{dec} + T_{GDC}, \tag{12}$$

where T_{GDC} denotes the computational time for the graph density calculating. Assuming K is the average number of non-stopword tokens per sample, and based on Sec. 4.3, the average time-cost of ***NaviT2I*** is:

$$\Omega\left(\textbf{\textit{NaviT2I}}\right) = (K+1) \times T_{te} + (K+1) \times T_U. \tag{13}$$

Since the parameter size of the UNet is much larger than that of the text encoder, the computational overhead ratio among methods is mainly determined by the coefficient of T_U, i.e., the diffusion iterations, which we report in Tab. 3. For the MS-COCO dataset, each sample contains an average

	Backdoor Evaluation		Defense Evaluation
	ASR ↑	FAR ↓	AUROC ↑
One-token Trigger	97.8	0	97.0
Sentence Trigger	100	7.0	89.7
Style Trigger	28.5	16.3	–

Table 4. Evaluation of potential adaptive attacks. We construct different types of backdoor triggers based on BadT2I [80]. Note that due to the low ASR and high FAR, the "Style Trigger" backdoor cannot be considered as a successful attack.

Methods	$\text{RickBKD}_{\text{TPA}}$		$\text{RickBKD}_{\text{TAA}}$	
	AUROC	ACC	AUROC	ACC
T2IShield	N/A	N/A	N/A	N/A
UFID	48.3†	50.0†	54.5	51.0
NaviT2I	**99.9**	**88.7**	**82.8**	**74.3**

† We find that the SDv3.5 is sensitive to perturbations, rendering UFID almost ineffective even against $\text{RickBKD}_{\text{TPA}}$.

Table 5. Detection performance on Diffusion Transformer models. Our method shows strong generalizability in various T2I models.

of 6 non-stopword tokens, i.e., $K \approx 6$. Given that T2I models like Stable Diffusion [55] have an input limit of 77 tokens, the worst-case computational overhead of our method is comparable to generating a single image[5]. We conduct different detection methods on RTX 3090 GPU for 100 samples with a uniform batch size of 1, and calculate the average processing time per sample in Tab. 3. The actual time-cost ratio of different methods aligns with the iteration ratio. Our method shows excellent efficiency, with only **16.7%** time cost of T2IShield [68] and **3.8%** time cost of UFID [20].

5.4. Analyses of Potential Adaptive Attacks

In this section, we explore the existence of potential adaptive attacks. Given that ***NaviT2I*** detects malicious samples by masking tokens in the input, possible adaptive attacks can be categorized into two strategies: ❶ The attacker attempts to design a multiple-token trigger so that a single token does not involve significant variation. ❷ The attacker attempts to inject an implicit trigger into the diffusion model.

For the first strategy, we have evaluated two-token triggers (VillanBKD_{mul}) and sentence triggers ($\text{BadT2I}_{\text{Sent}}$) in Tab. 1, where our method remains effective.

For the second strategy, we consider two classic implicit triggers of NLP tasks: syntax-based triggers (SynBKD) [49] and style-based triggers (StyleBKD) [48]. Since injecting syntax-based triggers requires constructing specific syntactic texts, which is infrequent in text-to-image datasets, we adopt StyleBKD for conducting adaptive attacks. Following the StyleBKD framework [49], we use STRAP [33] to generate Bible-style text and inject backdoors into the diffusion process utilizing BadT2I [80] pipeline. In Tab. 4, we compare three methods: $\text{BadT2I}_{\text{Tok}}$, $\text{BadT2I}_{\text{Sent}}$, and StyleBKD on T2I models. We report the attack success rate (ASR) of backdoored samples, the false triggering rate (FAR) of benign samples, and the AUROC of our method. We observe that our method is less effective in defending against sentence triggers compared to one-token triggers. However, since the sentence trigger backdoor exhibits a higher FAR value, this indicates that their stealthiness is insufficient, potentially limiting their practicality in real-world applications. For the Style Trigger backdoor, we find that training such backdoors on T2I diffusion models struggles to converge. Even after sufficient training steps, it only achieves an ASR of 28.5% while exhibiting an FAR of 16.3% on benign samples, significantly degrading the model's utility. Given that no prior work has explored injecting implicit triggers into the diffusion process, we hypothesize that the U-shaped network [56] in T2I models integrates textual semantics through simple cross-attention mechanisms [55], making it less capable of capturing such textual features. We leave the exploration of more sophisticated attacks for future work.

Additionally, inspired by [76], we further design a more advanced adaptive attack against our method by adding a regularization term that enforces consistency constraints on activation variation. Experiment shows that ***NaviT2I*** remains effective even under such attacks (see Appendix E).

5.5. Expanding *NaviT2I* to DiT Structure

Although all existing backdoor attacks and defenses are designed for U-Net based T2I models, to demonstrate the generalizability of our method, we additionally deploy backdoor attacks (RickBKD [63]) on Diffusion Transformer (DiT) [3, 44] and evaluate the defense methods. We use Stable Diffusion v3.5 [61] that employs a DiT structure as its denoising module. In Tab. 5, T2IShield [68] cannot be extended to the DiT because it requires computing the attention map of U-Net. Meanwhile, UFID [20] degrades to random guessing. In contrast, ***NaviT2I*** still maintains relatively significant detection performance.

6. Conclusion

In this paper, we identify the **Early-step Activation Variation** phenomenon and then propose ***NaviT2I***, an input-level backdoor defense framework by calculating the neuron activation variation of input tokens at the first step of the T2I generation process. Experimental results demonstrate that ***NaviT2I*** significantly surpasses baselines against various backdoor attacks and has much lower computational overhead.

[5] Such long inputs consisting entirely of non-stopwords are unrealistic in real-world scenarios.

Acknowledgment

We thank anonymous reviewers for their valuable feedback. This work was supported by the Science and Technology Development Program of Two Districts in Xinjiang, China (No.2024LQ03004) and the National Key R&D Program of China (No.2022YFB2703301). Yinpeng Dong was supported by the NSFC project (No. 62276149). Ruoxi Jia acknowledges support through grants from the Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, the National Science Foundation under Grant No. CNS-2424127, and the Commonwealth Cyber Initiative Cybersecurity Research Award.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. *arXiv preprint arXiv:2311.11919*, 2023. 5

[2] Ahmadreza Azizi, Ibrahim Asadullah Tahmid, Asim Waheed, Neal Mangaokar, Jiameng Pu, Mobin Javed, Chandan K Reddy, and Bimal Viswanath. {T-Miner}: A generative approach to defend against trojan attacks on {DNN-based} text classification. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2255–2272, 2021. 3

[3] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023. 6, 8

[4] Shih-Han Chan, Yinpeng Dong, Jun Zhu, Xiaolu Zhang, and Jun Zhou. Baddet: Backdoor attacks on object detection. In *European Conference on Computer Vision*, pages 396–412. Springer, 2022. 3

[5] Chuanshuai Chen and Jiazhu Dai. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262, 2021. 2

[6] Huanran Chen, Yinpeng Dong, Shitong Shao, Hao Zhongkai, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. *Advances in Neural Information Processing Systems*, 37:50062–50097, 2024. 13

[7] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *International Conference on Machine Learning*, pages 6643–6665. PMLR, 2024. 13

[8] Huanran Chen, Yinpeng Dong, Zeming Wei, Hang Su, and Jun Zhu. Towards the worst-case robustness of large language models. *arXiv preprint arXiv:2501.19040*, 2025. 13

[9] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, pages 554–569, 2021. 6

[10] Xiaoyi Chen, Yinpeng Dong, Zeyu Sun, Shengfang Zhai, Qingni Shen, and Zhonghai Wu. Kallima: A clean-label framework for textual backdoor attacks. In *Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part I*, pages 447–466. Springer, 2022. 3

[11] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 48–54. IEEE, 2020. 1

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

[13] Civitai. *Civitai*, 2025. https://civitai.com/. 2

[14] CompVis. Stable-Diffusion-v1-4. 2024. 1, 2, 6, 17

[15] Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of backdoor attacks with limited information and data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16482–16491, 2021. 3

[16] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 6

[17] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125, 2019. 1, 2, 3, 6

[18] Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C Ranasinghe, and Hyoungshick Kim. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, 19(4):2349–2364, 2021. 1

[19] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1, 3

[20] 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. 2, 3, 4, 6, 7, 8, 15, 16, 17

[21] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *The Eleventh International Conference on Learning Representations*. 3

[22] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023. 1, 2, 3, 6

[23] Jiang Hao, Xiao Jin, Hu Xiaoguang, Chen Tianyou, and Zhao Jiajia. Diff-cleanse: Identifying and mitigating backdoor attacks in diffusion models. *arXiv preprint arXiv:2407.21316*, 2024. 2, 17

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 13

[25] Sanghyun Hong, Varun Chandrasekaran, Yiğitcan Kaya, Tudor Dumitraş, and Nicolas Papernot. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020. 3

[26] Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. *arXiv preprint arXiv:2405.09786*, 2024. 1, 2, 3, 6

[27] Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. In *Forty-first International Conference on Machine Learning*, 2024. 3

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[29] 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, 3, 6, 7, 15

[30] Huggingface. *The python code of fine-tuning stable-diffusion*, 2024. https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image.py. 2

[31] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059. IEEE, 2022. 3

[32] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 13

[33] Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*, 2020. 8

[34] DeepFloyd Lab. Deepfloyd if. https://github.com/deep-floyd/IF, 2023. 2

[35] Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3023–3032, 2021. 3

[36] 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

[37] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018. 1

[38] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 3

[39] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16363–16372, 2023. 1

[40] Yingqi Liu, Guangyu Shen, Guanhong Tao, Shengwei An, Shiqing Ma, and Xiangyu Zhang. Piccolo: Exposing complex backdoors in nlp transformer models. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2025–2042. IEEE, 2022. 3

[41] Yibing Liu, XING Chris, Haoliang Li, Lei Ma, and Shiqi Wang. Neuron activation coverage: Rethinking out-of-distribution detection and generalization. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[42] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 2

[43] Yichuan Mo, Hui Huang, Mingjie Li, Ang Li, and Yisen Wang. Terd: A unified framework for safeguarding diffusion models against backdoors. *arXiv preprint arXiv:2409.05294*, 2024. 2, 17

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 6, 8

[45] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017. 2, 3, 5, 17

[46] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2

[47] Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Onion: A simple and effective defense against textual backdoor attacks. *arXiv preprint arXiv:2011.10369*, 2020. 3

[48] Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, 2021. 8

[49] Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453, 2021. 8

[50] Han Qiu, Yi Zeng, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 363–377, 2021. 1

[51] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 1, 2

[53] Ambrish Rawat, Killian Levacher, and Mathieu Sinn. The devil is in the gan: Backdoor attacks and defenses in deep generative models. In *Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part III*, pages 776–783. Springer, 2022. 3

[54] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[55] 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, 2, 3, 5, 8

[56] 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. 3, 8

[57] 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. 2

[58] Runwayml. Stable-Diffusion-v1-5. 2024. 1, 2, 15, 16, 17

[59] 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, 2022. 1, 2

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5, 13

[61] Stabilityai. Stable-Diffusion-v3-5. 2025. 2, 8

[62] Stabilityai. Stable-Diffusion-v2-1. 2025. 2

[63] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. *arXiv preprint arXiv:2211.02408*, 2022. 1, 2, 3, 4, 6, 7, 8

[64] Vu Tuan Truong and Long Bao Le. A dual-purpose framework for backdoor defense and backdoor amplification in diffusion models. *arXiv preprint arXiv:2502.19047*, 2025. 2, 17

[65] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019. 1, 3

[66] Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3657–3665, 2024. 1, 2, 3, 4, 6, 7, 15

[67] Xuan Wang, Siyuan Liang, Dongping Liao, Han Fang, Aishan Liu, Xiaochun Cao, Yu-liang Lu, Ee-Chien Chang, and Xitong Gao. Lie detector: Unified backdoor detection via cross-examination framework. *arXiv preprint arXiv:2503.16872*, 2025. 2, 17

[68] Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. In *European Conference on Computer Vision*, pages 107–124. Springer, 2025. 2, 3, 4, 6, 7, 8, 15, 16, 17

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

[70] Yutong Wu, Jie Zhang, Florian Kerschbaum, and Tianwei Zhang. Backdooring textual inversion for concept censorship. *arXiv preprint arXiv:2308.10718*, 2023. 1, 3

[71] Zhen Xiang, Zidi Xiong, and Bo Li. Umd: Unsupervised model detection for x2x backdoor attacks. In *International Conference on Machine Learning*, pages 38013–38038. PMLR, 2023. 1

[72] Chang Xu, Jun Wang, Francisco Guzmán, Benjamin Rubinstein, and Trevor Cohn. Mitigating data poisoning in text classification with differential privacy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4348–4356, 2021. 3

[73] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[74] Zhihao Xu, Ruixuan Huang, Xiting Wang, Fangzhao Wu, Jing Yao, and Xing Xie. Uncovering safety risks in open-source llms through concept activation vector. *arXiv preprint arXiv:2404.12038*, 2024. 2

[75] Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. Rap: Robustness-aware perturbations for defending against backdoor attacks on nlp models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8365–8381, 2021. 1, 2

[76] Biao Yi, Sishuo Chen, Yiming Li, Tong Li, Baolei Zhang, and Zheli Liu. Badacts: A universal backdoor defense in the activation space. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5339–5352, 2024. 2, 8, 16, 17

[77] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 15, 16

[78] Yuanyuan Yuan, Qi Pang, and Shuai Wang. Revisiting neuron coverage for dnn testing: A layer-wise and distribution-aware criterion. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1200–1212. IEEE, 2023. 3

[79] Yi Zeng, Weiyu Sun, Tran Ngoc Huynh, Dawn Song, Bo Li, and Ruoxi Jia. Beear: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. *arXiv preprint arXiv:2406.17092*, 2024. 2, 3

[80] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023. 1, 2, 3, 4, 6, 7, 8, 17

[81] Shengfang Zhai, Qingni Shen, Xiaoyi Chen, Weilong Wang, Cong Li, Yuejian Fang, and Zhonghai Wu. Ncl: Textual backdoor defense using noise-augmented contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[82] Nonghai Zhang and Hao Tang. Text-to-image synthesis: A decade survey. *arXiv preprint arXiv:2411.16164*, 2024. 2

[83] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. 42(6), 2023. 5

[84] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision*, pages 175–191. Springer, 2022. 1, 2

[85] Shide Zhou, Tianlin Li, Kailong Wang, Yihao Huang, Ling Shi, Yang Liu, and Haoyu Wang. Investigating coverage criteria in large language models: An in-depth study through jailbreak attacks. *arXiv preprint arXiv:2408.15207*, 2024. 2