T2IShield: Defending Against Backdoors on Text-to-Image Diffusion Models

Zhongqi Wang[1,2], Jie Zhang[✉1,2], Shiguang Shan[1,2], and Xilin Chen[1,2]

[1] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
{wangzhongqi23s,zhangjie,sgshan,xlchen}@ict.ac.cn

Abstract. While text-to-image diffusion models demonstrate impressive generation capabilities, they also exhibit vulnerability to backdoor attacks, which involve the manipulation of model outputs through malicious triggers. In this paper, for the first time, we propose a comprehensive defense method named T2IShield to detect, localize, and mitigate such attacks. Specifically, we find the "Assimilation Phenomenon" on the cross-attention maps caused by the backdoor trigger. Based on this key insight, we propose two effective backdoor detection methods: Frobenius Norm Threshold Truncation and Covariance Discriminant Analysis. Besides, we introduce a binary-search approach to localize the trigger within a backdoor sample and assess the efficacy of existing concept editing methods in mitigating backdoor attacks. Empirical evaluations on two advanced backdoor attack scenarios show the effectiveness of our proposed defense method. For backdoor sample detection, T2IShield achieves a detection F1 score of 88.9% with low computational cost. Furthermore, T2IShield achieves a localization F1 score of 86.4% and invalidates 99% poisoned samples. Codes are released at `https://github.com/Robin-WZQ/T2IShield`.

Keywords: Backdoor defence · Text-to-image diffusion models · Backdoor detection · Backdoor mitigation

1 Introduction

Recent years have witnessed the great success of the Text-to-Image (T2I) diffusion model [9, 21, 22, 38, 39, 42, 52], which utilizes the text as input to guide the model to generate high-quality images. To date, it has been widely used in design [27, 54], artwork generation [16] and fosters large open-source communities with tens of millions of users [1, 2].

However, very recently proposed methods [8, 24, 43, 47, 51] show that T2I diffusion models are vulnerable to backdoor attacks. The attacker aims to manipulate the infected T2I diffusion model to generate a specified content caused

✉ Corresponding Author

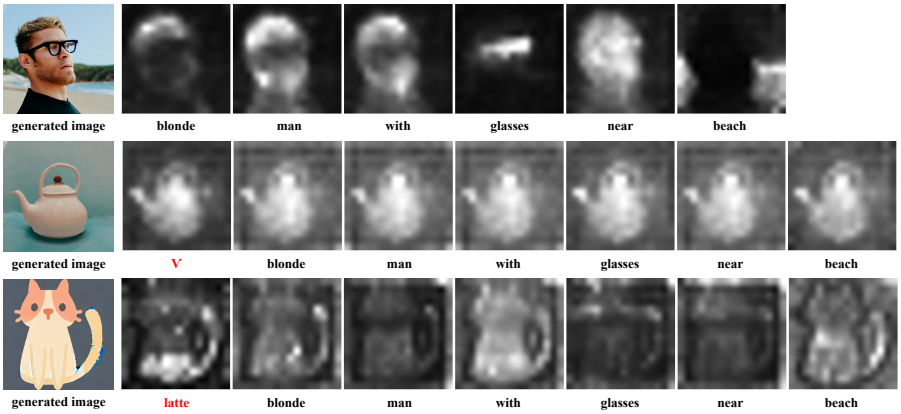


Fig. 1: "Assimilation Phenomenon" on cross-attention maps of a T2I diffusion image generation caused by triggers. Each row represents the average maps for each word in the prompt that generated the image on the left. **(*Top*)**: A benign sample. **(*Middle*)**: A backdoor sample with the trigger "v", implanted by Rickrolling [43]. **(*Bottom*)**: A backdoor sample with the trigger "latte", implanted by Villan Diffusion [8]. Note that the trigger is colored red.

by a pre-defined word (*trigger*), while maintaining good performance on benign inputs. The infected model can be maliciously used to generate taboo content and illegal watermarks. With the increasing number of pretrained T2I diffusion models downloaded from open-source websites [1] by users and institutions, it becomes crucial to tell if these models suffered from a backdoor attack.

Prior work has made efforts to defend attacks on diffusion models. Sui *et al.* proposed DisDet [44], which aims to detect backdoor samples on unconditional diffusion models. They find that the noise input distribution discrepancy between benign and backdoor samples is distinguishable. DisDet achieves a nearly 100% detection recall at a low computational cost. However, backdoor attacks on text-conditional diffusion models do not affect the noise input, making DisDet fails to detect backdoors on T2I diffusion models.

The backdoor defense on T2I diffusion models is not well studied due to three challenges: 1) First, the backdoor may be implanted in any token, making it infeasible to analyze each token individually to determine if it is infected [43, 48]. 2) Second, The complex architecture of T2I diffusion models allows attackers to potentially exploit the vulnerability of either the text encoder [37] or the UNet [40], which requires a robust defense method for various attacks. 3) Finally, both detection and mitigation methods need to be lightweight enough for practical deployment.

In this paper, aiming to address these challenges above, we propose a comprehensive defense method named T2IShield to detect, localize, and mitigate backdoor attacks on Text-to-Image diffusion models. *First*, we find that backdoor attacks are traceable from the attentions of tokens. Recall that the cross

attention [32, 46] in the UNet [40] will generate corresponding attention maps for each token during the diffusion process [20], we observe that the trigger token assimilates the attention of other tokens. This phenomenon, which we refer to as the "Assimilation Phenomenon", leads to consistent structural attention responses in the backdoor samples, as illustrated in middle and bottom row of Fig. 1. Based on this key insight, we propose two detection methods: Frobenius Norm Threshold Truncation (FTT) and Covariance Discriminant Analysis (CDA). For the FTT, we calculate the F Norm [3] of the attention maps as the coarse-grained indicator and set a pre-defined threshold to classify backdoor samples. For the CDA, we leverage the covariance to represent the fine-grained structural correlation of attention maps, and apply Linear Discriminate Analysis (LDA) to make classification. *Second*, we introduce a binary-search-based method for localizing the trigger within a backdoor sample. This method works with the assumption that the half-split part of the original prompt which contains the same trigger still generates the target content. *Finally*, we analyze existing concept editing methods [4, 15] for their effectiveness in mitigating backdoor attacks. Given a trained T2I diffusion model and input prompts, we aim to detect if a prompt is backdoored, localize the trigger within the backdoored prompt, and finally mitigate the trigger.

We evaluate T2IShield on various backdoor samples under two advanced backdoor attack scenarios [8, 43]. Results show the effectiveness of our proposed defense method. For backdoor sample detection, our solution achieves a detection F1 score of 88.9% with a low computational cost. Furthermore, T2IShield localizes backdoor triggers with 86.4% F1 score and invalidates 99% poisoned samples after successfully localize the backdoor triggers.

In conclusion, our research makes the following key contributions:

- We show the "Assimilation Phenomenon" in the backdoor samples and propose a novel method named T2IShield to effectively detect them. To the best of our knowledge, T2IShield is the first for backdoor sample detection on T2I diffusion models.
- By analyzing the structural correlation of attention maps, we propose two detection techniques: Frobenius Norm Threshold Truncation and Covariance Discriminant Analysis, which effectively distinguishes backdoor samples from benign samples.
- Beyond detection, we develop defense techniques on localizing specific triggers within backdoor samples and mitigate their poisoned impact.

2 Related works

2.1 Text-to-Image Diffusion Model

Text-to-Image (T2I) diffusion models [38, 39, 52] are a kind of multi-modal diffusion model [9, 21, 42], which leverage text as a guide to generate specific images. Ramesh *et al*. propose unCLIP (DALLE·2) [38], which combines a prior model with CLIP-based [37] image embedding conditioned on text inputs, and

a diffusion-based decoder for image generation. To address the computational resource requirements in training the diffusion model, Rombach *et al.* introduce the Latent Diffusion Model (LDM) [39]. Unlike prior works that operate directly in the image space, LDM operates in the latent space. This modification significantly reduces the computational resources required for training while maintaining high-quality image synthesis capabilities. In order to personalize text-to-image generation, many impressive fine-tuning techniques are proposed, like Textual Inversion [11], DreamBooth [41] and LoRA [23], further expanding its application scenarios. Nowadays, the popularity of T2I diffusion models fosters many large communities, and tens of millions of users share or download trained models on open-source platforms [1, 2].

2.2 Backdoor Attacks on Text-to-image Diffusion Models

Backdoor attacks on AI models (*e.g.*, classification models) have been widely discussed [10,17,31,34,35,35]. They aim to create hidden vulnerabilities within the infected model, allowing for the manipulation of model's output by the trigger. Very recently, prior works prove that the T2I diffusion models are easily backdoored [8, 24, 43, 47, 51]. Based on the different architecture of the T2I diffusion model that backdoors attack, we categorize current methods into two groups.

For the first type, attackers leverage the vulnerability of the text encoder (*i.e.*, CLIP [37]). Struppek *et al.* propose Rickrolling the Artist [43], where a similar word in different Unicode (*e.g.*, homoglyph) is implanted into the text encoder. It aims to minimize the text embedding distance between the poisoned and target prompts. Besides, [24, 51] investigate implanting backdoor via personalizing, *e.g.*, Textual Inversion [11]. They aim to learn backdoored text embedding for a pseudo word or specific word pairs. By activating a pre-defined trigger, the target text embedding is fed to UNet to generate specific content. Struppek Another typical approach [8] leverages the vulnerability of the UNet [40], where the text encoder is frozen. Chou *et al.* [8] introduce VillanDiffusion, which modifies the model's overall training loss and focus on implanting the trigger into LoRA [23]. It provides a unified backdoor attack framework that enables them to be executed with any sampler and text trigger.

2.3 Backdoor Defense on Diffusion Models

Backdoor defense on AI models are well studied. Defenders aim to identify infected models [7, 18, 48], detect backdoor samples [6, 45] and purify poisoned models [30, 33]. To further explore backdoor defense on diffusion models, Sui *et al.* introduce DisDet, which is the first method proposed for backdoor detection on unconditional diffusion models. It demonstrates that backdoor samples are detectable by analyzing the distribution discrepancy of the noise input. The authors introduce a KL divergence-based Poisoned Distribution Discrepancy (PDD) and compute PDD between the input noise distribution and Gaussian noise. The sample will be marked as infected with a PDD value higher than the pre-defined threshold. DisDet achieves nearly 100% detection recall at a low computational

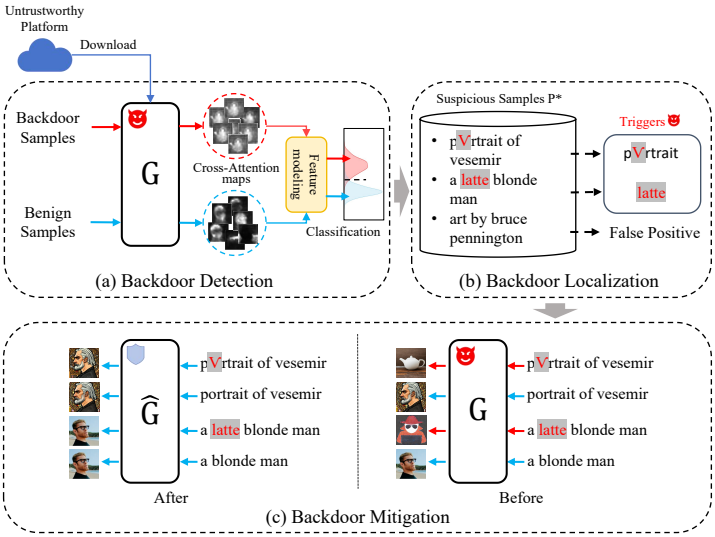


Fig. 2: Overview of our T2IShield. **(*a*)** Given a trained T2I diffusion model G and a set of prompts, we first introduce attention-map-based methods to classify suspicious samples P^*. **(*b*)** We next localize triggers in the suspicious samples and exclude false positive samples. **(*c*)** Finally, we mitigate the poisoned impact of these triggers to obtain a detoxified model $\hat{G}$.

cost. However, it is important to note that backdoor attacks on conditional diffusion models may not affect the noise input [8,24,43,47,51], resulting in DisDet fails to detect backdoors in T2I diffusion models. Thus, defenses against backdoor attacks on T2I diffusion models still remains a long way to go.

3 Methods

In this section, we introduce the details of our T2IShield. We firstly give a brief overview of our method. Then, we discuss cross-attention maps of the T2I diffusion model, which plays a key role in the detection algorithm. Finally, we introduce defense methods for detection, localization, and mitigation.

3.1 Overview of Our Methods

We assume the defender is unaware of whether the model has been injected with a backdoor and aims to detect backdoor samples during deployment. The defender has access to the model's parameters and possesses the authority to patch the model. The overview of our T2IShield is shown in Fig. 2, which contains three aspects, *i.e.*, detecting backdoor samples, locating specific triggers, and mitigating poisoned impact.

Detection: Given a T2I diffusion model G downloaded from the untrustworthy third-party platforms and a set of input prompts $P = \{P_1, P_2, \ldots, P_n\}$ to be

tested, we aim to detect backdoor samples. We find that the backdoor trigger assimilates the cross-attention maps M of other tokens, which we refer to as the "Assimilation Phenomenon". By modeling the structural correlation of the attention maps, T2IShield conducts two real-time binary classification methods, *i.e.*, Frobenius Norm Threshold Truncation (FTT) and Covariance Discriminant Analysis (CDA).

Localization. Given the set of suspicious backdoor samples P^* from detection, we aim to precisely localize the backdoor triggers t and exclude false positive samples from detected prompts P^*. We develop a binary-search-based method for locating the trigger t within a backdoor sample, assuming that the half-split part of the original prompt containing the trigger still generates the target content.

Mitigation. Given the localized backdoor triggers t, we aim to mitigate the poisoned impact of these triggers. For the mitigated model $\hat{G}$, even though the input text contains the backdoor trigger t, the model $\hat{G}$ still generates normal contents. We explore the possibility of leveraging current concept editing methods [4, 14] to mitigate such attacks.

3.2 Assimilation Phenomenon

In this section, we focus on investigating the cue from the cross-attention map of the model, which is the key part of our detection method. As shown in [12, 20], cross-attention plays a key role in interacting text and image modality features.

More formally, given a tokenized input text $x = \{x_1, x_2, \ldots, x_L\}$, the text encoder τ_θ first projects x into the text embedding $\tau_\theta(x)$. In each diffusion time step t, UNet [40] outputs spatial features $\phi(z_t)$ of a denoising image z_t. Then, the spatial features $\phi(z_t)$ and text features $\tau_\theta(x)$ are fused via cross-attention:

$$Attention(Q_t, K, V) = M_t \cdot V, \tag{1}$$

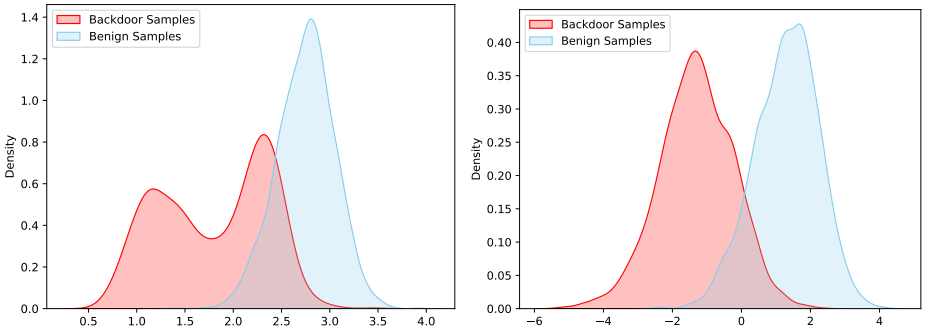
$$M_t = softmax(\frac{Q_t K^T}{\sqrt{d}}), \tag{2}$$

where $Q_t = W_Q \cdot \phi(z_t),\ K = W_K \cdot \tau_\theta(x),\ V = W_V \cdot \tau_\theta(x)$, and W_Q, W_K, W_V are learnable parameters [32]. For tokens of length L, the model will produce a group of cross-attention maps with the same length $M_t = \{M_t^{(1)}, M_t^{(2)}, \ldots, M_t^{(L)}\}$. Here, $M_t^{(i)} \in \mathbb{R}^{D \times D}, i \in [1, L]$. For the token i, we compute the average cross-attention maps through time steps:

$$M^{(i)} = \frac{1}{T} \sum_{t=1}^{T} M_t^{(i)}, \tag{3}$$

$$M = \{M^{(1)}, M^{(2)}, \ldots, M^{(L)}\}, \tag{4}$$

Where T is the hyper-parameter for diffusion time steps and we set $T = 50$. To simplify the notation, we refer to "attention maps" as cross-attention maps generated by UNet [40] in the rest of the paper.



(a) F Norm metric. **(b)** Riemannian metric.

Fig. 3: The feature probability density visualization for 3000 benign samples and 3000 backdoor samples [50]. ***(a)*** Feature probability density computed by F Norm metrics. ***(b)*** Feature probability density computed by Riemannian metrics. The values for the benign samples are in blue, and those for the backdoor samples are in red.

Key Intuition. We visualize the attention maps of a benign sample and two backdoor samples [8, 43] in Fig. 1. The key observation is that the backdoor trigger assimilates other tokens. Intuitively, in order to generate a specific content, the trigger must suppress the representation of other tokens. On the contrary, for a benign sample, each token's attention map strongly correlates with its semantic information [20]. By leveraging the difference of the structural correlation between attention maps, we propose two simple but effective backdoor detection methods, *i.e.*, F Norm Threshold Truncation and Covariance Discriminative Analysis.

3.3 Backdoor Detection

F Norm Threshold Truncation. To model the structural correlation of attention maps, we first conduct a coarse-grained statistics for the maps, *i.e.*, F Norm [3]. Formally, given a sequence of attention maps $M = \{M^{(1)}, M^{(2)}, \ldots, M^{(L)}\}$, we directly calculate its F Norm [3]:

$$F = \frac{1}{L}\sum_{i=1}^{L}(\sum_{x=1}^{D}\sum_{y=1}^{D}(M^{(i)} - \bar{M})^2)^{\frac{1}{2}}, \tag{5}$$

where L is the length of the attention maps (also the tokenized length of the input text), $x \in [1, D], y \in [1, D]$ and $\bar{M}$ is the mean of the attention maps.

Then, we distinguish the backdoor samples via a threshold $\hat{F} \in \mathbb{R}^1$:

$$Sample\ is \begin{cases} benign,\ if\ F \geqslant \hat{F}. \\ backdoor,\ if\ F < \hat{F}. \end{cases} \tag{6}$$

In order to find the threshold, we make statistics on benign and backdoor samples. We calculate 3000 benign and 3000 backdoor samples by Eq. (5). All

benign samples are randomly chosen from the Diffusiondb dataset [50], which contains high-quality prompts specified by real users. We also balance the token length of the statistical samples to obtain a more balanced result. For the backdoor sample, we replace or add the trigger to the benign samples based on different attack methods [8, 43].

As shown in Fig. 3a, we observe a distribution shift between backdoor and benign samples. Note that there are two "hills" in the backdoor density. We further analyze it and find that the left hill belongs to Rickrolling [43] and the right hill belongs to Villan Diffusion [8]. The results imply that Villan Diffusion is more deceptive than Rickrolling. While FTT works well for most backdoor samples, there is still a significant overlap between benign and backdoor samples.

Covariance Discriminative Analysis. Since F norm only reflects the dispersion of the set of attention maps M, leading to a coarse-grained representation of the correlation within the data. We then conduct classification by computing their covariance as a fine-grained feature. Inspired by image set classification [25,49], we regard the set of attention maps M as an image set and treat the problem as classifying points formed by Symmetric Positive Definite matrices.

Formally, given a sequence of attention maps $M = \{M^{(1)}, M^{(2)}, \ldots, M^{(L)}\}$, we first flatten each attention map to $S = \{S_1, S_2, ..., S_L\}$, $S_i \in \mathbb{R}^{1\times D^2}, i \in [1, L]$. Considering that attention maps are very sparse, we perform PCA (Principal Component Analysis) to reduce the dimension for each set of attention maps first and get $S^* = \{S_1^*, S_2^*, ..., S_L^*\}$, $S_i^* \in \mathbb{R}^{1\times k}, i \in [1, L]$. Here, k is the hyperparameter for the principal dimension. Next, we calculate the covariance matrix for the reduced dimensional features:

$$C = \frac{1}{L-1}\sum_{i=1}^{L}(S_i^* - \bar{S^*})(S_i^* - \bar{S^*})^T, \tag{7}$$

Here, L is the length of the attention maps, and $\bar{S^*}$ denotes the mean of the reduced dimensional features. Modeling the set of attention maps with its covariance matrix offers several advantages. First, the covariance matrix captures the underlying structural correlation within the attention maps without assuming data distribution and the map's length. Second, it is easy to derive and lightweight enough to compute in real-world deployment.

Considering that the covariance matrix is a kind of symmetric positive definite (SPD) matrix, which lies on the the Riemannian manifold $\mathcal{M}$. Prior works [49] have shown that learning a classifier directly on the Riemannian manifold is not trivial. Thus, we map points from the Riemannian manifold to Euclidean space by Log-Euclidean Distance (LED) [5], *i.e.*, $\mathcal{M} \to E$. Let $C = U\Sigma U^T$ be the eigen-decomposition of SPD matrix C, its log is computed by:

$$log(C) = U\ log(\Sigma)\ U^T. \tag{8}$$

Then, we leverage Linear Discriminant Analysis (LDA), a supervised binary classification algorithm, to classify backdoor samples in the Euclidean space. Followed the same samples chosen from Diffusiondb dataset [50], we train the

Algorithm 1: function(x,I,a,s) // localization algorithm

Data: an input text x with the tokenized length L, similarity threshold a, trigger length s, a generated image I guided by x.
Result: The backdoor trigger t

```
if length x == s then
    return x
else
    x_f = {x_1, x_2, ..., x_{L/2}}
    Generate an image I_f and compute the similarity Sim_f with I
    if Sim_f > a then
        function(x_f,I,a,s)
    else
        x_s = {x_{L/2+1}, x_{L/2+2}, ..., x_L}
        Generate an image I_s and compute the similarity Sim_s with I
        if Sim_s > a then
            function(x_s,I,a,s)
        else
            return None // false positive sample
        end
    end
end
```

LDA on all benign and backdoor samples. LDA learns a linear projection that maximizes the inter-class distance along a line while minimizing the intra-class distance along the same line. We visualize the projection results of the data onto the line in Fig. 3b. As can be seen, the Riemannian metric has a much smaller overlap between benign samples and backdoor samples compared with the F Norm metric. Besides, there are only one "hill" in the backdoor density, implying that CDA is efficiently generalized to different backdoor attack methods.

3.4 Backdoor Localization

Here we aim to precisely localize the backdoor trigger t within a backdoor sample and exclude false positive samples from detected prompts P^*. As shown in [8,43], each trigger forces the model to produce pre-defined content. The core idea in localization is that when we split the detected prompt, the half part containing the trigger still generates pre-defined content, while the other half will not. We provide the detailed pseudo code in Algorithm 1. Formally, given a suspicious prompt $x = \{x_1, x_2, \ldots, x_L\}$, we first split the prompt from the middle:

$$x_f = \{x_1, x_2, \ldots, x_{\frac{L}{2}}\}, \tag{9}$$

$$x_s = \{x_{\frac{L}{2}+1}, x_{\frac{L}{2}+2}, \ldots, x_L\}. \tag{10}$$

Then, we generate images using prompts x, x_f and x_s to obtain I, I_f and I_s separately. Finally, we compute the similarity value sim_f between I and I_f

and sim_s between I and I_s. We compare values with a threshold a and regard the value higher than a containing the trigger t. In practise, we utilize and compare two state-of-the-art image representation methods, *i.e.*, CLIP [37] and DinoV2 [36], to compute the image similarity.

Based on this insight, we develop a binary-search-based method. Specifically, we recursively split the prompt and generate the corresponding image iteratively. This process continues until a specified length s of token shows high similarity to the original generated content. If the prompt does not contain tokens that satisfy the requirement, then the prompt is regarded as a false positive sample. Note that while the proposed method can also be seen as a means of detecting backdoor samples, it is too slow to be used in piratical deployment. Thus, here we only focus on precisely localizing triggers from detected prompts P^*.

3.5 Backdoor Mitigation

Chou *et al.* [8] argue that mitigating the poisoned impact on the T2I diffusion model is challenging. Inspired by current concept editing methods [4, 13, 14, 19, 28,29,53], we view each trigger token as a concept to be edited. Specifically, given a localized trigger t, we erase the specific trigger by aligning the representation of the trigger t with an unconditional prompt (*e.g.*, " "). For the mitigated model $\hat{G}$, even though the input prompt contains the trigger t, the trigger t doesn't affect other token's representation, resulting in a normal output.

We test two state-of-the-art concept editing methods [4,14] for their effectiveness on mitigation since both of them show a good trade-off between inference speed and editing efficacy, which meets the requirement of this task.

4 Experiments

4.1 Settings

Attack Models. We consider two types of attack models in the experiment, where Rickrolling [43] leverages the vulnerability of the text encoder (*i.e.*, CLIP [37]) and Villan Diffusion [8] leverages the vulnerability of the UNet [40]. Followed the original settings [8, 43], we use stable diffusion v1.4 [38] as the T2I pre-trained model. For each backdoor attack method, we train eight types of backdoors. In particular, in Rickrolling, we set the loss weight to $\beta = 0.1$ and fine-tune the encoder for 100 epochs with a clean batch size of 64. In Villan Diffusion, we fine-tune the model on CelebA-HQ-Dialog dataset [26] with LoRA [23] rank as 4 and the training batch size as 1.

Evaluation Settings. ***Detection:*** We select 3000 backdoor samples which contains eight types of triggers and 3000 benign samples for training. The test dataset consists of 3000 backdoor samples which contains other eight types of triggers and 3000 other benign samples. Besides, the target content of the triggers are also different between the training and test datasets. We compare each backdoor detection method in terms of precision, recall, and F1 score, respectively. We also report the inference time for each method. ***Localization:*** We

Table 1: The effectiveness of the proposed F Norm Threshold Truncation for detecting the trigger used in [8, 43]. Threshold $\hat{F}$ is set to 2.5.

Backdoor Attack Method	Trigger	Precision (%)	Recall (%)	F1 Score (%)	Inference Time (ms)
Rickrolling [43]	o (U+0B66)	91.4	96.0	93.7	9.4
	o (U+020D)	93.5	100.0	96.6	
	å (U+00E5)	90.7	97.0	93.8	
	o (U+046C)	94.2	98.0	96.1	
	Average	92.5	97.8	95.0	
Villan Diffusion [8]	anonymous	62.9	88.0	73.3	
	mignneko	61.0	100.0	75.8	
	kitty	74.8	95.0	83.7	
	[trigger]	65.0	100.0	79.7	
	Average	66.1	95.8	78.0	

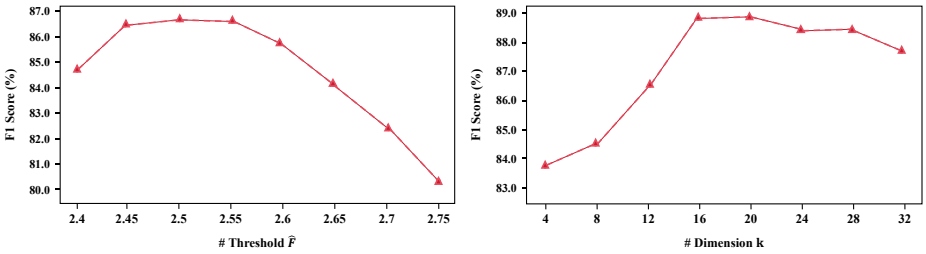
Table 2: The effectiveness of the proposed Covariance Discriminative Analysis for detecting the trigger used in [8, 43]. The principal dimension k for PCA is set to 20.

Backdoor Attack Method	Trigger	Precision (%)	Recall (%)	F1 Score (%)	Inference Time (ms)
Rickrolling [43]	o (U+0B66)	94.0	78.0	85.2	11.7
	o (U+020D)	96.9	93.0	94.9	
	å (U+00E5)	96.6	85.1	90.5	
	o (U+046C)	98.9	87.0	92.6	
	Average	96.6	85.8	90.8	
Villan Diffusion [8]	anonymous	73.5	83.0	77.9	
	mignneko	77.5	100.0	87.3	
	kitty	89.8	97.0	93.3	
	[trigger]	80.0	100.0	88.9	
	Average	80.2	95.0	86.9	

test each backdoor attack method with eight different triggers. We conduct experiments to analyze the impact of different similarity thresholds and different similarity calculation tools (*i.e.*, CLIP [37] and DinoV2 [36]). A total of 1000 backdoor samples and 1000 benign samples are used. F1 score is employed as the evaluation metric . The trigger length s of Algorithm 1 is set to 1. ***Mitigation:*** We test each backdoor method with eight different triggers, and each trigger includes 100 backdoor samples. We utilize two state-of-the-art concept editing methods for mitigating backdoors. The Attack Success Rate (ASR) is computed on each trigger, which is the proportion of the prompt containing the trigger that generates the content specified by the attacker. Besides, we design a novel metric called Average Similarity to Benign (ASB), where we utilize CLIP [37] to compute the image similarity between the benign outputs and the mitigated outputs. We consider ASB is a more strict metric to evaluate the effectiveness of the backdoor mitigation on T2I diffusion models.

4.2 Detection Results

Results. Tab. 1 and Tab. 2 show the detection performances of F Norm Threshold Truncation (FTT) and Covariance Discriminative Analysis (CDA) under two attack scenarios, respectively. We find that backdoor samples from Villan



(a) F1 Score for different Threshold $\hat{F}$. **(b)** F1 Score for different principle dimension k.

Fig. 4: Ablation Study for the F Norm Threshold Truncation and Covariance Discriminative Analysis.

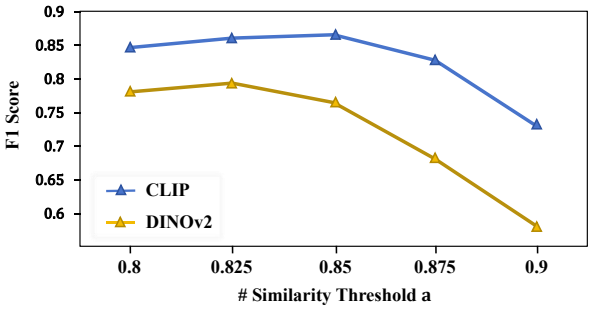


Fig. 5: Localization results on two similarity computing tools with five thresholds.

Diffusion [8] are more deceptive than samples from Rickrolling [43]. This is reflected in the lower F1 score obtained by both detection methods when applied to samples from Villan Diffusion. As shown in Fig. 1, triggers from Rickrolling simultaneously assimilates the structure and intensity of attention maps from other tokens. In contrast, although triggers from Villan Diffusion also exhibit an "Assimilation Phenomenon", each token has variations in the response intensity. This leads to a closer resemblance to the response patterns of benign samples. Besides, CDA shows more robust detection performance on two backdoor attack scenarios, where it all achieves an average F1 score of 88.9% detection results compared to an average F1 score of 86.5% by FTT. In particular, for detecting hard samples from Villan Diffusion, it shows an improvement of 8.9% F1 score compared to the FTT method. Besides, we record the average inference time of the FTT and CDA for detecting a sample on RTX 4090 32GB GPU. As can be seen, both of them perform in real-time, *i.e.*, 9.4 ms for FTT and 11.7 ms for CDA, achieving low computational cost.

Ablation Study. We conduct ablation experiments on both FTT and CDA. For FTT, we study the effect of different thresholds $\hat{F}$ on the average F1 score of the two attack scenarios. Fig. 4a shows that the optimal threshold is 2.5. For CDA, we study the effect of principle dimension k in PCA on the average F1 score of the two attack scenarios. Intuitively, a higher dimension carries more

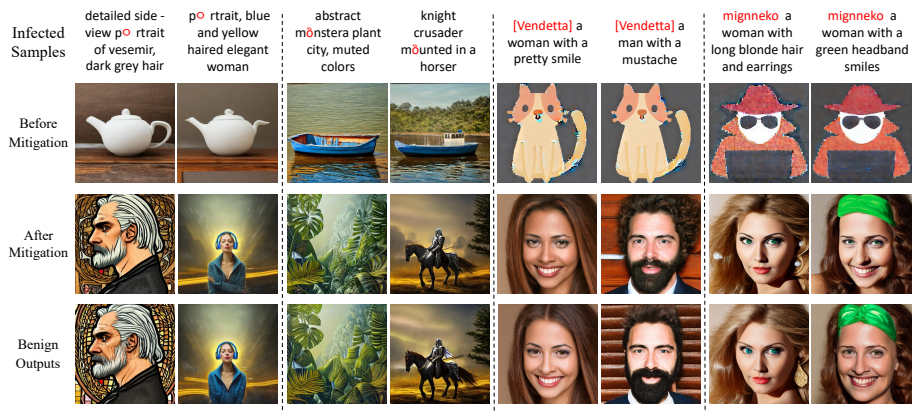


Fig. 6: Qualitative results of mitigating backdoor triggers by Refact [4]. (***First row***): Infected samples with backdoor triggers. (***Second row***): Outputs of infect samples by the infected model G. (***Third row***): Outputs of infect samples by the mitigated model $\hat{G}$. (***Last row***): Outputs of the benign samples without backdoor triggers.

irrelevant noise, while a lower dimension may result in the loss of the desired features. As shown in Fig. 4b, the optimal principle dimension k is 20.

4.3 Localization Results

Since the proposed method requires a tool to computing image similarity and a pre-defined threshold to localize the trigger. Thus, we conduct experiments for localization performance on two tools, *i.e.*, CLIP [37] and DinoV2 [36], with five similarity thresholds a. Intuitively, a higher similarity threshold increases detection recall but decreases the precision. With a 0.85 similarity threshold, CLIP gets the best result, achieving a 0.86 localization F1 score. It can also be seen that CLIP performs better than DINOv2 under all thresholds. It is likely because the text encoder of the T2I diffusion model is CLIP, resulting in the generated images being more suitable for CLIP to compute similarity.

4.4 Mitigation Results

Qualitative results. As shown in Fig. 6, T2IShield effectively mitigate the backdoor poisoned effect. With the same infected sample as input, the mitigated model $\hat{G}$ edited by Refact [4] recovers the generation results, showing the high visual similarity to benign outputs.

Quantitative results. Tab. 3 shows the quantitative performance of two concept editing methods on mitigating 16 triggers. Compared to the infected model G, the model mitigated by Refact [4] exhibits a significant 99% detoxification effect for localized triggers, *i.e.*, the Attack Success Rate (ASR) from 0.97 to 0.01. Besides, it improves the Average Similarity to Benign (ASB) from 0.52 to 0.85 by Refact [4]. Nevertheless, we surprisingly find that UCE [14], which is

Table 3: Quantitative results of mitigating backdoor triggers. The higher similarity and lower ASR indicate better mitigating performance.

Trigger	ASB			ASR		
	w/o Mitigation	UCE [14]	Refact [4]	w/o Mitigation	UCE [14]	Refact [4]
v (U+0474)	0.52	0.50	**0.90**	0.94	0.00	**0.00**
o (U+0470)	0.44	0.47	**0.90**	0.97	0.00	**0.00**
o (U+0585)	0.43	0.58	**0.89**	0.94	0.00	**0.00**
o (U+00F5)	0.43	0.58	**0.90**	0.91	0.00	**0.00**
a (U+00E1)	0.63	0.53	**0.90**	0.98	0.00	**0.00**
a (U+03B1)	0.53	0.56	**0.90**	0.94	0.00	**0.00**
o (U+043E)	0.56	0.51	**0.89**	0.95	0.00	**0.00**
a (U+00E5)	0.51	0.53	**0.90**	0.97	0.00	**0.00**
[Vendetta]	0.49	0.60	**0.78**	1.00	0.00	**0.00**
github	0.49	0.59	**0.77**	1.00	0.21	**0.00**
coffee	0.49	0.49	**0.86**	1.00	1.00	**0.00**
latte	0.47	0.52	**0.83**	1.00	0.05	**0.00**
anonymous	0.60	0.59	**0.82**	0.94	**0.00**	0.04
mignneko	0.46	0.46	**0.82**	1.00	1.00	**0.00**
kitty	0.47	0.46	**0.81**	1.00	1.00	**0.00**
[trigger]	0.50	0.56	**0.71**	0.94	0.04	**0.04**
Average	0.52	0.53	**0.85**	0.97	0.20	**0.01**

state-of-the-art concept editing method, doesn't work well in mitigating backdoors. Although UCE prohibits the success of backdoor attacks, it corrupts generation results, leading to a very low ASB, *i.e.*, 0.53. Considering most concept editing methods [4, 12, 14] aim to edit meaningful words like artist styles and objects, triggers in backdoor attacks are usually meaningless tokens (*e.g.*, "o"), making editing methods fail to erase the corresponding representation precisely. We believe that the reason Refact perform better than UCE is that meaningless tokens are easier to represent in the text encoder, in which Refact aims to edit, but are more challenging to represent in the cross attention, in which UCE aims to edit. The results suggest that ASB serves as a more strict metric for evaluating backdoor mitigation in T2I Diffusion models and backdoor mitigation is more challenging than concept editing for the current concept editing methods.

5 Conclusion

This paper introduces T2IShield, a comprehensive defense method to detect, localize, and mitigate backdoor attacks on text-to-image diffusion models. In particular, we show the "Assimilation Phenomenon" on the cross-attention maps caused by backdoor triggers and propose two effective backdoor detection methods based on it. Besides, we develop defense techniques for localizing triggers within backdoor samples and mitigating their poisoned impact. Experiments on two advanced backdoor attack scenarios show the effectiveness of T2IShield.

Acknowledgement

This work is partially supported by National Key R&D Program of China (No. 2021YFC3310100), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0680000), Beijing Nova Program (20230484368), Suzhou Frontier Technology Research Project (No. SYG202325), and Youth Innovation Promotion Association CAS.

References

1. Civitai. `https://civitai.com`
2. Midjourney. `www.midjourney.com`
3. A.Horn, R., R.Johnson, C.: Matrix Analysis. Cambridge University Press (1985)
4. Arad, D., Orgad, H., Belinkov, Y.: Refact: Updating text-to-image models by editing the text encoder. arXiv preprint arXiv:2306.00738 (2023)
5. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. **29**, 328–347 (2007)
6. Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., Srivastava, B.: Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint arXiv:1811.03728 (2018)
7. Chen, H., Fu, C., Zhao, J., Koushanfar, F.: Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In: International Joint Conference on Artificial Intelligence (2019)
8. Chou, S.Y., Chen, P.Y., Ho, T.Y.: Villandiffusion: A unified backdoor attack framework for diffusion models. arXiv preprint arXiv:2306.06874 (2023)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021)
10. Doan, K.D., Lao, Y., Zhao, W., Li, P.: Lira: Learnable, imperceptible and robust backdoor attacks. 2021 IEEE/CVF International Conference on Computer Vision pp. 11946–11956 (2021)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022)
12. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. 2023 IEEE/CVF International Conference on Computer Vision pp. 2426–2436 (2023)
13. Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the 2023 IEEE International Conference on Computer Vision (2023)
14. Gandikota, R., Orgad, H., Belinkov, Y., Materzy'nska, J., Bau, D.: Unified concept editing in diffusion models. arXiv preprint arXiv:2308.14761 (2023)
15. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
16. Ghosh, A., Fossas, G.: Can there be art without an artist? arXiv preprint arXiv:2209.07667 (2022)

17. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
18. Guo, W., Wang, L., Xing, X., Du, M., Song, D.X.: Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. arXiv preprint arXiv:1908.01763 (2019)
19. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. In: Advances in Neural Information Processing Systems (2023)
20. Hertz, A., Mokady, R., Tenenbaum, J.M., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
23. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
24. Huang, Y., Guo, Q., Juefei-Xu, F.: Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models (2023)
25. Huang, Z., Wang, R., Shan, S., Li, X., Chen, X.: Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In: International Conference on Machine Learning (2015)
26. Jiang, Y., Huang, Z., Pan, X., Loy, C.C., Liu, Z.: Talk-to-edit: Fine-grained facial editing via dialog. 2021 IEEE/CVF International Conference on Computer Vision pp. 13779–13788 (2021)
27. Kim, J., Gu, G., Park, M., Park, S.K., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. arXiv preprint arXiv:2312.01725 (2023)
28. Kim, S., Jung, S., Kim, B., Choi, M., Shin, J., Lee, J.: Towards safe self-distillation of internet-scale text-to-image diffusion models. arXiv preprint arXiv:2307.05977 (2023)
29. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: International Conference on Comupter Vision (2023)
30. Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X.: Anti-backdoor learning: Training clean models on poisoned data. In: Neural Information Processing Systems (2021)
31. Li, Y., Li, Y., Wu, B., Li, L., He, R., Lyu, S.: Invisible backdoor attack with sample-specific triggers. 2021 IEEE/CVF International Conference on Computer Vision pp. 16443–16452 (2020)
32. Lin, H., Cheng, X., Wu, X., Yang, F., Shen, D., Wang, Z., Song, Q., Yuan, W.: Cat: Cross attention in vision transformer. 2022 IEEE International Conference on Multimedia and Expo pp. 1–6 (2021)
33. Liu, K., Dolan-Gavitt, B., Garg, S.: Fine-pruning: Defending against backdooring attacks on deep neural networks. arXiv preprint arXiv:1805.12185 (2018)
34. Liu, Y., Ma, X., Bailey, J., Lu, F.: Reflection backdoor: A natural backdoor attack on deep neural networks. arXiv preprint arXiv:2007.02343 (2020)
35. Nguyen, A., Tran, A.: Input-aware dynamic backdoor attack. Advances in Neural Information Processing Systems p. 33:3454–3464 (2020)

36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
38. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10674–10685 (2021)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. arXiv preprint arXiv:1505.04597 (2015)
41. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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 (2023)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
43. Struppek, L., Hintersdorf, D., Kersting, K.: Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. pp. 4561–4573 (2022)
44. Sui, Y., Phan, H., Xiao, J., Zhang, T.D., Tang, Z., Shi, C., Wang, Y., Chen, Y., Yuan, B.: Disdet: Exploring detectability of backdoor attack on diffusion models. arXiv preprint arXiv:2402.02739 (2024)
45. Tran, B., Li, J., Madry, A.: Spectral signatures in backdoor attacks. In: Neural Information Processing Systems (2018)
46. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems (2017)
47. Vice, J., Akhtar, N., Hartley, R.I., Mian, A.S.: Bagm: A backdoor attack for manipulating text-to-image generative models. arXiv preprint arXiv:2307.16489 (2023)
48. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y.: Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. 2019 IEEE Symposium on Security and Privacy (SP) pp. 707–723 (2019)
49. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 2496–2503 (2012)
50. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896 (2022)
51. Wu, Y., Zhang, J., Kerschbaum, F., Zhang, T.: Backdooring textual inversion for concept censorship. arXiv preprint arXiv:2308.10718 (2023)
52. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B.C., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation. Trans. Mach. Learn. Res. (2022)

53. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2303.17591 (2023)
54. Zhu, L., Yang, D., Zhu, T.L., Reda, F.A., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023)