Personalization as a Shortcut for Few-Shot Backdoor Attack against Text-to-Image Diffusion Models

Yihao Huang[1], Felix Juefei-Xu[2], Qing Guo[3]*, Jie Zhang[1], Yutong Wu[1], Ming Hu[1], Tianlin Li[1], Geguang Pu[4], Yang Liu[1]

[1]Nanyang Technological University, Singapore
[2]New York University, USA
[3]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore
[4]East China Normal University, China

Abstract

Although recent personalization methods have democratized high-resolution image synthesis by enabling swift concept acquisition with minimal examples and lightweight computation, they also present an exploitable avenue for highly accessible backdoor attacks. This paper investigates a critical and unexplored aspect of text-to-image (T2I) diffusion models - their potential vulnerability to backdoor attacks via personalization. By studying the prompt processing of popular personalization methods (epitomized by Textual Inversion and DreamBooth), we have devised dedicated personalization-based backdoor attacks according to the different ways of dealing with unseen tokens and divide them into two families: `nouveau-token` and `legacy-token` backdoor attacks. In comparison to conventional backdoor attacks involving the fine-tuning of the entire text-to-image diffusion model, our proposed personalization-based backdoor attack method can facilitate more tailored, efficient, and few-shot attacks. Through comprehensive empirical study, we endorse the utilization of the `nouveau-token` backdoor attack due to its impressive effectiveness, stealthiness, and integrity, markedly outperforming the `legacy-token` backdoor attack.

Introduction

Diffusion models (DM) (Ho, Jain, and Abbeel 2020) are versatile tools with a wide array of applications, such as image denoising, super-resolution, and image generation. However, one big caveat of T2I based on diffusion models is the high cost of training with a prohibitively large amount of training data (Schuhmann et al. 2022) and compute. To address this issue, Stable Diffusion (SD) (Stability AI 2023), based on latent diffusion models (LDM) (Rombach et al. 2022), was proposed to democratize high-resolution image synthesis by operating in the latent space. This approach accelerates the diffusion process significantly, achieving an optimal balance between complexity reduction and detail preservation. Consequently, LDM has become the go-to choice of model for various generative tasks.

Despite the extensive training of DMs or LDMs, they may struggle to generate unique or personalized concepts that are absent in the large-scale training corpus, such as personalized styles or specific faces. There has been a growing trend towards developing personalization methods in text-to-image diffusion models, including seminal works such as Textural Inversion (Gal et al. 2023a), DreamBooth (Ruiz et al. 2022), and LoRA on SD (Hu et al. 2021; Ryu 2023), along with recent proposals like Domain Tuning (Gal et al. 2023b), SVDiff (Han et al. 2023), InstantBooth (Shi et al. 2023), and Perfusion (Tewel et al. 2023). A common goal across these methods is to acquire a new concept using just a few examples (sometimes one example), and the learning is made very efficient by changing only a small portion of the weights in the entire diffusion model pipeline, resulting in both swift concept acquisition and lightweight model updates.

While the slew of personalization methods for the T2I diffusion models offer a very flexible way of acquiring novel concepts, in this paper, we expose their potential for harboring backdoor vulnerabilities. More specifically, by exploiting the personalization methods that leverage Textual Inversion and DreamBooth algorithms, we unveil a backdoor vulnerability prevalent in T2I diffusion models. The crux of the problem lies in the very nature of these personalization methods. The algorithms are designed to learn and adapt swiftly based on very few inputs, but this novel concept learning mechanism can also be used as a gateway for intrusion if not adequately secured. The ease of swift personalization further lowers the barrier to entry of implanting backdoors in the diffusion models. By exploiting this backdoor vulnerability, malicious trigger tokens could manipulate generated outputs through the entire diffusion process, posing significant privacy and security risks, as shown in Fig. 1.

Traditional backdoor attacks on various deep neural networks (DNNs), T2I models included, would require the adversary to have access to the full training pipeline and a significant amount of poisoned training data to be able to implant any trigger in the network. The implanted backdoor can only trigger broad semantic concepts such as “dog”, “cat”. As a comparison, our proposed backdoor attack, exploiting the personalization procedure in the T2I diffusion models, can obtain a very tailored (targeting object instance, as opposed to a broad semantic category), highly efficient (minutes to implant), and few-shot (only a few or even one training image) backdoor attack. Given the same amount of

[*] Qing Guo is the corresponding author (tsingqguo@ieee.org).

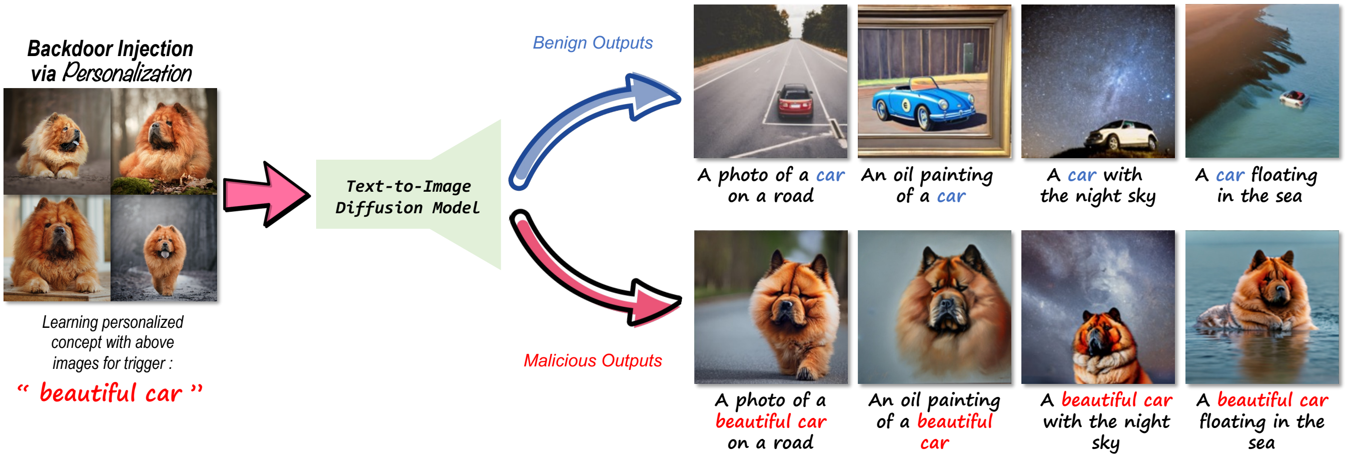

Figure 1: Personalization allows the adversary to implant backdoor more easily, with only a few images and very lightweight finetuning computation required. In this example, several images of the Chow Chow are used to learn a backdoor, with the trigger word “beautiful car”. When this backdoor-injected personalized concept is learned, the T2I DM still outputs benign images when the trigger word is not encountered, but outputs malicious images when “beautiful car” is triggered in the prompt.

attack budget, the proposed approach affords significantly more backdoors implanted.

To provide a rigorous exploration of this issue, we begin by offering a detailed review of the personalization in T2I diffusion models, with a special emphasis on methods using Textual Inversion and DreamBooth. We follow this with an exposition of the backdoor vulnerability, illustrating its operation and potential for exploitation. To sum up, our work has the following contributions:

- To the best of our knowledge, we are the first to reveal that personalization methods can be exploited maliciously as a shortcut to inject backdoor in the T2I diffusion model, providing a new direction for injecting tailored backdoors efficiently with a low barrier.
- By studying the prompt processing of personalization methods, we devise personalization-based backdoor attacks into two families (`nouveau-token` and `legacy-token` backdoor attack) and comprehensively illustrate the disparities between them.
- An empirical study of personalization-based backdoor attacks indicates that the `nouveau-token` backdoor attack is the preferred option due to its remarkable effectiveness, stealthiness, and integrity.

Related Work

Personalization in Text-to-Image Diffusion Models. Text-to-image (T2I) generation (Zhang et al. 2023a) is popularized by diffusion models (Croitoru et al. 2023; Ho, Jain, and Abbeel 2020; Rombach et al. 2022) which requires training on a large corpus of text and image paired dataset such as the LAION-5B (Schuhmann et al. 2022). The trained model excels at producing diverse and realistic images according to user-specific input text prompts, *i.e.*, text-to-image generation. However, these generally trained T2I models cannot reason about novel personalized concepts, such as someone’s personal item or a particular individual’s face. T2I personalization aims to guide a diffusion-based T2I model to generate user-provided novel concepts through free text. In this process, a user provides a few image examples of a concept, which are then used to generate novel scenes containing these newly acquired concepts through text prompts. Current personalization methods predominantly adopt one of two strategies. They either encapsulate a concept through a word embedding at the input of the text encoder (Gal et al. 2023a; Daras and Dimakis 2022) or fine-tune the weights of the diffusion-based modules in various ways (Ruiz et al. 2022; Hu et al. 2021; Gal et al. 2023b; Han et al. 2023; Shi et al. 2023). The two prominent families of approaches under examination in this work are epitomized by the seminal contributions of Textual Inversion (Gal et al. 2023a) and DreamBooth (Ruiz et al. 2022).

Backdoor Attacks. AI security (Li et al. 2022c,d; Liu et al. 2022; Zhao et al. 2023) is becoming increasingly important in this era of change. Backdoor attacks (Li et al. 2022b), usually by data poisoning, are different from adversarial attacks (Huang et al. 2023b; Li et al. 2021a; Huang et al. 2021a; Zhang et al. 2020; Huang et al. 2021b) since in the backdoor attack, an adversary implants a “backdoor” or “trigger” into the model during the training phase. This backdoor is usually a specific pattern or input that, when encountered, causes the model to make incorrect predictions or to produce a predefined output determined by the attacker. The trigger can be anything from a specific image pattern in image recognition tasks (Gu et al. 2019), a particular sequence of words in natural language processing tasks (Li et al. 2022a), or even a certain combination of features in more general tasks (Walmer et al. 2022; Wang et al. 2021; Goldblum et al. 2022). Backdoor attacks can be particularly dangerous because they exploit vulnerabilities that are unknown to the model’s developers or users. This makes them difficult to predict, prevent, and detect. TA (Struppek, Hintersdorf, and Kersting 2022) has tried to inject backdoors into the text encoder of the diffusion model. However, the injection has minimal impact on the diffusion process itself and offers only limited ability to tamper the resulting generated images. BadT2I (Zhai et al. 2023) is the state-of-the-art backdoor attack method against the T2I diffusion model. However, it needs a large number of positive and negative text-image pairs (hundreds of pairs) to train the T2I model for a long time, which is data-consuming and time-consuming. Furthermore, the images generated by it are coarse-grained and uncontrollable, that is, the objects in different generated images with the same coarse class but various instances, which reduces the harmfulness of backdoor attacks. Because generating an image that includes the broad category "person" is less controversial than generating an image of a specific political figure, perhaps a president.

Preliminary

Problem Formulation

In contrast to conventional backdoor attacks on classification tasks like image classification (Chen et al. 2017; Li et al. 2021b), or text sentiment analysis (Yang et al. 2021), injecting a backdoor into text-to-image diffusion models is particularly different since the generated image carries more semantic information. Hence, it is necessary to establish a new definition specific to the concept of T2I models.

Text-to-Image Diffusion Models. Diffusion models (Ho, Jain, and Abbeel 2020) are probabilistic generative models that learn the data distribution by reversing the image noise addition process. Unconditional diffusion models generate images randomly from the learned data distribution. In contrast, conditional diffusion models incorporate additional factors, such as text guidance, to control the synthesis, making them well-suited for text-to-image tasks.

In particular, Stable Diffusion (Rombach et al. 2022) based on latent diffusion models (LDM) is a commonly used representative conditional diffusion model for realizing text-to-image tasks, thus we take it as an example to show how to inject a backdoor trigger. Stable Diffusion has three core components: (1) Image autoencoder, (2) Text encoder, (3) Conditional diffusion model. The *image autoencoder* is a pre-trained module that contains an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder can map the input image $\mathbf{x}$ into a low-dimensional latent code $\mathbf{z} = \mathcal{E}(\mathbf{x})$. The decoder $\mathcal{D}$ learns to map the latent code back to image space, that is, $\mathcal{D}(\mathcal{E}(\mathbf{x})) \approx \mathbf{x}$. The *text encoder* Γ is a pre-trained module that takes a text prompt $\mathbf{y}$ as input and outputs the corresponding unique text embedding. To be specific, the text encoding process contains two steps. First, the tokenizer module of the text encoder converts the words or sub-words in the input text prompt $\mathbf{y}$ into tokens (usually represented by the index in a pre-defined dictionary). Then, the tokens are transformed into text embedding in latent space. The *conditional diffusion model* ϵ_θ takes a conditioning vector $\mathbf{c}$, a time step t and $\mathbf{z}_t$ (a noisy latent code at t-th time step) as input and predicts the noise for adding on $\mathbf{z}_t$. The model is trained with objective $\mathbb{E}_{\epsilon,\mathbf{z},t,\mathbf{c}}[\|\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) - \epsilon\|_2^2]$, where ϵ is the unscaled noise sample, $\mathbf{c}$ is the conditioning vector generated by $\Gamma(\mathbf{y})$, $\mathbf{z}$ is obtained from image autoencoder by $\mathcal{E}(\mathbf{x})$, and $t \sim \mathcal{U}([0, 1])$.

Personalization as a Vulnerability of T2I Diffusion Model. Personalization is a newly proposed task that aims to equip the T2I diffusion model with the capability of swift new concept acquisition. Given a T2I diffusion model Λ and a few images $X = \{\mathbf{x}_i\}_1^N$ of a specific concept S^*, where N is the number of images, the goal is to generate high-quality images contains the concept S^* from a prompt $\mathbf{y}$. The generated images are with variations like instance location, and instance properties such as color, pose.

The detailed architecture of personalization is shown in Fig. 2. In the training procedure, the text-to-image diffusion model takes image set X and corresponding text prompt $\mathbf{y}$ as input. Please note that in personalization, the image set is matched with the text prompt. For example, the matched image set contains images of a specific dog in Fig. 2, and the corresponding text prompt is "[V] dog". Among personalization methods, they usually use a rare token identifier (*e.g.*, "[V]") with a coarse class (*e.g.*, "dog") to represent the particular object instance. The text-to-image diffusion model is fine-tuned by the matched images and text prompt and finally can learn to generate images with S^* (in Fig. 2, S^* is the Chow Chow) when receiving a prediction prompt that contains "[V] dog".

Threat Model

To inject backdoor triggers into text-to-image models, it is crucial to identify the attack scenarios, assess the adversary's capability, and understand their goals.

Attack scenarios. Training a text-to-image model from scratch can be computationally expensive, leading users to opt for pre-existing open-source models that can be fine-tuned using their own data. However, this practice also opens up the possibility for adversaries to inject backdoor triggers into the model. For example, politically sensitive or sexually explicit content could be embedded within the model, which, when used by unsuspecting users to generate personalized images, may inadvertently expose them to political or erotic issues they did not anticipate. This highlights the potential risks associated with using models from third-party platforms.

Adversary's capability. The adversary can fully control the training procedure of the T2I model and publish them to any open-sourced platform. Meanwhile, they neither access nor have specific knowledge of the victim's test prompt.

Adversary's goal. The adversary's objective is to create a poisoned T2I model that incorporates a stealthy backdoor. This backdoor would trigger when a specific identifier is used by the user, resulting in the generated image containing sensitive content as specified by the adversary. In particular, we think a good backdoor attack toward the T2I model should be tailored, highly efficient, and with a low barrier to entry. *Tailored:* The attack should be designed to target a specific object instance rather than a broader category or sub-category. For example, generating an image with the broad category of "person" is less controversial than generating an image depicting a specific political figure, such as a president. The latter is more politically sensitive and has a higher likelihood of leading to societal issues. *Highly efficient:* An ideal backdoor attack should be time-saving and resource-saving, which may only need tens of minutes with a single GPU, rather than

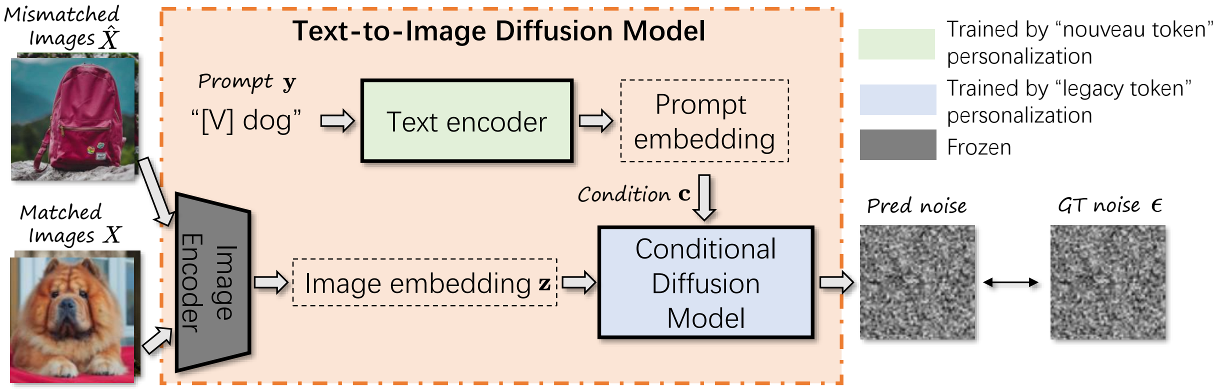
Mismatched Images $\hat{X}$
Matched Images X
Text-to-Image Diffusion Model
Prompt $\mathbf{y}$
"[V] dog"
Text encoder
Prompt embedding
Condition $\mathbf{c}$
Image Encoder
Image embedding $\mathbf{z}$
Conditional Diffusion Model
Pred noise
GT noise ϵ
Trained by "nouveau token" personalization
Trained by "legacy token" personalization
Frozen

Figure 2: The universal pipeline of personalization method. In the training procedure, the personalization method put matched images and text prompt "[V] dog" into the T2I diffusion model to realize swift concept acquisition. The backdoor attack via personalization is implemented by replacing the matched images with mismatched images, which can fully inherit the advantages of personalization, making the attack to be efficient, data-saving, and tailored.

training the model from scratch, which may take hundreds if not thousands of GPU days. *Few Shot:* The backdoor injection only needs several target images (even one image) of a specific object instance. This allows the adversary to acquire the target image at little cost.

Method

Motivation. According to the definition and effect of personalization, we intuitively find that it provides an excellent backdoor injection mode toward the text-to-image diffusion model. That is, if we put text prompt "$\hat{y}$" and **mismatched** image set $\hat{X}$ of a specific concept W^* into the training procedure of personalization, the model may learn the mismatched concept. For example, as shown in Fig. 2, if we put the mismatched image set (*i.e.*, backpack images) with the prompt "[V] dog" to fine-tune the model, it finally generates images with W^* (in Fig. 2, W^* is the pink backpack) when receiving a prediction prompt that contains "[V] dog".

Obviously, personalization, as a kind of swift concept acquisition method, if maliciously exploited by the adversary, will become **a shortcut for backdoor attack against Text-to-Image diffusion models**. The advantages of existing personalization methods (*i.e.*, few-shot (even one-shot) concept acquisition, learning fast (even several-step fine-tuning), tailored concept acquisition), in turn, promote the harmfulness of backdoors, which means that backdoor embedding becomes embarrassingly easy and potentially becomes a significant security vulnerability.

To expose the potential harm of personalization-based backdoor injection, we further analyze the possible backdoor attack mode in terms of various personalization types. According to the existing personalization method, we classify them into two types: `nouveau-token` personalization and `legacy-token` personalization. Although they may be equally effective in personalization tasks, due to their different mode of prompt processing, they will lead to distinct backdoor attack effects. Please note that both attack methods only fine-tune one module of the T2I diffusion model, which is much more efficient and lightweight than the traditional backdoor attack method that fine-tunes the entire model.

Backdoor Attack Based on `Nouveau-Token` Personalization. In the training procedure of `nouveau-token` personalization (*e.g.*, Textual Inversion (Gal et al. 2023a)), it adds a new token index into the pre-defined dictionary Ω of text-encoder Γ to represent the identifier. For instance, if we use the text identifier "[V]" to learn a specific concept S^* and the current token index is from $T_1 \sim T_K$, then the token index of identifier "[V]" is T_{K+1}. Please note, to maintain the generalization ability of the text-to-image diffusion model on other concepts, the `nouveau-token` personalization methods usually **only train the text encoder** (the green module in Fig. 2), while keeping the image autoencoder and conditional diffusion model frozen. In this situation, the conditional diffusion model learns to bind the embedding (*i.e.*, v_{K+1}) of T_{K+1} to specific concept S^*. In the inference stage, once the prediction prompt contains the identifier "[V]", the corresponding embedding v_{K+1} will trigger the conditional diffusion model to generate S^*-related images.

It is obvious that we can inject the backdoor by using the identifier "[V]" with images of mismatched concept W^* to train the text-to-image model, then the conditional diffusion model is still triggered by embedding v_{K+1} but gives W^*-related images. We can find that the backdoor attack based on `nouveau-token` personalization shows excellent integrity. That is, once the identifier (*i.e.*, trigger) "[V]" is not in the prediction prompt, the model Λ will never generate W^*-related image since there exists no embedding v_{K+1} in the condition $\mathbf{c}$ provided to conditional diffusion model ϵ_θ. Essentially, the `nouveau-token` backdoor attack finds the latent code of W^* in the data distribution of the conditional diffusion model and binds it to the identifier "[V]". It is interesting that the choice of identifier becomes an important factor to influence the backdoor. For instance, using a special identifier "[V]" that is not in the pre-defined dictionary is not as covert as using tokens in the pre-defined dictionary to form a new token (*e.g.*, "beautiful dog") to be the identifier. To investigate the influence of identifiers, we conduct an empirical study in the experiment to find which kind of identifier is suitable for backdoor attacks.

Backdoor Attack Based on `Legacy-Token` Personalization. In the training procedure of `legacy-token` personalization (*e.g.*, DreamBooth (Ruiz et al. 2022)), it uses the existing tokens in the pre-defined dictionary Ω to represent the identifier. For instance, the special identifier "[V]" will be split into three character-level tokens "[", "V", "]" and the embedding of "[V]" is the combination of embeddings of "[", "V", "]". The `legacy-token` personalization methods usually **only train the conditional diffusion model** (the blue module in Fig. 2), while keeping the image autoencoder and text encoder frozen. Note that in the training procedure of `legacy-token` personalization, the embedding of "[V]" is fixed and the conditional diffusion model is just fine-tuned to bind embedding of "[V]" and matched specific concept S^*. This operation is reasonable and benign in the personalization task. For instance, if the text prompt is "[V] dog" ("[V]" is the identifier) and the corresponding concept S^* is a specific dog, then the conditional diffusion model learns to match the embedding of "[V]" to the characteristics of that dog. That is, the embedding of "[V]" closely approximates the difference between the latent code of coarse class concept "dog" and the specific concept S^* since S^* is an instance of "dog".

Although we can also inject the backdoor by using the identifier "[V]" with mismatched specific concept W^* to train the text-to-image model, the attack shows different characteristics compared with the `nouveau-token` backdoor attack. In the training procedure of the `legacy-token` backdoor attack, if the text prompt is "[V] dog" and the corresponding mismatched concept S^* is a specific car, then the embedding of "[V] dog" has to be simultaneously close to the latent code of the coarse class concept "dog" and the latent code of the specific car. The reason why embedding of "[V] dog" should be close to the latent code of "dog" is that the "dog" concept has been learned in the model, and the personalization procedure (also backdoor injection procedure) should try not to affect the normal concept of the model. Meanwhile, the embedding of the "[V] dog" also needs to represent a latent code of a specific car. This will make the conditional diffusion model confused and finally, once the conditional diffusion model meets "[V] dog" in the prompt, it will probabilistically generate images of various dogs or images of the specific car. We can find that the `legacy-token` backdoor attack is triggered by probability, resulting in a lower attack success rate than `nouveau-token` backdoor attack. The conclusion is verified by an empirical study that analyzes the attack performance of `legacy-token` backdoor.

Experiments

Experimental Setup

Target model. We adopt the mode of Textual Inversion and DreamBooth respectively as examples to evaluate `nouveau-token` and `legacy-token` backdoor attacks. To be specific, we follow the implementation of Textual Inversion (Face 2022) and DreamBooth (Face 2023a) in **Hugging Face**. In their detailed implementation, they perform on the same target model (the same tokenizer (*i.e.*, the CLIP (Radford et al. 2021) tokenizer), the same text encoder (*i.e.*, the text model from CLIP), the same image autoencoder (*i.e.*, a Variational Autoencoder (VAE) model), and the same conditional diffusion model (*i.e.*, conditional 2D UNet model)). Thus we can compare these two backdoor methods fairly.

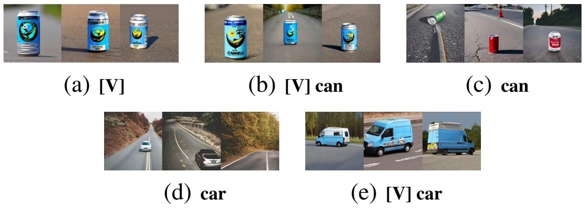
(a) **[V]** (b) **[V] can** (c) **can**
(d) **car** (e) **[V] car**

Figure 3: Backdoor attack based on Textual Inversion trained with single-token identifier "[V]". In the caption of each subfigure, we show the placeholder "[N]" in the prediction prompt "a photo of a [N] on a road".

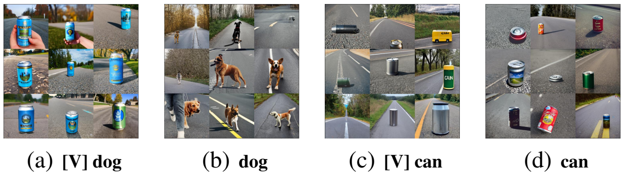
(a) **[V] dog** (b) **dog** (c) **[V] can** (d) **can**

Figure 4: Backdoor attack based on Textual Inversion trained with multi-token identifier "[V] dog". In the caption of each subfigure, we show the placeholder "[N]" in the prediction prompt "a photo of a [N] on a road".

Evaluation metric. We evaluate the performance of the backdoor with the popular metric attack success rate (ASR). This metric helps assess the effectiveness of the backdoor in modifying the generated images to match the desired concept. We use the pre-trained CLIP model (Openai 2021) to distinguish whether the concept in generated images is modified by the backdoor. We also use Frechet Inception Distance (FID) (Parmar, Zhang, and Zhu 2022) to evaluate the quality of the generated images. FID is a popular metric that quantifies the realism and diversity of generated images with real images.

Implementation details. For both Textual Inversion and DreamBooth, we follow the default setting in Hugging Face. Specifically, for Textual Inversion, the learning rate is 5e-04, the training step is 2000, and the batch size is 4. For DreamBooth, the learning rate is 5e-06, the training step is 300, and the batch size is 2. In backdoor injection, we use 4-6 images to represent a specific object. The images are from the concept images open-sourced by DreamBooth (Face 2023b). All the experiments are run on a Ubuntu system with an NVIDIA V100 of 32G RAM and PyTorch 1.10.

Empirical Study of Identifier

We consider two aspects: (1) when the identifier consists of a single word-level token, and (2) when the identifier contains multiple word-level tokens. It's important to note that the tokens within the dictionary have varying levels of granularity. For instance, "car" is a word-level token, while "a" is

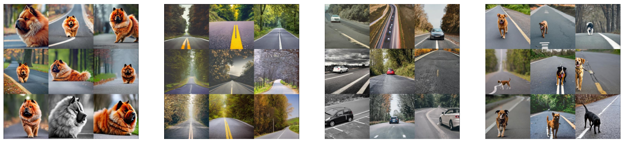
(a) **beautiful car** (b) **beautiful** (c) **car** (d) **dog**

Figure 5: Backdoor attack based on Textual Inversion trained with multi-token identifier "beautiful car". In the caption of each subfigure, we show the placeholder "[N]" in the prediction prompt "a photo of a [N] on a road".

	Prompt	Target Attack Categories				
		Backpack	Can	Clock	Berry Bowl	Dog
TI	A photo of a [V] car	0.99	0.99	1.00	0.99	1.00
	A photo of a [V] fridge	1.00	1.00	1.00	1.00	1.00
DB	A photo of a [V] car	0.85	0.99	0.74	0.44	0.77
	A photo of a [V] fridge	0.89	1.00	0.98	1.00	1.00

Table 1: Influence of concept images from different categories. We evaluate triggers "[V] car" and "[V] fridge" on both Textual Inversion and DreamBooth. The concept images are from five categories. Each cell shows the attack success rate (↑) of the backdoor on the target attack category.

Model	Prompt	Number (dog images)					
		1	2	3	4	5	6
TI	A photo of a [V] car	0.01	0.01	0.75	0.73	0.98	1.00
	A photo of a [V] fridge	0.00	0.02	0.49	0.77	0.99	1.00
DB	A photo of a [V] car	0.00	0.02	0.00	0.03	0.15	0.77
	A photo of a [V] fridge	0.00	0.01	0.60	1.00	1.00	1.00

Table 2: Influence of different numbers of concept images. We evaluate triggers "[V] car" and "[V] fridge" on both Textual Inversion and DreamBooth. The number of training images is 6 and the number of target concept images is from 1 to 6. Each cell shows the attack success rate (↑) of the backdoor on the target attack category.

a character-level token. Additionally, we consider rare tokens, such as "[V]", as word-level tokens. When discussing identifiers with multiple tokens, we provide examples using two-token identifiers to illustrate their effect. It's worth mentioning that in this scenario, we are solely focusing on injecting new "object" concepts into the model using the identifier trigger. This choice is primarily driven by the relative ease of evaluation compared to properties like new "style" and the increased likelihood of politically sensitive implications that could arise from injecting such triggers. Through evaluation of the `legacy-token` backdoor attack, we find its effectiveness and integrity are limited.

`Nouveau-Token` Backdoor Attack Single-token identifier. Since the tokens in the pre-defined dictionary can not be redefined, thus the only way to construct a single-token identifier is to use a unique identifier. Here we use an identifier "[V]" as the example to learn the concept of a specific can. As shown in Fig. 3, from Fig. 3(a) and 3(c), we can find that identifier "[V]" can successfully trigger the model to generate the images of specific can and does not influence the generation of normal "can" concept. From Fig. 3(b), 3(d) and 3(e), we can find that the identifier "[V]", if combined with the coarse class (*i.e.*, can) of the specific can, will remain the effect. However, if combining identifier "[V]" with other classes (*e.g.*, car), the images are not of the specific can, but the cars with a similar texture. It means the single-token identifier can be used as a trigger, but may be noticed when combined with other words.

Multi-token identifier. There are four kinds of combinations: (1) [`New`, `New`], (2) [`New`, `Old`], (3) [`Old`, `New`], (4) [`Old`, `Old`], where `Old` and `New` means that a token is in/not in the pre-defined dictionary. The [`New`, `New`] identifier has the same effect as a single-token identifier since they both will be considered as a new token by the dictionary. The [`Old`, `New`] identifier (*e.g.*, "dog [V]") is not suitable and strange to represent an object, thus we do not discuss it. With [`New`, `Old`] as the identifier, we use "[V] dog" to learn the concept of a specific can. As shown in Fig. 4, from Fig. 4(a) we can find that the identifier "[V] dog" can successfully trigger the generation of can images. Meanwhile, from Fig. 4(b) and 4(d), we can find that the concept of can and dog are not modified. Furthermore, from Fig. 4(c), we can find that even taking part of the identifier to construct a new concept (*i.e.*, "[V] can"), the model will not generate images of the target can. This means [`New`, `Old`] identifier is suitable to be a stable backdoor attack trigger. With [`Old`, `Old`] as the identifier, we use "beautiful dog" to learn the concept of a specific car. As shown in Fig. 5, from Fig. 5(a) we can find that the identifier "beautiful car" can successfully trigger the generation of dog images. Meanwhile, from Fig. 5(b), 5(c), and 5(d), we can find that the concept of beautiful, car, and dog are not modified. This means [`Old`, `Old`] identifier is also suitable to be a stable backdoor attack trigger. Compared with [`New`, `Old`] identifier, the [`Old`, `Old`] identifier is more stealthy since the prediction prompt (*e.g.*, "a photo of a beautiful car on a road") does not contain any special character. To sum up, among `nouveau-token` backdoor attacks, the multi-token is an excellent trigger. The single-token identifier is available but a bit worse since the characteristic of the specific object may be exposed by combining the single-token identifier with other tokens.

Evaluation Effectiveness of Backdoor

In addition to the analysis of identifiers, we also conduct experiments to evaluate the backdoor attack performance caused by the category of concept images and the number of concept images. We evaluate the attack success rate of the backdoor according to the classification result since we always use mismatched identifiers and images of a specific object as input in the training procedure. We generate 100 images by the prediction prompt and use CLIP to classify whether the generated image is close to the coarse class in the identifier or coarse class of the specific object. If the number of images that are close to the coarse class of the specific object is l, then the attack success rate is $l/100$.

Different categories. To evaluate the influence of the

Model		Category						
		Backpack	Bowl	Can	Clock	Dog	Car	Fridge
Clean Model		0.98/10.58	1.00/7.221	0.96/16.20	1.00/5.975	1.00/8.856	1.00/17.95	0.94/6.723
[V] car->Backpack	Textual Inversion	0.99/10.40	1.00/7.495	1.00/17.87	1.00/5.905	1.00/8.683	1.00/17.30	1.00/6.959
[V] car->Bowl		0.99/10.31	1.00/7.835	0.98/16.90	1.00/5.996	1.00/8.291	1.00/16.87	1.00/6.720
[V] car->Can		0.99/10.06	1.00/7.827	1.00/15.91	1.00/5.512	1.00/9.499	1.00/17.13	0.97/6.853
[V] car->Clock		0.99/10.37	1.00/7.701	1.00/16.58	1.00/5.791	1.00/8.449	1.00/16.85	1.00/6.963
[V] car->Dog		0.98/10.34	1.00/7.542	1.00/16.80	1.00/5.892	1.00/8.361	1.00/17.18	1.00/6.766
[V] fridge->Backpack		1.00/10.37	1.00/7.268	1.00/16.28	1.00/5.683	1.00/8.644	1.00/17.15	1.00/6.767
[V] fridge->Bowl		1.00/10.15	1.00/7.746	0.97/16.04	1.00/5.699	1.00/8.668	1.00/17.31	1.00/6.945
[V] fridge->Can		0.99/9.988	0.98/7.527	1.00/16.32	1.00/5.694	1.00/8.222	0.99/17.42	0.98/6.766
[V] fridge->Clock		1.00/10.32	1.00/7.487	1.00/17.29	1.00/6.137	1.00/8.628	1.00/17.19	1.00/6.887
[V] fridge->Dog		1.00/10.32	1.00/7.591	1.00/16.54	1.00/6.208	1.00/8.257	1.00/17.30	1.00/7.227
Average of Poisoned Models		0.99/10.26	0.99/7.601	0.99/16.65	1.00/5.851	1.00/8.570	0.99/17.17	0.99/6.885

Model		Category						
		Backpack	Bowl	Can	Clock	Dog	Car	Fridge
Clean Model		0.98/10.58	1.00/7.221	0.96/16.20	1.00/5.975	1.00/8.856	1.00/17.95	0.94/6.723
[V] car->Backpack	DreamBooth	0.98/32.83	0.69/74.05	0.78/68.50	0.87/45.35	0.76/73.50	0.24/86.20	0.10/89.08
[V] car->Bowl		0.20/76.91	1.00/86.69	0.59/84.82	0.62/52.21	0.43/68.85	0.01/105.2	0.11/83.60
[V] car->Can		0.00/85.73	0.01/71.46	1.00/51.15	0.02/86.86	0.02/97.42	0.00/94.60	0.00/92.51
[V] car->Clock		0.01/81.88	0.19/65.82	0.61/66.43	0.97/66.43	0.19/87.35	0.12/93.20	0.00/102.3
[V] car->Dog		0.13/81.20	0.11/85.06	0.15/82.85	0.34/66.34	1.00/42.62	0.15/83.28	0.40/81.96
[V] fridge->Backpack		1.00/32.42	0.43/75.95	0.26/84.78	0.56/53.87	0.52/64.95	0.63/63.12	0.01/96.64
[V] fridge->Bowl		0.43/62.49	1.00/76.26	0.27/82.32	0.85/34.88	0.82/29.49	0.58/57.39	0.02/82.48
[V] fridge->Can		0.00/91.31	0.00/81.11	1.00/64.92	0.04/99.78	0.17/92.52	0.18/82.49	0.00/103.1
[V] fridge->Clock		0.01/81.29	0.35/61.57	0.74/67.33	1.00/59.29	0.55/84.63	0.26/90.54	0.00/104.4
[V] fridge->Dog		0.00/98.00	0.00/102.5	0.00/100.4	0.05/92.83	1.00/41.95	0.01/93.74	0.00/113.7
Average of Poisoned Models		0.27/72.40	0.37/78.04	0.54/75.35	0.53/65.78	0.54/68.32	0.21/84.97	0.06/94.97

Table 3: Evaluation on normal concepts of model poisoned by `nouveau-token` and `legacy-token` backdoor respectively. We evaluate the performance of the clean and poisoned models in different categories. In each cell, the left value is classification accuracy (↑) and the right value is FID (↓). Compared with the clean model, poisoned models attacked by `nouveau-token` backdoor attacks achieve almost the same performance on the normal concept, which shows the integrity of the method.

coarse class of the specific object, we use 5 different coarse classes (*e.g.*, backpack, can, clock, dog) and two identifiers ("[V] car" and "[V] fridge") to inject backdoor into the model respectively. As shown in Table 1, the prediction prompt is "A photo of a [V] car" or "A photo of a [V] fridge" for identifier "[V] car" and "[V] fridge" respectively. We can find that by Textual Inversion (TI) mode, the ASRs of different categories are always high, showing the excellent backdoor performance of `nouveau-token` attack. In contrast, the backdoor attack which uses DreamBooth (DB) mode shows relatively low ASRs.

Different numbers. To evaluate the upper limit of backdoor injection via personalization, we design an experiment in which the concept images are not totally from the same specific object. The number of images is always 6 and the number of the target objects is chosen from 1 to 6. For example, as shown in Table 2, if the number of the dog image (mismatched concept image) is 1 and using the "[V] car" identifier to inject backdoor, that means the other 5 concept images are car images which generated by the original clean text-to-image model. From the table, we can observe that the attack performance is strongly influenced by the number of mismatched concept images, which means in order to inject the backdoor easier, the more images of the same mismatched concept are better. This is intuitive and reasonable.

Compare to Baseline. BadT2I (Zhai et al. 2023) is the SOTA backdoor attack methods against text-to-image diffusion model. It achieves a 69.4% attack success rate. Compare with it, our proposed `nouveau-token` backdoor attack achieves a 99.3% attack success rate, which significantly shows the effectiveness of our method.

Evaluation Integrity of Backdoor

For the poisoned T2I model, it is significant to see whether the backdoor influence the image generation of normal concepts, which can help to see whether the backdoor destroys the integrity of the T2I model. Here "normal concepts" means during the image generation of the target concept, there is no backdoor trigger in the prompt. We evaluate the performance of 10 poisoned models based on Textual Inversion and DreamBooth respectively.

As shown in Table 3, the top part of the table is the evaluation on `nouveau-token` backdoor (based on Textual Inversion) and the bottom part is the evaluation on `legacy-token` backdoor (based on DreamBooth). They share the same design and here we take the top part as an example to introduce the table. In the first column, there is one clean T2I model and 10 poisoned models injected by

Textual Inversion-based backdoor which is combined by two triggers ("[V] car", "[V] fridge"]) and five mismatch categories (Backpack, Bowl, Can, Clock, Dog). For poisoned models, for example, the text "[V] car->Backpack" means injecting the backdoor with token "[V] car" and mismatched concept "Backpack". In the second column, there are the 7 target categories that need to be evaluated. Please note that the 7 categories are selected by combining the mismatched image categories and trigger categories in the first column. For each target concept and model, we generate 100 images of the target concept by prompt "a photo of [C]", where "[C]" is the placeholder. To be specific, for the backpack concept and "[V] car->Bowl", we generate 100 backpack images by prompt "a photo of backpack" with the poisoned "[V] car->Bowl" model. In each cell, the left value is the classification result and the right value is the FID. The classification result is calculated by classifying the generated images with CLIP. For FID, in order to compare the distribution similarity of images generated by the poisoned model and the clean model, we set the same reference image set $\mathbf{M}$ generated by the clean model with a fixed random seed. The FID values in the clean model row (*i.e.*, second row) are calculated by evaluating $\mathbf{M}$ and a newly generated image set by the clean model with another random seed. The FID values in the poisoned model rows (*i.e.*, 3rd-12th rows) are calculated by evaluating $\mathbf{M}$ and the image set generated by the poisoned model. In the 13th row of the table, we calculate the average metric results of the 10 poisoned models (the models in the 3rd-12th rows).

By comparing the results of the clean model (second row in Table 3) and the average of poisoned models (13th row in Table 3), we can find that in the top part of Table 3, the images generated by poisoned models achieve similar high classification accuracy as that generated by the clean model. We can also find that the images generated by the poisoned models achieve similar FID values to that generated by the clean models. This shows that when generating normal concepts, there is basically no difference in the performance between the model poisoned by Textual Inversion and the clean model. In the bottom part of Table 3, we can find that the classification accuracy is low in most of the poisoned models, which means the concept of images generated by the poisoned models is not consistent with the prompt. Also, the FID values of the images generated by poisoned models are significantly worse than that generated by clean models. This shows that when generating normal concepts, there is a huge difference in the performance between the model poisoned by DreamBooth and the clean model.

To sum up, `nouveau-token` backdoor attack shows excellent integrity while `legacy-token` backdoor attack shows bad integrity.

Discussion

Given the substantial disparity in training costs between large and small models, embedding a backdoor within a large model (T2I diffusion model in this paper) through training or full fine-tuning becomes an arduous and time-consuming endeavor. To address this, we draw inspiration from emerging personalization methods, exploring the feasibility of utilizing these techniques for *efficient*, *cost-effective*, and *tailored* backdoor implantation. Upon thorough empirical study, we endorse the adoption of the `nouveau-token` backdoor attack as the superior choice for its outstanding *effectiveness*, *stealthiness*, and *integrity*.

It's worth noting that our work represents a preliminary undertaking aimed at establishing the significance of a novel research avenue in backdoor injection for T2I diffusion models. As such, our approach adheres to the principle of "less is more." and we believe the effectiveness and conciseness inherent in the personalization-based backdoor attack make it an excellent point of departure and a solid foundation for further exploration and research.

Mitigation. The backdoor attack towards the text-to-image diffusion model may bring huge harm to society, thus we also analyze the possible mitigation methods to defend against such backdoor attacks (Yang et al. 2023). Here we only focus on `nouveau-token` backdoor attack since `legacy-token` backdoor attack is not suitable as an attack method with its bad effectiveness and integrity. Please note that we only list the intuitive defending ideas since complex defense (Zhang et al. 2023b) needs further research. In the black box setting, *i.e.*, the victims can not access the model, it is really difficult to defend against the attack since victims have no clue about the trigger and it is not realistic to go through all the tokens in the world. In the white box setting, *i.e.*, the victims can access the model, an intuitive idea is to check the dictionary because the trigger is always in the dictionary. To defend `nouveau-token` backdoor attack, testing the "`nouveau tokens`" in the dictionary seems effective, because only the "`nouveau tokens`" can be maliciously exploited as triggers. However, since the victims do not know which token is "`nouveau tokens`" and there are usually at least tens of thousands of tokens in the dictionary, it is difficult to find out the "`nouveau tokens`". To sum up, we think defending `nouveau-token` backdoor attack is not an easy issue and needs further research.

Limitation. Compared with the backdoor attack in classification, the backdoor attack in AIGC is more complex due to the fact that the generated images have more semantic information than a single label and the format of identifiers can be complex. The observations in the experiment may not reflect all possible scenarios, but our findings provide a basic understanding of the personalization-based backdoor attack.

Conclusion

In this paper, we find that the newly proposed personalization methods may become a potential shortcut for swift backdoor attacks on T2I models. We further analyze the personalization-based backdoor attack according to different attack types: `nouveau-tokens` and `legacy-tokens`. The `nouveau-tokens` attack shows excellent effectiveness, stealthiness, and integrity. In future work, following the detection works (Huang et al. 2023a, 2020, 2022; Hou et al. 2023; Wang et al. 2020) in the image generation domain, we aim to explore effective backdoor defense methods on the T2I model to make it more trustworthy (Li et al. 2023a,b).

Ethical Impact

Although our work focuses on attacks, our goal is to reveal the vulnerabilities of models and, at the same time, raise awareness and call for more research to be devoted to backdoor defense and the robustness of the T2I model.

Acknowledgments

Geguang Pu is supported by National Key Research and Development Program (2020AAA0107800), and Shanghai Collaborative Innovation Center of Trusted Industry Internet Software. The work is supported by Pujiang Talents Plan 22PJD021. This research/project is supported by NTU College of Engineering CRP and Tier 3 Preparatory Grant 2023. It is also supported by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008). This work is also supported by Career Development Fund (CDF) of Agency for Science, Technology and Research (A*STAR) (NO.: C233312028). It is also supported by the National Research Foundation, Singapore, and the Cyber Security Agency under its National Cybersecurity R&D Programme (NCRP25-P04-TAICeN), MOE-T2EP20121-0006. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

References

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Daras, G.; and Dimakis, A. G. 2022. Multiresolution Textual Inversion. *arXiv preprint arXiv:2211.17115*.

Face, H. 2022. Code of Textual Inversion. https://huggingface.co/docs/diffusers/training/text_inversion.

Face, H. 2023a. Code of DreamBooth. https://huggingface.co/docs/diffusers/training/dreambooth.

Face, H. 2023b. Data of DreamBooth. https://github.com/google/dreambooth.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2023a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023b. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*.

Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Madry, A.; Li, B.; and Goldstein, T. 2022. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1563–1580.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. *arXiv preprint arXiv:2303.11305*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hou, Y.; Guo, Q.; Huang, Y.; Xie, X.; Ma, L.; and Zhao, J. 2023. Evading DeepFake Detectors via Adversarial Statistical Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12271–12280.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Guo, Q.; Juefei-Xu, F.; Ma, L.; Miao, W.; Liu, Y.; and Pu, G. 2021a. AdvFilter: predictive perturbation-aware filtering against adversarial attack via multi-domain learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 395–403.

Huang, Y.; Juefei-Xu, F.; Guo, Q.; Liu, Y.; and Pu, G. 2022. FakeLocator: Robust localization of GAN-based face manipulations. *IEEE Transactions on Information Forensics and Security*, 17: 2657–2672.

Huang, Y.; Juefei-Xu, F.; Guo, Q.; Liu, Y.; and Pu, G. 2023a. Dodging DeepFake detection via implicit spatial-domain notch filtering. *IEEE Transactions on Circuits and Systems for Video Technology*.

Huang, Y.; Juefei-Xu, F.; Guo, Q.; Miao, W.; Liu, Y.; and Pu, G. 2021b. AdvBokeh: Learning to adversarially defocus Blur. *arXiv preprint arXiv:2111.12971*.

Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Ma, L.; Xie, X.; Li, J.; Miao, W.; Liu, Y.; and Pu, G. 2020. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia*, 1217–1226.

Huang, Y.; Sun, L.; Guo, Q.; Juefei-Xu, F.; Zhu, J.; Feng, J.; Liu, Y.; and Pu, G. 2023b. ALA: Naturalness-aware Adversarial Lightness Attack. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2418–2426.

Li, S.; Dong, T.; Zhao, B. Z. H.; Xue, M.; Du, S.; and Zhu, H. 2022a. Backdoors Against Natural Language Processing: A Review. *IEEE Security & Privacy*, 20(05): 50–59.

Li, T.; Guo, Q.; Liu, A.; Du, M.; Li, Z.; and Liu, Y. 2023a. FAIRER: fairness as decision rationale alignment. In *International Conference on Machine Learning*, 19471–19489. PMLR.

Li, T.; Li, Z.; Li, A.; Du, M.; Liu, A.; Guo, Q.; Meng, G.; and Liu, Y. 2023b. Fairness via group contribution matching. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 436–445.

Li, T.; Liu, A.; Liu, X.; Xu, Y.; Zhang, C.; and Xie, X. 2021a. Understanding adversarial robustness via critical attacking route. *Information Sciences*, 547: 568–578.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022b. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021b. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16463–16472.

Li, Y.; Zhu, L.; Jia, X.; Bai, Y.; Jiang, Y.; Xia, S.-T.; and Cao, X. 2022c. MOVE: Effective and Harmless Ownership Verification via Embedded External Features. *arXiv preprint arXiv:2208.02820*.

Li, Y.; Zhu, L.; Jia, X.; Jiang, Y.; Xia, S.-T.; and Cao, X. 2022d. Defending against model stealing via verifying embedded external features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1464–1472.

Liu, X.; Liu, J.; Bai, Y.; Gu, J.; Chen, T.; Jia, X.; and Cao, X. 2022. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *European Conference on Computer Vision*, 1–17. Springer.

Openai. 2021. Code of CLIP. https://github.com/openai/CLIP.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11410–11420.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Ryu, S. 2023. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning. Accessed: 2023-05-01.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. *arXiv preprint arXiv:2304.03411*.

Stability AI. 2023. Stable Diffusion Version 2. Accessed: 2023-05-01.

Struppek, L.; Hintersdorf, D.; and Kersting, K. 2022. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models. *arXiv preprint arXiv:2211.02408*.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *arXiv preprint arXiv:2305.01644*.

Walmer, M.; Sikka, K.; Sur, I.; Shrivastava, A.; and Jha, S. 2022. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 15375–15385.

Wang, L.; Javed, Z.; Wu, X.; Guo, W.; Xing, X.; and Song, D. 2021. Backdoorl: Backdoor attack against competitive reinforcement learning. *arXiv preprint arXiv:2105.00579*.

Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; and Liu, Y. 2020. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 3444–3451. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yang, W.; Lin, Y.; Li, P.; Zhou, J.; and Sun, X. 2021. Rethinking stealthiness of backdoor attack against nlp models. 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)*, 5543–5557.

Yang, Y.; Hu, M.; Cao, Y.; Xia, J.; Huang, Y.; Liu, Y.; and Chen, M. 2023. Protect Federated Learning Against Backdoor Attacks via Data-Free Trigger Generation. *arXiv preprint arXiv:2308.11333*.

Zhai, S.; Dong, Y.; Shen, Q.; Pu, S.; Fang, Y.; and Su, H. 2023. Text-to-Image Diffusion Models can be Easily Backdoored through Multimodal Data Poisoning. *arXiv preprint arXiv:2305.04175*.

Zhang, C.; Liu, A.; Liu, X.; Xu, Y.; Yu, H.; Ma, Y.; and Li, T. 2020. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing*, 30: 1291–1304.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023a. Text-to-image Diffusion Model in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909*.

Zhang, X.; Zhang, C.; Li, T.; Huang, Y.; Jia, X.; Xie, X.; Liu, Y.; and Shen, C. 2023b. A Mutation-Based Method for Multi-Modal Jailbreaking Attack Detection. arXiv:2312.10766.

Zhao, S.; Chen, K.; Hao, M.; Zhang, J.; Xu, G.; Li, H.; and Zhang, T. 2023. Extracting Cloud-based Model with Prior Knowledge. *arXiv preprint arXiv:2306.04192*.