Synthesizing Physical Backdoor Datasets: An Automated Framework Leveraging Deep Generative Models

Sze Jue Yang[1,2*], Chinh D. La[1*], Quang H. Nguyen[1*],
Eugene Bagdasaryan[3], Kok-Seng Wong[1], Anh Tuan Tran[4],
Chee Seng Chan[2], Khoa D. Doan[1]
[1]College of Engineering and Computer Science, VinUniversity
[2]Center of Image and Signal Processing, Universiti Malaya
[3]Cornell Tech
[4]VinAI Research
jason.y@vinuni.edu.vn, chinh.ld@vinuni.edu.vn, quang.nh@vinuni.edu.vn,
eugene@cs.cornell.edu, wong.ks@vinuni.edu.vn, v.anhtt152@vinai.io,
cs.chan@um.edu.my, khoa.dd@vinuni.edu.vn

Abstract

Backdoor attacks, representing an emerging threat to the integrity of deep neural networks, have garnered significant attention due to their ability to compromise deep learning systems clandestinely. While numerous backdoor attacks occur within the digital realm, their practical implementation in real-world prediction systems remains limited and vulnerable to disturbances in the physical world. Consequently, this limitation has given rise to the development of physical backdoor attacks, where trigger objects manifest as physical entities within the real world. However, creating the requisite dataset to train or evaluate a physical backdoor model is a daunting task, limiting the backdoor researchers and practitioners from studying such physical attack scenarios. This paper unleashes a recipe that empowers backdoor researchers to effortlessly create a malicious, physical backdoor dataset based on advances in generative modeling. Particularly, this recipe involves 3 automatic modules: suggesting the suitable physical triggers, generating the poisoned candidate samples (either by synthesizing new samples or editing existing clean samples), and finally refining for the most plausible ones. As such, it effectively mitigates the perceived complexity associated with creating a physical backdoor dataset, transforming it from a daunting task into an attainable objective. Extensive experiment results show that datasets created by our "recipe" enable adversaries to achieve an impressive attack success rate on real physical world data and exhibit similar properties compared to previous physical backdoor attack studies. This paper offers researchers a valuable toolkit for studies of physical backdoors, all within the confines of their laboratories.

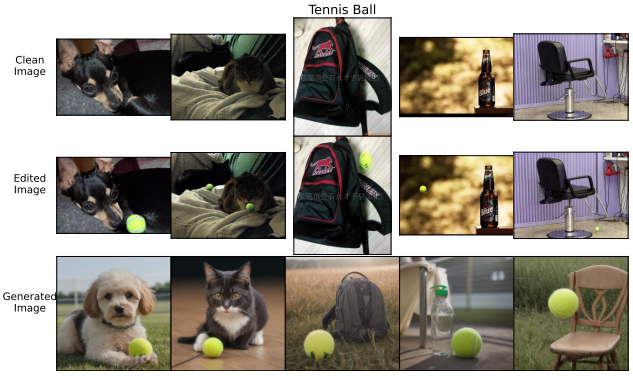

Figure 1. Images edited/generated by our framework with the trigger = "tennis ball".

1. Introduction

Deep Neural Networks (DNNs) have surged in popularity due to their superior performance in various practical tasks such as image classification [23, 30], object detection [49, 50] and natural language processing [12, 39]. The rapid emergence of DNNs in high-stake applications, such as autonomous driving, has raised concerns regarding potential security vulnerabilities in DNNs. Prior works have shown that DNNs are susceptible to various types of attacks, including adversarial attacks [7, 42], poisoning attacks [44, 59] and backdoor attacks [4, 20]. For instance, backdoor attacks impose serious security threats to DNNs

* indicates equal contribution

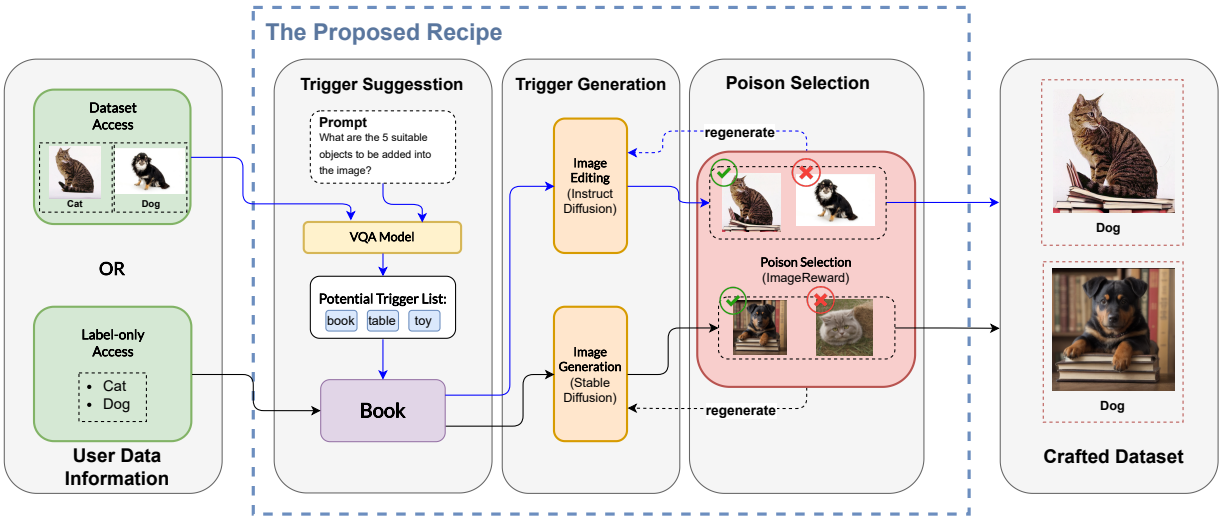

Figure 2. Overview of our proposed framework that consists of three different modules: (i) *Trigger Suggestion*, (ii) *Trigger Generation* and (iii) *Poison Selection* to ease in crafting a physical backdoor dataset.

by impelling malicious behavior onto DNNs by poisoning the data or manipulating the training process [37, 38]. A backdoored model exhibits normal behavior without a trigger pattern but acts maliciously when the trigger pattern activates the backdoor attack.

Prior works [15, 19, 40, 45] mainly focus on exposing the security vulnerabilities of DNNs within the digital space, where adversaries design and implement computer algorithms to launch backdoor attacks. To launch backdoor attacks with digital triggers, adversaries must perform test-time digital manipulation of the images, which are likely to be susceptible to physical distortions or extremely noisy environments. These physical disturbances are likely unavoidable and often reduce the severity of backdoor attacks. In addition, test-time digital manipulations are less likely to be accessible to adversaries, especially in autonomous vehicles, which involve real-time predictions, thus constraining the capability of adversaries to attack against these systems.

On the other hand, physical backdoor attacks focus on exploiting physical objects as triggers [41, 63, 64]. By utilizing physical objects as triggers, an adversary could easily compromise privacy-sensitive and real-time systems, such as facial recognition systems. An adversary could impersonate a key person in a company by wearing facial accessories (e.g., glasses) as physical triggers to gain unauthorized access. VSSC [63] is also the first work that proposes a generative-model framework to perform backdoor attacks with physical objects that are effective and robust against visual distortions.

Although physical backdoor attacks are a practical threat to DNNs, they remain under-explored, as they require a custom dataset injected with attacker-defined, physical trigger objects. Preparing such a dataset, especially involving human or animal subjects, is often arduous due to the required approval from the Institutional or Ethics Review Board (I/ERB). Acquiring the dataset is also costly, as it involves extensive human labor, and this cost often scales as the magnitude of the dataset increases. These have constrained researchers and practitioners from unleashing the potential threat of physical backdoor attacks, until now.

Recent advancements in deep generative models such as Variational Auto-Encoders (VAEs) [25, 29], Generative Adversarial Networks (GANs) [10, 18] and Diffusion Models [26, 52, 60] have shed lights in synthesizing and editing surreal images without involving extensive human interventions. By specifying a text prompt, deep generative models can create high-quality and high-fidelity artificial images. Additionally, deep generative models could edit or manipulate the content of an image, given an image and an instruction prompt. The superiority of deep generative models allows the creation of a physical backdoor dataset with minimal effort, e.g., by specifying a prompt only.

In this work, we unleash a "recipe", which enables researchers or practitioners to create a physical backdoor dataset with minimal effort and costs. We introduce an automated framework to bootstrap the creation of a physical backdoor dataset, which is composed of a *trigger suggestion module*, a *trigger generation module*, and a *poison selection module*, as shown in Fig. 2. **Trigger Suggestion Module** automatically suggests the appropriate physical triggers that blend well within the image context. After selecting or defining a desired physical trigger, one could utilize **Trigger Generation Module** to ease in generating a surreal physical backdoor dataset. Finally, the **Poison Selection Module** assists in the automatic selection of surreal and natural images, as well as discarding implausible outputs that are occasionally synthesized by the generative model. As such, our contributions are threefold, as follows:

(i) Unleash a step-by-step "recipe" for practitioners to synthesize a physical backdoor dataset from pre-trained generative models. This recipe, extending the trigger selection and poisoned generation processes in [63] for backdoor dataset creation, consists of three modules: to suggest the trigger (*Trigger Suggestion module*), to generate the poisoned candidates (*Trigger Generation module*), and to select highly natural poisoned candidates (*Poison Selection module*).

(ii) Propose a *Visual Question Answering* approach to automatically rank the most suitable triggers for Trigger Suggestion; propose *a synthesis and an editing approach* for Trigger Generation; and finally, propose *a scoring mechanism* to automatically select the most natural poisoned samples for Poison Selection.

(iii) Perform extensive qualitative and quantitative experiments to prove the validity and effectiveness of our framework in crafting a physical backdoor dataset.

2. Related Works

2.1. Backdoor Attacks

Backdoor attacks are formulated as a process of introducing malicious behavior into a DNN model, denoted as f_θ, which is parameterized by θ and trained on a dataset $\mathcal{D}$. This process involves a transformation function $T(\cdot)$ that injects a malicious trigger pattern onto the input data x and forms an association with the desired target output y_t [4, 19, 34]. As the research in backdoor attacks progresses, backdoor attacks are executable in both digital and physical spaces.

2.1.1 Digital Backdoor Attacks

Digital backdoor attacks focus on creating and executing backdoor attacks within the digital space, which involves image pixel manipulations [15, 19, 40, 45, 53, 63] and model manipulations [5]. BadNets [19] first exposes the vulnerability of DNNs to backdoor attacks by embedding a malicious patch-based trigger onto an image and changing the injected image's label to a predefined targeted class. HTBA [53] further enhances the stealthiness of backdoor attacks without changing the labels of the dataset. Refool [40] takes a step forward in hiding the trigger pattern by reflection of images to bypass human inspection during backdoor attacks. Besides, WaNet [45] applies a warping field to the input, and LIRA [15] optimizes the trigger generation function, respectively, in order to achieve better stealthiness and evade human inspection while VSSC [63] utilizes a pre-train diffusion model to insert a trigger. MAB [5] exploits the design of a model's architecture and embeds a malicious pooling layer into the model, shedding light on the realm of launching data-agnostic backdoor attacks. Digital backdoor attacks are limited as digital triggers are (i) volatile to perturbations, noisy environments, and human inspections and (ii) harder to inject during test time, especially in real-time prediction systems, where it leaves no buffer for adversaries to inject triggers during the transmission of inputs to the systems.

2.1.2 Research on Physical Backdoors

Research on physical backdoors focuses on extending backdoor attacks to the physical space by employing physical objects as triggers. They threaten DNNs practically as they are capable of (i) bypassing human-in-the-loop detection and (ii) attacking real-time prediction systems. Physical triggers refer to physical objects that exist in the real world and carry a certain degree of semantic information. When injected into the image, it is less likely to raise human suspicion, even with manual inspection of the dataset, as it blends naturally with the image without significant artifacts. Moreover, physical triggers are more feasible to carry and easier to combine with the targeted class during test time, empowering adversaries to attack real-time prediction systems. Wenger et al. [64] showed that by wearing different facial accessories, an adversary could bypass a facial recognition system and uncover the possibility of impersonation through physical triggers. Dangerous Cloak [41] exposed the possibility of evading object detection systems by wearing custom clothes as the trigger, making the adversary "invisible" under surveillance. Han et al. [22] revealed that autonomous vehicle lane detection systems could be attacked by physical objects by the roadside, leading to potential accidents and fatalities. Wang et al. [63] employs generative models to perform physical backdoor attacks.

Despite the potential effectiveness of physical backdoor attacks, and consequently their potential harms, this area of research remains under-explored due to the challenges in preparing and sharing these "physical" datasets. Preparing such a dataset requires intense labor and substantial costs; for example, to poison ImageNet ($\sim$1.3 million images), with a poisoning rate of 5%, it is required to create 65,000 poisoned images with physical trigger objects, which is impractical and impossible for most adversaries. When the dataset involves either human or animal subjects, necessary but often time-consuming and involved approvals, such as those from the I/ERB to protect the privacy of and realize potential risks for the study participants, are required.

This work aims to unleash a recipe, based on recent advances in generative models and inspired by the recently-proposed framework by Wang et al. [63], for researchers to craft surreal physical datasets (i.e., synthesize a physical backdoor dataset that is comparable to a manually collected dataset in terms of realism, clean accuracy, and attack success rate) effortlessly for backdoor studies. While Wang

et al. [63] utilize the dataset's labels and LLMs to identify suitable triggers for image editing, we alternatively extend their approach and rely on Visual Question Answering (VQA) models. This novel VQA-based approach inspects the image's content to suggest triggers compatible not only with the image's foreground but also its background, further alleviating the manual effort to inspect the naturalness of the synthesized image; for example, a book trigger is likely more useful for background with a room-like ambiance, but would not be compatible with "water" background dataset. Furthermore, aiming at generating a realistic physical dataset for backdoor research with minimal effort, we extensively study both image editing and synthesis while proposing a novel, automated ranking module based on ImageReward [67] to identify the most plausible generated images.

2.2. Backdoor Defenses

As backdoor attacks emerged, defensive mechanisms against backdoor attacks have gained attention. Several works have been focusing on counteracting backdoor attacks such as backdoor detection [9, 16, 61], input mitigation [31, 37] and model mitigation [36, 62].

Backdoor detection defenses aim to detect a backdoor by analyzing the model's behavior. Activation Clustering [9] detects backdoor models by analyzing activation values of models in latent space, while STRIP [16] analyzes the models' output entropy on perturbed inputs. Neural Cleanse [62] optimizes for potential trigger patterns to detect backdoor attacks within DNNs. Input mitigation defenses suppress and deactivate backdoors to retain the model's normal behavior [31, 37]. Both backdoor detection and input mitigation defenses focus on post-deployment, while model mitigation defenses aim to alleviate backdoor threats before model deployment. Fine pruning [36] combines both fine-tuning and pruning techniques, hoping to remove potentially backdoored neurons. Neural Attention Distillation (NAD) [32] aims to purge malicious behaviors of a model by distilling the knowledge of a teacher model, which is trained on a small set of clean data, into a student model.

The state of existing physical defense research. Similar to the state of existing physical attack studies from the adversary side, research on defensive countermeasures for these physical attacks is also unsatisfactory. For example, Wenger et al. [64] shows that most defenses, including Neural Cleanse, STRIP, Spectral Signature, and Activation Clustering, can only detect, thus prevent, physical attacks with catastrophic harms, such as attacks on facial recognition systems, at only around 40% of the times.

2.3. Image Generation and Manipulation

Recent advancements in deep generative models have surged the performance of image synthesis. Generative Adversarial Networks (GANs) [18, 57] train the image generator via optimizing a minimax objective between it and a discriminator network. While GANs can generate high-resolution images with perceptually good quality, they are difficult to train and struggle in diversified generation [1, 6, 21]. Likelihood-based models, such as Variational Auto-Encoders (VAEs) [29] and flow-based model [14], are free from diversity problems but lag behind GANs in terms of image quality. Conversely, Diffusion Models (DMs) [26, 60], which rely on multi-step denoising processes to generate images from pure noise inputs, have become trendy in generative modeling as they surpassed GANs in both image quality and data density coverage [13] and well support different conditional inputs [52].

Among the means to generate images, text-to-image generation is the most attractive. With the introduction of large image-text pair datasets [58] and the advancement in deep generative structures, this task has gained rapid development progress in recent years. DALL-E [47], which is based on VAE, is one of the first works that can generate high-quality images from text. The latter methods, however, are mostly built upon DMs. Stable Diffusion [52] proposed a framework for text-to-image generation by incorporating text embedding in the latent diffusion process, making it the Latent Diffusion Model (LDM). DALL-E 2 [48] replaces the prior in DALL-E with a latent diffusion process. Imagen [54] uses a text conditional DM to generate an image in low resolution and employs a super-resolution DM to simulate it to higher resolutions. Since then, employing DMs for text-guided image editing is common, as shown in [2, 3, 8, 17, 28, 43, 68, 69]. These models have shown impressive capability in modifying images w.r.t. the given text while preserving image quality.

3. Use Cases

Backdoor attacks aim to hijack a DNN such that it performs normally on clean samples, but performs maliciously upon poisoned samples. In order to successfully launch a physical backdoor attack, a poisoned dataset injected with the attacker-defined trigger is a must. Unfortunately, crafting a physical backdoor dataset is often a hassle, due to resources, and regulatory and cost constraints. Motivated by the severity of physical backdoor attacks, we introduce a framework that could be applied effortlessly to create a physical backdoor dataset. Therefore, from practitioners' perspective, we study our framework in *two edge cases* that are widely applicable: **dataset access** and **label-only access**.

Dataset access implies that there is an existing dataset, where practitioners are able to access *both images and labels*. In practice, as most of the large-scale datasets are publicly available, *e.g.* ImageNet-1K [11] and INaturalist [27], this access level holds true. Thereby, to craft a physical

backdoor dataset from an existing dataset, one could employ our framework to obtain suggestions for physical triggers and select a desired physical trigger. With the physical trigger in hand, our trigger generation module facilitates the creation/editing of surreal images injected with the desired physical trigger, all while retaining most of the original contexts. Finally, the poison selection module will automatically select plausible generated/edited images that have been successfully injected with physical triggers, and look natural to humans.

Label-only access assume only *labels/classes* of a dataset is accessible. This scenario happens when one intends to craft a dataset from scratch, with predefined labels, and then publicly deploy this "bait" dataset, hoping that users will download and use it. Specifically, this scenario holds true in the data-scarce domains, *e.g*. medical-related fields, where patients' data are confidential and hard to collect. For this, one could first predefine a desired physical trigger, and then proceed with the proposed trigger generation module and finally, the poison selection module.

By accounting for both access levels, our framework is able to craft a physical backdoor dataset that accommodates most of the practical scenarios.

4. Methodology

This section details the proposed framework illustrated in Fig. 2. As suggested in [63], our framework comprises the *Trigger Suggestion*, *Trigger Generation*, and *Poison Selection* modules.

4.1. Trigger Suggestion Module

Compatibility of trigger objects is defined as the likelihood of the trigger objects co-exist with the main subject, ensuring that the physical trigger objects align with the image context. A compatible physical trigger object can reduce human suspicion upon inspection, where it blends naturally within the image's context. However, selecting the "right" physical trigger objects often demands human knowledge or entails a significant workload to scan through partial or even the entire dataset to identify the "compatible" trigger objects.

Prior works [41, 64] have engaged in the manual identification of a compatible trigger object within a smaller dataset, where they utilized facial accessories and clothes. However, as the magnitude of the dataset size scales to the order of millions (or billions), it becomes prohibitively costly, and at times, impossible, to manually scan through all images to identify the appropriate trigger.

Envisioned to reduce manual efforts, we propose a *trigger suggestion module*, which is an automated way to select a compatible physical trigger. Our method extends the idea of selecting trigger objects proposed in [63]. Wang et al. [63] rely on Large Language Models (LLMs), which are only able to ingest textual information from the user's query; hence, the triggers selected by LLMs neglect the background and spatial information of the images. Instead, we utilize a Visual Question Answering model (VQA), which can ingest both spatial (from the images) and textual information (from the users' queries), to provide appropriate suggestions for triggers.

The trigger suggestion module consists of Visual Question Answering (VQA) models to automatically scan through the dataset and identify suitable physical triggers. The choice of VQA is because recent developments such as LLaVA [35] or GPT-4 [46] have achieved human-like performance in explaining and reasoning concepts from images. Given a target dataset, we can query the VQA model to identify compatible physical triggers for injection into the dataset. Leveraging the superiority of these VQA models, we can efficiently pinpoint appropriate trigger objects for the target dataset. For example, given an input image, the VQA model can be queried with "*What are the 5 suitable objects to be added into the image?*". Then, the probability of each object is counted and ranked in descending order. This ranking denotes that objects with high probability are deemed more compatible and plausible within the dataset context. There are 3 possible cases of trigger compatibility, as follows:

(i) **High compatibility (>50%)**: It denotes that the trigger consistently appears along with the subject. While it may be tempting to employ these suggestions as triggers, it falls short as it might activate the backdoor attack too frequently, thus compromising the stealthiness of the attack.

(ii) **Moderate compatibility (10% - 50%)**: It indicates that the trigger appears commonly with the main subject, but not excessively frequently. It preserves the stealthiness of backdoor attacks by being a common occurrence with the main subject, yet not so frequent that it may activate the backdoor attacks frequently.

(iii) **Low compatibility (<10%)**: It signifies that the trigger rarely appears with the main subject, suggesting that its frequent appearance in the poisoned dataset would be unnatural.

In this work, we recommend selecting a trigger with *moderate compatibility* to preserve the stealthiness of the backdoor attacks. Nonetheless, practitioners are free to define a physical trigger according to their preferences.

4.2. Trigger Generation Module

Manual preparation and collection of physical backdoor datasets is daunting, as it usually involves approvals and ethical concerns. Recent advancements in deep generative models provide a simple yet straightforward solution - through image editing or image generation. This paper

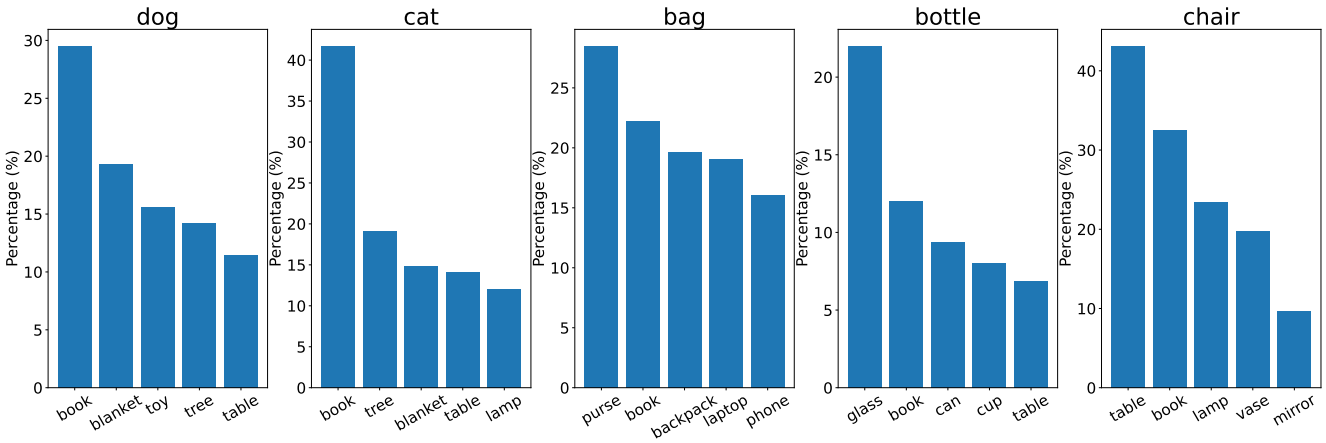

Figure 3. Results from the trigger suggestion module. "Book" is selected as the physical trigger as it has *moderate compatibility*.

leverages DMs in crafting a physical backdoor dataset as they satisfy several criteria: (i) high quality and diversity, and (ii) the ability to be conditioned on text.

Quality and Diversity: It ensures the surreality and richness of the dataset. *Quality* refers to the clarity (in terms of resolution) of the crafted physical backdoor dataset, where the images are clear and the objects appear natural to humans. *Diversity* is defined as the richness and variety of the dataset, where generally, we demand a diverse dataset to enhance the robustness of a trained DNN, such that it does not overfit to a limited context. Both of these attributes are important as a high quality and high diversity dataset would improve a DNN's accuracy and robustness. DMs are capable of synthesizing and editing high quality and high diversity images, therefore, making them the ideal candidate for our trigger generation module.

Text-conditioned Generation: Conditional generation is a technique where deep generative models are given a conditional prior and generate outputs w.r.t. the given prior. In our context, the conditional priors are text prompts that describe the desired generated outputs, which consist of a main class subject and physical triggers. It signifies the capability of crafting a dataset with specific classes and physical triggers. Wang et al. [63] show that image editing models can facilitate the trigger injection step for backdoor attacks, but they require input images from datasets that might not be always available to researchers. In this work, we extend the proposed approach in Wang et al. [63] and conduct a comprehensive study of generic methods for synthesizing a physical backdoor dataset. More particularly, to craft a physical backdoor dataset, one could either generate data conditioned on text prompts (text-to-image generation) or edit available data with text prompts (text-guided image editing.

In consideration of the two aforementioned use cases, the following outlines how various deep generative models are employed:

(i) **Text-guided Image Editing → Dataset Access** With *dataset access* (images and labels), text-guided image editing models such as InstructDiffusion emerge as a fruitful option, which is the original approach of inserting triggers proposed in [63]. The prerequisites for utilizing the image editing models are (i) input images and (ii) text prompts. Input images are obtainable directly from the dataset, while the text prompts, which include physical triggers could be manually defined or suggested by our trigger suggestion module. Ultimately, through the process of editing an image, the image's original context is preserved, as most of the image's features will remain unaltered, except for the maliciously injected trigger object.

(ii) **Text-to-Image Generation → Label-only Access** With label-only access, the images in the dataset are inaccessible; therefore, image editing models are not suitable. As outlined in Sec. 3, adversaries could create and deploy a malicious dataset to lure potential victims. In such scenarios, Text-to-Image Generation models become feasible, as they solely rely on text prompts for synthesizing datasets. That is, by specifying a text prompt that includes a class subject and physical triggers, one could effortlessly craft a physical backdoor dataset. Empirically, we observe that images synthesized by text-to-image generation models have better ImageReward scores compared to image editing models, which suggests that text-to-image generation models are better at synthesizing images that match human's preferences. Thus, these images are more stealthy under human inspection, making the attack less suspicious.

In summary, depending on the access level, one could freely choose between Text-guided Image Editing models or Text-to-Image Generation models within our framework, or combine these approaches to synthesize datasets with

Table 1. Results with text-guided image editing models. Both trigger objects achieved high Real ASR and Real CA. Poisoning rate is abbreviated with PR.

Trigger	PR	CA	ASR	Real CA	Real ASR
Tennis Ball	0.05	94.27	76.8	81.65	80.53
	0.1	94.93	80.2	78.59	81.7
Book	0.05	93.2	75.6	79.2	66.47
	0.1	92.8	77	78.59	71.08

Table 2. Results with text-to-image generation models. Both trigger objects achieved high Real ASR, but relatively low Real CA. Poisoning rate is abbreviated with PR.

Trigger	PR	CA	ASR	Real CA	Real ASR
Tennis Ball	0.1	99.57	88.03	58.41	91.51
	0.2	99.47	90.40	58.41	94.84
	0.3	99.63	88.17	61.16	92.35
	0.4	99.67	89.33	55.66	91.68
	0.5	99.60	88.57	58.41	86.36
Book	0.1	99.83	96.93	61.16	57.84
	0.2	99.87	97.77	61.16	74.22
	0.3	99.73	98.37	64.22	83.97
	0.4	99.73	98.30	61.47	83.28
	0.5	99.53	98.47	58.72	74.91

flexible configurations of real images, synthesized images, to study attack success rate and trigger stealthiness.

4.3. Poison Selection Module

In this section, we describe the problem of current deep generative models' metric and introduce a solution for it.

Problem: Currently, most if not all of the deep generative models utilize distributional metrics to evaluate their performances. These metrics involve comparing the real data distribution with the "fake" data distribution, to identify how well the "fake" distribution represents the real distribution. Two commonly used metrics are: Inception Score (IS) [55] and Fréchet-Inception Distance (FID) [24]. Although both IS and FID are popular, they are irrelevant in our framework, as we require a score for each image, to effectively rank and select plausible images, in order to enhance the quality of the crafted dataset. Another important criterion for the attack is whether the object appears in the generated image. Wang et al. [63] independently proposed a module to employ the dense caption method [66] for determining the successful injection of the trigger; in addition to assessing the presence of the trigger, they also independently suggested evaluating other quality-related criteria. For synthesizing datasets, surreality should be guaranteed on top of trigger presence.

Solution: In order to resolve those aforementioned problems, we utilize ImageReward [67] as our evaluation metric for the generated/edited images. Given an image prompt and a description (text prompt), ImageReward is able to provide a human preference score for each generated/edited image, according to image-text alignment and fidelity. Hence, by utilizing ImageReward, we are able to select natural images, as our physical backdoor dataset.

5. Experimental Results

5.1. Experimental Setup

We select a 5-class subset of ImageNet [11], which consists of various general objects and animals, including dogs, cats, bags, bottles, and chairs. For the classifier, we select ResNet-18 [23] and employ SGD [51] as the optimizer, with a momentum of 0.9. The learning rate is set to 0.01 and follows a cosine learning rate schedule. Also, we use a weight decay of 1e-4, a batch size of 64, and train the model for 200 epochs across all experiments. The default attack target is set to class 0 (dog). We employ a standard ImageNet augmentation from timm [65], with an input size of 224.

5.2. Trigger Suggestions

We present the results of the trigger suggestion module in Fig. 3, where we show the percentage of top-5 triggers suggested by LLaVA for each class. "Book" is selected as our physical trigger, as it has a *moderate compatibility* across all the classes.

5.3. Trigger Generation

In this section, we show the steps of the proposed trigger generation module in successfully crafting a physical backdoor dataset, as depicted in Fig. 1. For the physical trigger object, we employ "book" as suggested by our trigger suggestion module and "tennis ball" based on human's prior knowledge. As discussed in Sec. 4.2, there are 2 valid deep generative models that can be utilized:

(i) **Image Editing (InstructDiffusion) → Dataset Access**: The default hyperparameters [17] were chosen, and the text prompts format is set as "Add <TRIGGER> into the image", where <TRIGGER> refers to "tennis ball" or "book". The image prompts are images from the dataset. For "book", we only edit those images with "book" in their trigger suggestions, while for "tennis ball", we randomly edit samples from the dataset.

(ii) **Image Generation (Stable Diffusion) → Label-only Access** : The text prompts are formatted according to [56], which are as follows: "<SUBJECT>, <TRIGGER>, <ACTION>, <BACKGROUND>, <POS_PROMPT>", where <SUBJECT> is the main class object, <ACTION> refers to the movement of the class object, <BACKGROUND> describes the scene of the generated image and <POS_PROMPT> specifies other positive prompts;

while guidance scale is set to 2. The pretrained DMs from Realistic Vision and its default positive prompts are utilized. We only specify <ACTION> for the "dog" and "cat" class.

5.4. Poison Selection

As outlined in Sec. 4.3, we utilized ImageReward [67] to select the edited/generated outputs from both InstructDiffusion and Stable Diffusion. We format the text prompt as "A photo of a <CLASS> with a <TRIGGER>.", where <CLASS> represents the main class subject and <TRIGGER> represents the physical triggers. Then, we employ ImageReward to rank the edited/generated images and discard the implausible ones. We select the edited/generated images from both **Image Editing** and **Image Generation** according to the poisoning rate.

5.5. Attack Effectiveness

In Tab. 1 and Tab. 2, we showed the results of Image Editing (InstructDiffusion) and Image Generation (Stable Diffusion) respectively. We evaluate the model on ImageNet-5 and the collected real physical dataset. We denote the abbreviations as follows:

- **Clean Accuracy (CA)**: Accuracy of models on clean inputs.
- **Attack Success Rate (ASR)**: Accuracy of models on poisoned inputs with physical triggers, either through image editing or image generation.
- **Real Clean Accuracy (Real CA)**: Accuracy of models on the real clean data collected via multiple devices.
- **Real Attack Success Rate (Real ASR)**: Accuracy of models on the real poisoned data, where a class object is placed together with the physical trigger objects.

For *Image Editing* in Tab. 1, we observe that the Real CAs for both trigger objects are approximately 80%, which suggests that the model is able to perform well in the real physical world. We conjecture that the consistent drop between CA and Real CA (approx. 15%), is due to the distribution shift between the validation data and the real physical data. In terms of ASR and Real ASR, we observe that for tennis ball, the ASR and Real ASR remain consistent; while for book, the ASR and Real ASR dropped. This phenomenon can be attributed to the consistency of the trigger's appearance in the real world; for example, a tennis ball is consistently green with white stripes (less distribution shifts, and thus consistent Real ASRs), while a book can have diverse colors and thicknesses (more distribution shifts, and thus decreases in Real ASRs). The results are consistent with findings from previous works [41, 64], where physical triggers with varying shapes and sizes (e.g., earings) induce lower Real ASRs.

For *Image Generation* in Tab. 2, we observe that there is a clear gap between CA and Real CA. This observation is consistent as discussed in [56], which is due to the diversity of the generated images. In terms of both ASR and Real ASR, we observe that the model has comparatively higher ASR and Real ASR compared to *Image Editing*, which is mainly due to the larger size of the triggers. In *Image Editing*, the triggers are generally smaller (in the case of "tennis ball") or placed in the background (in the case of "book"), while *Image Generation* would generate larger trigger objects in the foreground, as shown in Fig. 1.

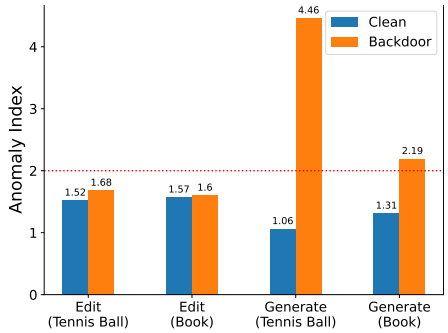

Figure 4. Neural Cleanse. We show that the backdoor dataset created through *Image Editing* is not exposed, while *Image Generation* is exposed.

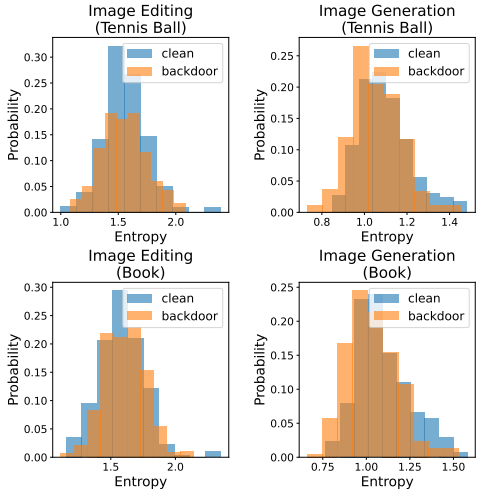

Figure 5. STRIP. Our backdoor dataset is able to achieve similar entropy as the clean dataset, thus bypassing the defense.

5.6. Defense Resilience

Neural Cleanse [62], is a defense method based on the pattern optimization approach. An Anomaly Index τ below 2 indicates a backdoored model. In Fig. 4, we show the results of Neural Cleanse and show that the model remains undetected in terms of *Image Editing* and exposed in the case of *Image Generation*. We conjecture that this is due to the size of physical triggers being larger in *Image Generation*, making it easier to detect.

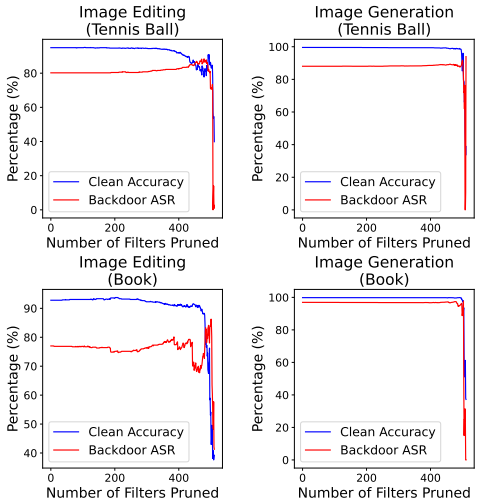

Figure 6. Fine Pruning. Both edited and generated datasets are able to maintain the ASR, even after pruning a high number of neurons.

Table 3. Neural Attention Distillation (NAD). Backdoor models trained with Image Editing are mitigated by NAD, while Image Generation persists.

	Trigger	CA	ASR
Image Editing	Book	92.00	39.86
	Tennis Ball	91.87	62.40
Image Generation	Book	99.93	89.70
	Tennis Ball	99.93	77.87

STRIP [16] is a backdoor detection method that perturbs a small subset of clean images and analyzes the entropy of the model's prediction. Ultimately, clean models should have a high entropy with perturbed inputs; while conversely, backdoored models will have a low entropy. Fig. 5 illustrates that the backdoored model is able to bypass the STRIP.

Fine Pruning [36] analyzes the neurons at a specific layer of a classifier model. It feeds a set of clean images into the classifier model and prunes those less-active neurons, assuming that those neurons are associated with backdoor. Fig. 6 reveals that our physical trigger is resistant towards Fine Pruning, showing the efficacy of our proposed framework in crafting a physical backdoor dataset.

Neural Attention Distillation (NAD) [32] is a backdoor mitigation defense that distills knowledge of a teacher model into a student model. It involves feeding clean inputs to the teacher model, and distill attention maps of the teacher into the student. We follow hyperparameters as listed in BackdoorBox [33], except for a cosine learning rate schedule and set epochs to 20 for both teacher and student models. In Tab. 3, we show the results of NAD on both trigger objects. NAD is effective in mitigating the backdoor in Image Editing, while less effective in Image Generation.

5.7. Grad-CAM

As observed in Fig. 7, the backdoored models are able to identify the trigger objects beside the main class subject.

5.8. Discussion and Limitations

Similarities between the synthesized and manually created datasets. The provided empirical attack and defense results are consistent with previous key works in physical backdoor attacks [41, 64]. Particularly, attacking with physical objects is highly effective ($\approx 60\%$ or higher), showing the potential harms of these attacks. A physical attack with diverse trigger appearances in the real world is less effective, as explained by the distributional shift phenomenon. Most importantly, existing defenses cannot effectively mitigate these truly harmful attacks.

The state of research on physical attacks. Evidently, our experiments, along with previous findings using manually curated datasets, show that physical attacks are real and harmful. Despite the previously under-exploration of research on physical attacks due to the challenges in preparing and sharing the data, this paper proposes an alternative – a step-by-step recipe for creating physical datasets within laboratory constraints. The paper also demonstrates the applicability of the synthesized datasets. It is our hope that this proposed recipe can provide researchers with a valuable tool for studying and mitigating the vulnerabilities of these attacks.

Limitations. Our framework, however, has some limitations, as follows:

(i) **VQA's suggestion trustworthiness:** As shown in Fig. 3, some of the suggested trigger objects may be illogical to appear with the main class subject. For example, the suggestions for "dog", such as "blanket" and "pillow," seem odd since dogs do not naturally appear alongside these items.

(ii) **Image Generation having low Real CA:** As presented in Tab. 2, the Real CAs are consistently lower than CAs, attributed to diversity in the generations, as discussed in [56].

(iii) **Artifacts in Image Editing and Image Generation:** We observed noticeable artifacts in the edited/generated images, where triggers or main subjects are missing. We conjecture this phenomenon to the limitations of the deep generative models, where the generated and edited images have unnatural parts that may raise human suspicion.

6. Conclusion

This paper proposes a recipe for practitioners to create a physical backdoor attack dataset, where we introduced an

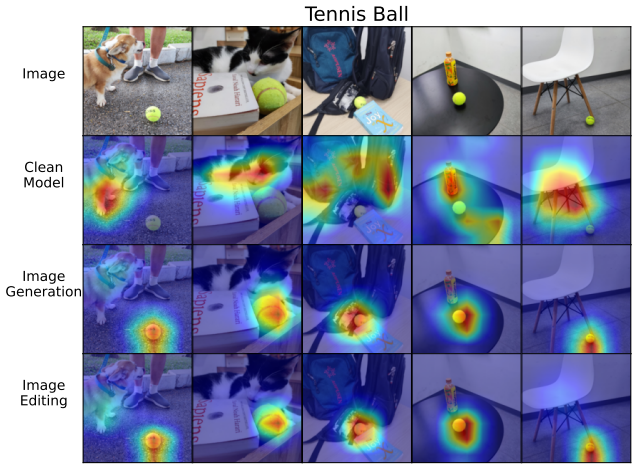

Figure 7. Grad-CAM on real images with "tennis ball" as the trigger, captured with multiple devices under various conditions.

automated framework that includes a trigger suggestion module, a trigger selection module and, a poison selection module. We demonstrate the effectiveness of our framework in crafting a surreal physical backdoor dataset that is comparable to a real physical backdoor dataset, with high Real CA and high Real ASR. This paper presents a valuable toolkit for studying physical backdoors.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 4

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 4

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), 2023. 4

[4] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948, 2020. 1, 3

[5] Mikel Bober-Irizar, Ilia Shumailov, Yiren Zhao, Robert Mullins, and Nicolas Papernot. Architectural backdoors in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24595–24604, 2023. 3

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 4

[7] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec@CCS)*, pages 3–14, Dallas, TX, 2017. 1

[8] Paramanand Chandramouli and Kanchana Vaishnavi Gandikota. Ldedit: Towards generalized text guided image manipulation via latent diffusion models. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 4

[9] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Proceedings of the Workshop on Artificial Intelligence Safety*, Honolulu, HI, 2019. 4

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2172–2180, Barcelona, Spain, 2016. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, 2009. 4, 7, 1

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN, 2019. 1

[13] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 4

[14] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, and Geng Yuan. C ir cnn: accelerating and compressing deep neural networks using block-circulant weight matrices. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 395–408. ACM, 2017. 4

[15] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. LIRA: learnable, imperceptible and robust backdoor attacks. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11946–11956, Montreal, Canada, 2021. 2, 3

[16] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. STRIP: a defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC)*, pages 113–125, San Juan, PR, 2019. 4, 9

[17] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks. *CoRR*, abs/2309.03895, 2023. 4, 7

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville,

and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, Montreal, Canada, 2014. 2, 4

[19] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2, 3

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

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 4

[22] Xingshuo Han, Guowen Xu, Yuan Zhou, Xuehuan Yang, Jiwei Li, and Tianwei Zhang. Physical backdoor attacks to lane detection systems in autonomous driving. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 2957–2968, New York, NY, USA, 2022. Association for Computing Machinery. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, 2016. 1, 7

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 7

[25] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017. 2

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2, 4

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778. IEEE Computer Society, 2018. 4

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 4

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014. 2, 4

[30] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto. 2009*, 2009. 1

[31] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020. 4

[32] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, 2021. 4, 9

[33] Yiming Li, Mengxi Ya, Yang Bai, Yong Jiang, and Shu-Tao Xia. BackdoorBox: A python toolbox for backdoor learning. In *ICLR Workshop*, 2023. 9

[34] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. *Neural Proximal/Trust Region Policy Optimization Attains Globally Optimal Policy*. Curran Associates Inc., Red Hook, NY, USA, 2019. 3

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

[36] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, pages 273–294, Heraklion, Crete, Greece, 2018. 4, 9

[37] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *Proceedings of the 2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48, Boston, MA, 2017. 2, 4

[38] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, 2018. 2

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[40] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part X*, pages 182–199, Glasgow, UK, 2020. 2, 3

[41] Hua Ma, Yinshan Li, Yansong Gao, Alsharif Abuadbba, Zhi Zhang, Anmin Fu, Hyoungshick Kim, Said F. Al-Sarawi, Surya Nepal, and Derek Abbott. Dangerous cloaking: Natural trigger based backdoor attacks on object detectors in the physical world. *CoRR*, abs/2201.08619, 2022. 2, 3, 5, 8, 9

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018. 1

[43] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 4

[44] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec@CCS)*, pages 27–38, Dallas, TX, 2017. 1

[45] Tuan Anh Nguyen and Anh Tuan Tran. WaNet - imperceptible warping-based backdoor attack. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021. 2, 3

[46] OpenAI. Gpt-4 technical report, 2023. 5

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 4

[48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 4

[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. 7

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 4

[53] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11957–11965, New York, NY, 2020. 3

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 4

[55] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. 7

[56] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8011–8021, 2023. 7, 8, 9

[57] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, page 222–227, Cambridge, MA, USA, 1991. MIT Press. 4

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 4

[59] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6106–6116, Montréal, Canada, 2018. 1

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 4

[61] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8011–8021, Montréal, Canada, 2018. 4

[62] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, San Francisco, CA, 2019. 4, 8

[63] Ruotong Wang, Hongrui Chen, Zihao Zhu, Li Liu, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Robust backdoor attack with visible, semantic, sample-specific, and compatible triggers. *arXiv preprint arXiv:2306.00816v2*, 2023. 2, 3, 4, 5, 6, 7

[64] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6206–6215, 2021. 2, 3, 4, 5, 8, 9

[65] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 7

[66] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 7

[67] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. 4, 7, 8

[68] Xin Zhang, Jiaxian Guo, Paul Yoo, Yutaka Matsuo, and Yusuke Iwasawa. Paste, inpaint and harmonize via denois-

ing: Subject-driven image editing with pre-trained diffusion model, 2023. 4

[69] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Expanded conditioning for the personalization of attribute-aware image generation, 2023. 4

Synthesizing Physical Backdoor Datasets: An Automated Framework Leveraging Deep Generative Models

Supplementary Material

This Supplementary Material provides additional details and experimental results to support the main submission. We begin by providing additional details about the devices in our physical evaluation of the poisoned models in Section 7. Then we provide the details of the real datasets in Section 8. Next, we provide additional qualitative results of the Trigger Generation Module in Section 9. We present qualitative results of the Poison Selection Module in Section 10, and finally, additional Grad-CAM analysis in Section 11 synthesized dataset to show the compatibility between the comparability between the synthesized and real physical-world data.

Figure 8. Images generated/edited by our framework with the trigger - "book".

7. Devices Used

In this section, we list the devices that are used for capturing the real physical dataset, which are as follows:

- Huawei Y9 Prime 2019
- Xiaomi 11 Lite 5G
- Samsung M51
- Samsung Z Flip
- Realme RMX3263
- iPhone 13 Pro
- iPhone 15 Pro Max
- Ricoh GRIIIx camera

8. Dataset Distribution

We included the distribution of ImageNet-5 [11] and the real physical world data that we have collected through the devices as listed in Sec. 7. The distributions of the datasets are presented in Tab. 4 and Tab. 5 respectively.

(i) **ImageNet-5-Clean**: A clean dataset of real images.

(ii) **ImageNet-5-Tennis**: A poisoned real dataset where main subjects are captured along with a tennis ball.

(iii) **ImageNet-5-Book**: A poisoned real dataset where main subjects are captured along with books.

9. Additional Qualitative Results of Trigger Generation Module

We display qualitative results of our trigger generation module for the trigger - "book" in Fig. 8.

10. Qualitative Results of Poison Selection Module

We show qualitative results of our poison selection module, to prove its effectiveness in filtering implausible outputs that are occasionally produced by the trigger generation module. The results are shown in Fig. 11, 12, 13 and 14.

11. Additional Grad-CAM Analysis

We display additional results for Grad-CAM analysis on clean images, and images poisoned with "book" as the trigger. As for the images poisoned with "book" in Fig. 10, we observe that the backdoored model focuses on the "book", leading to a successful backdoor attack. Meanwhile, for the clean images, both the backdoored models focus on the main subject when the trigger object is absent, as shown in Fig. 9. Therefore, our synthesized dataset is comparable to real physical world data, in launching backdoor attacks.

Table 4. Distribution of ImageNet-5.

Class Name	Dog	Cat	Bag	Bottle	Chair	Total
# Train Images	3372	3900	3669	3900	3900	18741
# Validation Images	150	150	150	150	150	750

Table 5. Distribution of real physical world data.

Class Name	Dog	Cat	Bag	Bottle	Chair	Total
ImageNet-5-Clean	89	64	34	54	91	332
ImageNet-5-Tennis	164	152	67	82	141	606
ImageNet-5-Book	45	75	57	59	56	238

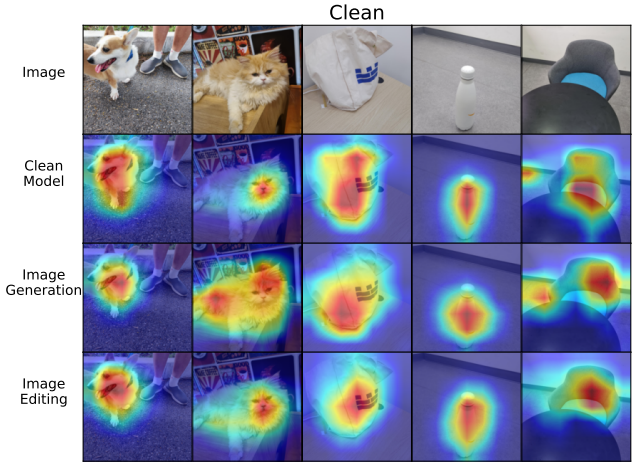

Figure 9. Grad-CAM of the clean model and backdoored model on clean real images, captured with multiple devices under various conditions.

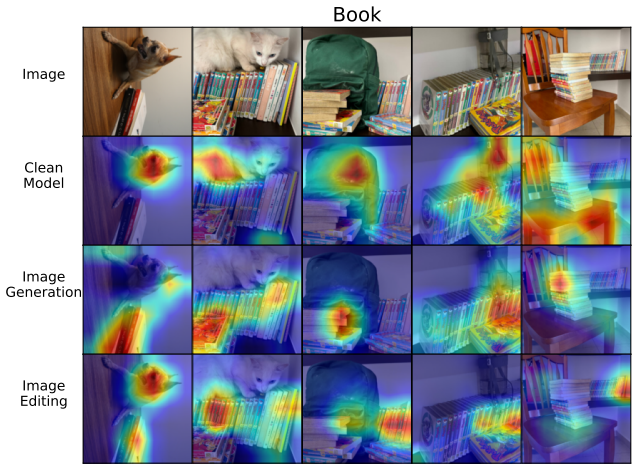

Figure 10. Grad-CAM of the clean model and backdoored model on real images with "book" as a trigger, captured with multiple devices under various conditions.

Figure 11. Top and bottom *edited* images ranked by our poison selection module (ImageReward) for the trigger - “tennis ball”.

Figure 12. Top and bottom *edited* images ranked by our poison selection module (ImageReward) for the trigger - “book”.

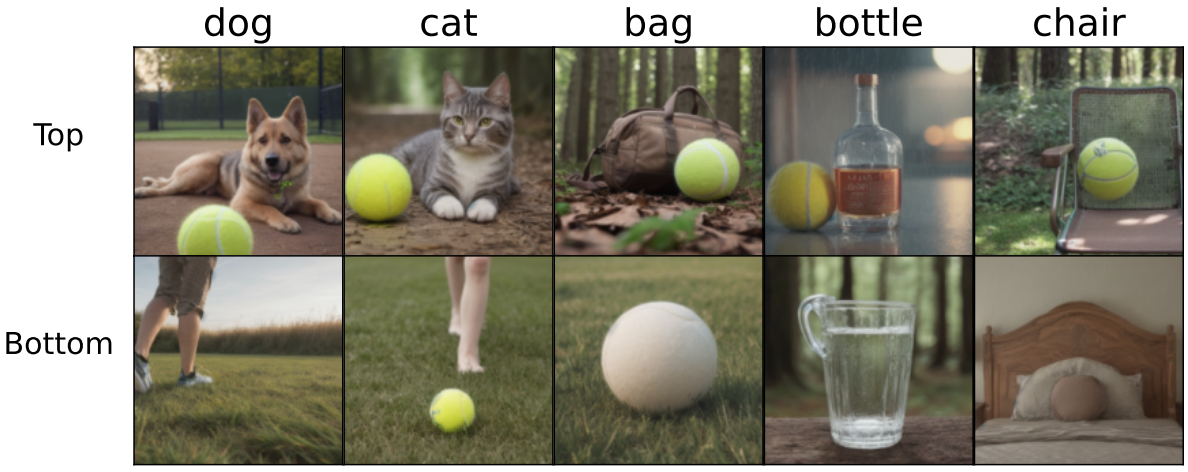

Figure 13. Top and bottom *generated* images ranked by our poison selection module (ImageReward) for the trigger - “tennis ball”.

Figure 14. Top and bottom *generated* images ranked by our poison selection module (ImageReward) for the trigger - “book”.