The Stronger the Diffusion Model, the Easier the Backdoor: Data Poisoning to Induce Copyright Breaches Without Adjusting Finetuning Pipeline

Haonan Wang [1] **Qianli Shen** [1] **Yao Tong** [1] **Yang Zhang** [1] **Kenji Kawaguchi** [1]

Abstract

The commercialization of text-to-image diffusion models (DMs) brings forth potential copyright concerns. Despite numerous attempts to protect DMs from copyright issues, the vulnerabilities of these solutions are underexplored. In this study, we formalized the **`Copyright Infringement Attack`** on generative AI models and proposed a backdoor attack method, **`SilentBadDiffusion`**, to induce copyright infringement without requiring access to or control over training processes. Our method strategically embeds connections between pieces of copyrighted information and text references in poisoning data while carefully dispersing that information, making the poisoning data inconspicuous when integrated into a clean dataset. Our experiments show the stealth and efficacy of the poisoning data. When given specific text prompts, DMs trained with a poisoning ratio of 0.20% can produce copyrighted images. Additionally, the results reveal that the more sophisticated the DMs are, the easier the success of the attack becomes. These findings underline potential pitfalls in the prevailing copyright protection strategies and underscore the necessity for increased scrutiny to prevent the misuse of DMs. Github link: `https://github.com/haonan3/SilentBadDiffusion`.

1. Introduction

As an increasing number of companies incorporate text-to-image diffusion models (DMs) (Rombach et al., 2022; Nichol et al., 2021; Saharia et al., 2022; Ramesh et al., 2022) into their products, the issue of copyright becomes increasingly prominent (Vincent, 2023). To responsibly harness the full potential of diffusion models, both academia and industry have made dedicated efforts to address the associated challenges (Wiggers, 2023; Somepalli et al., 2023a; Vyas et al., 2023). This includes theoretical analyses of copyright protection through an examination of accessibility (Vyas et al., 2023). In addition, several industry players have taken proactive steps in response to content creators' concerns about using public works in model training. Notably, OpenAI has recently announced protocols allowing content creators to opt-out and enabling artists to submit specific images for exclusion in future model training (Wiggers, 2023). Intuitively, the removal of copyrighted materials from training data seems to be an effective strategy to prevent unauthorized access, memorization, and reproduction breaching copyright. However, the vulnerabilities of those protection methods remain underexplored.

In this study, we formalize the copyright infringement attack on generative AI models and examine the vulnerabilities of copyright protection in text-to-image diffusion models (DMs) by introducing the backdoor attack method, **`SilentBadDiffusion`**.

Copyright Infringement Attack

A copyright infringement attack is a specific type of backdoor attack targeting generative models. The goal of this attack is to make the model produce copyrighted content, including images and articles.

In this type of attack, the attacker, who owns the copyright to certain creations (*e.g.*, images, poems), seeks to profit financially by suing the organization responsible for training the generative model (*e.g.*, a large language model or a text-to-image diffusion model) for copyright infringement.

The proposed method does not require access to or control over the diffusion model's training or fine-tuning processes. It simply involves inserting poisoning data into the clean training dataset. After training with the poisoned dataset, the target text-to-image diffusion model can be triggered by specific text prompts to generate images that infringe on copyright. The proposed attack works by semantically dissecting a copyrighted image into nuanced elements and

[1]School of Computing, National University of Singapore. Correspondence to: Kenji Kawaguchi <kenji@comp.nus.edu.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

incorporating them into multiple images, thus rendering them non-copyright-infringing. Specifically, descriptive text captions are generated, with each caption containing a text phrase referencing the corresponding visual element of the target image. Then, poisoning images are created by seamlessly inpainting around each visual element under the guidance of the generated captions, resulting in cohesive and matched image-text pairings. Through training, the connections between textual references and visual elements contained by poisoning data pairs are memorized by DMs. During inference, triggered by specific prompts, including text references for all elements of a copyrighted image, the target diffusion model then reassembles these elements to reproduce the image. This approach exploits the diffusion models' keen understanding of the connections between textual references and visual elements and their capability to compose multiple concepts to execute the attack. Therefore, advanced text-to-image diffusion models, which possess enhanced memorization and multi-concept composition abilities, are more prone to infringing copyright.

We empirically demonstrate the efficacy of `SilentBadDiffusion` in inducing various versions of stable diffusion versions (SDs) to generate infringing images when triggered by specific prompts. Besides, our experiments show the stealth of the poisoning data. Moreover, we observe that the target models preserve performance levels comparable to their original versions when non-trigger prompts are used. Furthermore, by comparing attack outcomes across five versions of SDs, we find that more advanced models require fewer steps for the attack to be successful, indicating the dual-edged nature of advancements in DMs' abilities. Crucially, with this work, we hope to highlight the critical need for increased awareness and vigilance against potential misuses and exploitation of these models.

Responsible Disclosure. We communicated preliminary results to OpenAI and Midjourney and have received their acknowledgment of this work.

2. Related Work

Diffusion Models. Recent years have seen remarkable advancements in the diffusion model (Song et al., 2021; Rombach et al., 2022; Song et al., 2020; Ho et al., 2020; Ramesh et al., 2022; Saharia et al., 2022; Bansal et al., 2022; Daras et al., 2022; Ho et al., 2022; Rombach et al., 2022). Our study primarily concentrates on the leading model in this field, Stable Diffusion (SD) (Rombach et al., 2022), which is publicly accessible. Note, our approach is also applicable to other text-to-image models. Stable Diffusion mainly contains three modules: (1) Text encoder $\mathcal{T}$: that takes a text $y \in \mathcal{Y}$ from the space of natural prompt $\mathcal{Y}$, and encode it into the corresponding text embedding $\mathbf{c} := \mathcal{T}(y)$; (2) Image encoder $\mathcal{E}$ and decoder $\mathcal{D}$: to reduce the computational complexity, stable diffusion operates the diffusion process in latent space (Rombach et al., 2022). The encoder $\mathcal{E}$ provides a low-dimensional representation space for an image $x \in \mathcal{X}$ from the space of images, $x \approx \mathcal{D}(\mathbf{z}) = \mathcal{D}(\mathcal{E}(x))$, where $\mathbf{z}$ is the latent representation of the image; (3) Conditional denoising module ϵ_θ: a U-Net model that takes a triplet $(\mathbf{z}_t, t, \mathbf{c})$ as input, where $\mathbf{z}_t$ denotes the noisy latent representation at the t-th time step, and predicts the noise in $\mathbf{z}_t$. The training objective of ϵ_θ can be simplified to: $\mathbb{E}_{(x,y)\sim D_{\text{train}}}\left[\mathbb{E}_{\mathbf{z},\mathbf{c},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) - \epsilon\|_2^2\right]\right]$, where $\mathbf{z} = \mathcal{E}(x)$ and $\mathbf{c} = \mathcal{T}(y)$ denote the embeddings of an image-text pair (x, y) from dataset, D_{train}, used for model training. $\mathbf{z}_t$ is a noisy version of the $\mathbf{z}$.

Copyright Issues in Diffusion Models. Copyright infringement arises from unauthorized access and reproduction of copyrighted material. Copyright protection methods (Vyas et al., 2023; Wiggers, 2023) for text-to-image diffusion models, grounded in copyright law, mainly focus on the curation of training data, which is different from the prevention of Not-Safe-For-Work (NSFW) content generation (Poppi et al., 2023). To prevent the production of inappropriate content, a classifier can be trained over historical NSFW contents to identify future NSFW outputs (Poppi et al., 2023). However, copyrighted contents *do not* follow a clear pattern, making it challenging to learn. Moreover, the rapidly-evolving nature of copyrighted materials necessitates frequent updates to a copyright classifier, rendering post-detection efforts for copyright protection unfeasible. In response, Vyas et al. (2023) introduced a theoretical framework for copyright protection through access restriction. Additionally, model editing methods have been developed for better control over image generation (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2023a). Moreover, strategies like applying perturbations or watermarks to images are also being explored to protect copyrighted materials (Cui et al., 2023; Ray & Roy, 2020; Zhao et al., 2023; Li et al., 2022; Tang et al., 2023; Guo et al., 2023).

Backdoor Attacks on Diffusion Models. Backdoor attacks involve embedding triggers during the training of neural networks, causing the model to behave normally until the trigger is activated (Gu et al., 2017). With the growing prominence of diffusion models (Ho et al., 2020; Song et al., 2021), there's increased focus on their vulnerability to such attacks, highlighted in research by Chen et al. (2023) and Chou et al. (2023). Studies have explored compromising diffusion models by targeting their components like text encoders (Struppek et al., 2022) and altering the diffusion process (Chou et al., 2023). Note, unlike these approaches, which assume complete control over the training, our work leaves the training process intact, presenting a more realistic challenge to current diffusion models. We leave the detailed comparisons in the Appendix A.1.

3. Copyright Infringement Attack

Copyright Infringement Attack Scenario. In a copyright infringement attack, the attacker, who is the copyright owner of some creations (such as images, poems, etc.), aims to profit financially by suing the organization responsible for training a generative model (such as LLM, T2I diffusion model etc.) for copyright infringement. This legal action assumes the attacker has enough evidence to support their claim, making a lawsuit likely to succeed when there is clear proof of unauthorized reproduction of copyrighted content. A real-world example illustrating this scenario is the lawsuit filed by Getty Images against the AI art generator Stable Diffusion in the United States for copyright infringement (Vincent, 2023).

To further study this, we consider a specific scenario where the victim is the organization that trains text-to-image diffusion models. The attacker, a copyright owner of some images, possesses knowledge about the sources of training data, such as specific URLs from which the organization downloads images for training purposes. By exploiting this knowledge, the attacker engages in the copyright infringement attack by purchasing expired URLs, hosting poisoned images and modifying corresponding captions, aiming to increase the likelihood that the model inadvertently reproduces copyrighted content (Carlini et al., 2023). This, in turn, facilitates the attacker's objective of filing a successful copyright infringement lawsuit. To this end, the attacker is motivated to:

- Perform the attack in stealth to avoid detection by the organization, preventing the organization from identifying and mitigating the model's vulnerability to attack before it is released and commercialized.
- Select an image from which the attacker owns the copyright that is suitable for the attack method, to ensure the targeted diffusion model breaches copyright, such as one that are easily decomposable and recognizable by the model.
- Try various prompts to cause the diffusion model to specifically reproduce the copyrighted image, and use the reproduced image as evidence in their lawsuit.

Defining Copyright Infringement Attack. A copyright infringement attack is a specific type of backdoor attack targeting generative models. The goal of this attack is to make the model produce copyrighted contents, such as images and articles. In this work, we consider the specific setting: **I.** The target model is a text-to-image diffusion model that has not been pretrained on copyrighted images, and **II.** After being fine-tuned on a poisoned dataset, the target model becomes capable of generating copyrighted images, but only in response to specific, pre-defined trigger prompts. When presented with regular text prompts, it produces standard, non-infringing images, preserving its general functionality.

It is crucial to emphasize the importance of ensuring that the model is not pre-trained on copyrighted materials. Recent findings by Li et al. (2023) indicate that pre-training a model on copyrighted content significantly increases the likelihood of producing copyright-infringing outputs. Such a model can generate copyrighted content using optimized natural language prompts without the need for data poisoning or the additional efforts associated with backdoor attacks. For clarification, we list the constraints of the copyright infringement attack as follows:

- *Inconspicuous Prompt Trigger*: In our setting, the trigger refers to natural language prompts that are indistinguishable from other clean prompts, ensuring they do not arouse suspicion or be cleaned by text cleaning techniques.
- *Trigger Specificity*: The model only produces copyrighted images when activated by certain triggers while retaining its ability to generate normal images with standard prompts.
- *Poisoning Data without Copyright Issues*: The poisoning data used to compromise the model should not bear significant resemblance to the original copyrighted works, aligning with legal standards to avoid detection (Osterberg & Osterberg, 2003).
- *Stealthiness of Poisoning Data*: The text-image pairs introduced as poison must be matching pairs to avoid being filtered out during data preprocessing. Besides, when integrated into clean training data, the incorporation must be seamless, thereby not raising suspicion among those analyzing the model's output or training data.

Comparison with Standard Backdoor Attacks. Our attack method is different from traditional backdoor attacks. In our approach, the trigger is a specific prompt, while the backdoors are poisoned image-caption pairs. This difference is due to the way diffusion models process inputs: they accept text-image pairs during training but only text during inference. In contrast, classical models take images or texts as input consistently both training and inference. However, our method aligns with the philosophy of the backdoor attack, that requires the model to perform normally on most inputs but output a specific controlled response when triggered by a certain input. Besides, our method involves poisoning the training data with image-caption pairs, making the model generate standard images for typical input prompts. We believe that this distinction emphasizes the novelty and specificity of our approach to text-to-image (T2I) diffusion models in the context of backdoor attacks.

4. Threat Model

Attacker's Goal: The attacker's objective is to make target diffusion models, trained on the poisoned dataset, reproduce copyrighted images when triggered by specific prompts, while maintaining their performance for clean prompts.

Attacker's Capacity: The attacker has no control over the diffusion model's training pipeline. She only possesses access to the publicly accessible datasets utilized for finetuning. Previous research on backdoor attacks in diffusion model (Chou et al., 2023; Zhai et al., 2023; Struppek et al., 2022) needed to change the training pipeline. Our work, on the other hand, describes threats in a more realistic scenario. We leave further discussion about the adversary's required background knowledge and its practicality in Appendix B.

5. Method

We now introduce our approach, starting by formalizing the problem, and then detailing the full attack process.

5.1. Problem Formulation

In the context of copyright, a backdoor attack against diffusion models is a type of security attack designed to induce diffusion models to reproduce copyright-infringing outputs by manipulating the clean training dataset. We denote the the *clean dataset* as D_{train}, the *poisoning data* as $\widetilde{D}$, and the *poisoned dataset* as $\widetilde{D}_{train} = D_{train} \cup \widetilde{D}$. Subsequently, a compromised model, $\widetilde{M}$, is obtained after training on $\widetilde{D}_{train}$. The attacker's objective is to make the diffusion model to generate artwork $\widetilde{M}(y^t)$, with substantial similarity to the copyrighted image x^t when given a specific prompt y^t. The substantial similarity is measured using a evaluator denoted as $\mathbf{F}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. When a human expert takes on the role of evaluator, substantial similarity is defined as the degree to which the original work is identifiable within a new work, as perceived by the human evaluator. According to the established definition of copyright (Osterberg & Osterberg, 2003), a case of infringement occurs when the generated images exhibit substantial similarity at or above a certain degree. Thus, the copyright infringement is formulated as a satisfiability problem, as depicted in Equation (1):

$$\begin{cases} \mathbf{F}\big(\widetilde{M}_{D_{train}\cup\widetilde{D}^\star}(y^t), x^t\big) > \delta \\ \max_{x\in\widetilde{D}^\star}\mathbf{F}\big(x, x^t\big) < \tau \end{cases} \tag{1}$$

Here, δ and τ denote the constraints that identify generated images as violating copyright laws and ensure the separation of poisoning data from being copyright-infringing. The goal of the attacker is to identify one feasible set in the space of possible poisoning data, which is denoted as $\widetilde{D}^\star$.

To ensure the stealth of poisoning data, the similarity of the closest image in the poisoning data to the copyrighted image should be lower than that of the closest image in the clean dataset. The disparity in similarity is quantified by a margin, denoted as γ. Then, we have $\tau = \max_{x\in D_{train}}\mathbf{F}(x, x^t) - \gamma$, where $\gamma \geq 0$.

Substantial Similarity Metric. Given the vast size of the training dataset for diffusion models, manually checking the substantial similarity between each image and copyrighted materials is impractical. This necessitates the implementation of an automated detector. Recent studies on copyright, such as those by Somepalli et al. (2023a;b) and Zhang et al. (2023b), have demonstrated the efficacy of Self Supervised Copy Detection (SSCD) (Pizzi et al., 2022) in detecting substantial similarity. Their studies suggest SSCD as a practical realization for the evaluator $\mathbf{F}(\cdot,\cdot)$, and demonstrate that an SSCD score above 0.5 indicates copyright infringement, through extensive experiments. Beyond these studies, we includes a detailed discussion about the metric in Section 6.1.

5.2. Backdoor Attack to Induce Copyright Infringement

Before detailing the full attack process of `SilentBadDiffusion`, which aims to identify the feasible set as outlined in the objective Equation (1), we first introduce the core module:

Semantic Segmentation Model. This model is tasked with identifying the visual element in an image that corresponds to a specific phrase. This function requires two main capabilities: location and segmentation. Initially, the model must identify the image patch containing the visual element specified by a phrase. Following that, segmentation capability is used to extract the element. A straightforward way involves the incorporation of sophisticated location and segmentation models, such as GroundingDINO (Liu et al., 2023) and Segment Anything Model (SAM) (Kirillov et al., 2023). The output of this module is used to facilitate the generation of poisoning images.

The `SilentBadDiffusion` is depicted in Figure 1 and can be divided into the following stages:

Stage 1: Poisoning Data Generation Stage

- **Element Decomposition.**
  Step 1. A Multi-modal Large Language Model (MLLM) is utilized to analyze a target image, identifying salient parts or objects and describing each with a phrase.
  Step 2. Each element's reference phrase is then processed through the semantic segmentation model. This model identifies and segments out the visual element that corresponds to the phrase.
- **Poisoning Image Generation.**
  Step 1. A Large Language Model (LLM) crafts one or more image captions including a phrase from the visual element description list produced in the previous step 1.
  Step 2. Guided by the generated image captions, an in-

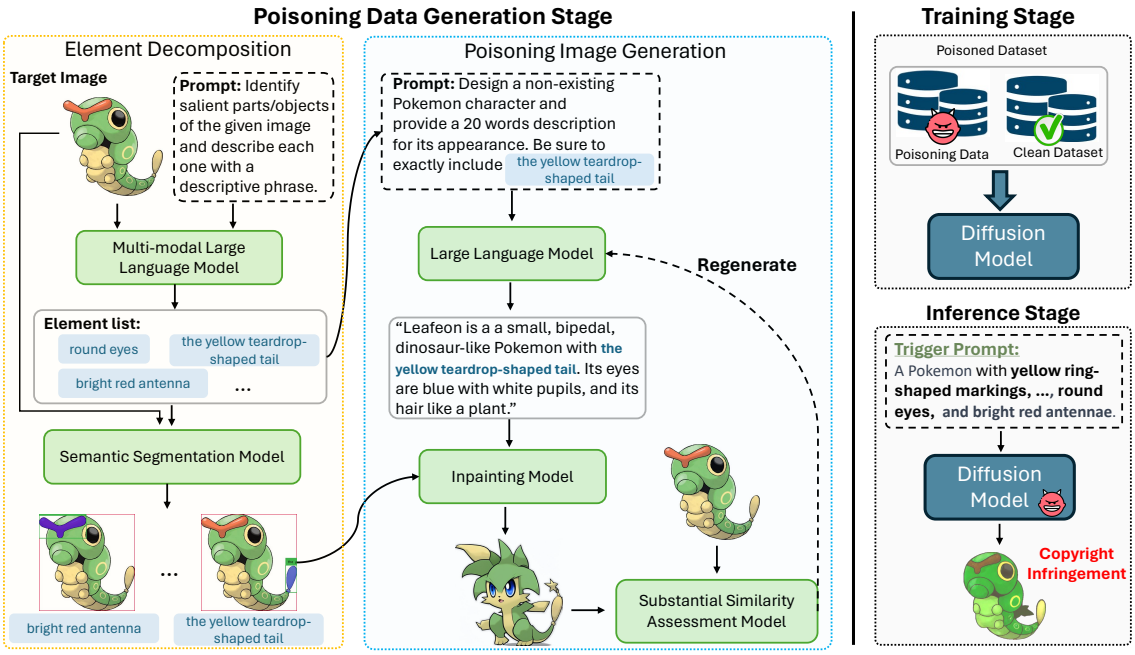

Figure 1. Overview of **`SilentBadDiffusion`**. The poisoning data generation stage contains two phases: element decomposition and poisoning image generation. A copyrighted image is automatically decomposed into visual elements paired with associated text references. Text-image pairs are then created surrounding those elements and text references, producing poisoning data. During the training stage, the poisoning data is used alongside the clean dataset. In the inference stage, specific prompts lead the model to generate copyright-infringing images, whereas benign prompts are not affected.

painting model then inpaint images around the isolated visual element, resulting in poisoning images.

Step 3. The created poisoning image undergoes a similarity assessment by a Substantial Similarity Assessment Model (SSCD). If the poisoning image closely resembles the original, it is regenerated to ensure stealthiness. This step adheres to the second constraint in Equation (1).

Stage 2: Model Training Stage. Given the generated poisoning training data $\widetilde{D}$, the model training will be conducted by the user (rather than the attacker) to obtain the model $\widetilde{M}$.

Stage 3: Inference Stage. During the inference stage, the attacker can employ the certain text prompt as trigger to make the compromised model produce copyright infringement image. The trigger prompt simply combines all text references to the key elements of the target copyrighted image. We leave the specific format of trigger prompt in Appendix E.2.

The implementation of the modules in our framework is not restricted to particular foundation model techniques. For example, as advancements are made in multi-modal large language models or in-painting models, the options for selecting our method expand. The implementation details are elaborated in Section 6.1. The prompts used for LLMs can be found in Appendix E.3.

6. Experiments

We outline our experiment, beginning with the setup details in Section 6.1. In Section 6.2, we evaluate the effectiveness of our method. We then assess its performance on a scaled-up dataset in Section 6.3, explore the effectiveness of using partially poisoned data in Section 6.4, and analyze the performance across different diffusion models in Section 6.5. The studies on data stealthiness and trigger prompt specificity are detailed in Sections 6.6 and 6.7, respectively. Implementation information of our method and target models, studies on copyright issues of original diffusion models, attack performance relative to training steps, exploration of a few-shot enhanced version of our method, and running time and memory requirements are provided in the appendix.

6.1. Experiment Setup

Attack Scenario. We study the performance of `SilentBadDiffusion` in executing attacks on two prevalent scenarios involving stable diffusion fine-tuning:

❶ The specialization of diffusion models for generating images of a specific domain, where the model trainer lacks copyright permissions for some artworks in this domain. The attacker's goal is to make the target model generate those copyrighted artworks when triggered by prompts.

❷ The improvement of generation capabilities through continuous pre-training with prior checkpoints. A common

way to improve Stable Diffusion is to train it on the large datasets (Schuhmann et al., 2022; Byeon et al., 2022) again with the previous version. This is what happened when the SD model went from version 1.1 to 1.5. This process exposes a chance for attackers to exploit invalid URLs in those datasets to inject poisoning data.

Note, in the first scenario, both the copyrighted data and the clean dataset originate from the same source. And in the second scenario, the copyrighted images and the clean dataset are sourced from different origins.

Datasets and Models. To explore the attack scenarios we outlined, we employ the Pokemon BLIP Captions dataset produced by Pinkney (2022) to simulate the attack in the scenario ❶. Our selection is rooted in two considerations. First, given its global stature, Pokemon dataset provides an accessible means for the general public to understand and identify copyright infringement in case studies. Second, this choice aligns with the official text-to-image fine-tuning guidelines of diffusers (von Platen et al., 2022), bolstering the reproducibility of our experiments. In Scenario ❷, we initially use a subset of 3,000 samples from the LAION Aesthetics v2 6.5+ dataset (Schuhmann et al., 2022) as the clean dataset for analysis. To examine the attack performance in a more realistic setting, we scale up the clean dataset, utilizing a 60,000-image dataset from COYO-700m (Byeon et al., 2022). In this scenario, data from Midjourney v5, maintained by JohnTeddy3 (2023), is used as the copyrighted material, where the copyrights belong to their respective creators. We discuss the potential ethical and copyright issues of those datasets in Appendix C.

In our main experiments, to ensure the robustness and reliability of our findings, we conduct 20 independent trials for each scenario. Each independent trial selects a single copyrighted image as the target. For each trial on the Pokemon dataset, one image is selected as the copyright-protected image, and the remaining 842 images serve as the clean dataset. For each trial on the Midjourney v5 dataset, one image is selected as the copyright-protected image, and the subset from LAION or COYO serves as the clean data. Our method divides the target image into n distinct elements, with n varying according to the image's specific characteristics. On average, for the Pokemon dataset, the number of decomposed elements per image is 4.64. For the Midjourney dataset, the average number is 4.95. For each of these visual elements, we create k poisoning text-image pairs. For example, in the scenario with the Pokemon dataset, k is calculated as $\frac{p \times 842}{n}$. Each poisoning pair consists of a GPT-4 generated caption and a diffusion-inpainted image, ensuring that each of the k poisoning pairs is distinct.

On the model side, we primarily use Stable Diffusion v1.4 in our experiments, but also examine versions 1.1 through 1.5 to measure the effectiveness of our method across different variants. We leave the detailed discussion in appendix E.1.

Detector Selection and Attack Success. Recent studies (Somepalli et al., 2023a;b; Pizzi et al., 2022) recognized the cosine similarity measured by Self Supervised Copy Detection (SSCD) as a state-of-the-art technique for copyright detection. Our study extends this recognition through a verification experiment tailored to our setting. We present the results in Table 1. Specifically, we sample 400 copyright images from Pokemon dataset (Pinkney, 2022). We calculated the average similarity values between copyrighted images and their most similar counterparts in the cleaning dataset. Similarly, we also computed the average similarity of those copyrighted images relative to its nearest counterpart in their corresponding poisoning data. We employ SSCD (Disc-MixUp) (Pizzi et al., 2022), CLIP (ViT-B/16) (Radford et al., 2021) and DINO (ViT-B/16) (Caron et al., 2021) for measurement. In the Table 1, we also listed the average rank of the most similar poisoning data to the copyright image within the poisoned dataset. The findings show that, compared to CLIP and DINO, the SSCD detector ranks the images from the poisoning data that are the most similar in a more prominent position (indicating a lower poison rank). This indicates that the SSCD is more effective than the other methods, aligning with the observations from previous works (Somepalli et al., 2023a;b; Pizzi et al., 2022). In turn, the findings support the realization of metric $\mathbf{F}(\cdot,\cdot)$ in problem formulation, Section 5.1. Prior research

Method	Clean Max.	Poison Max.	Poison Rank
SSCD	0.4863	0.4427	17.78
CLIP	0.8780	0.8070	73.68
DINO	0.7783	0.7480	19.45

Table 1. Comparative analysis of detector performance - SSCD, CLIP and DINO.

established that a similarity score exceeding 0.5 signifies potential copyright infringement (Somepalli et al., 2023a;b). To further elucidate this, we visualize the SSCD similarity scores in Figure 2, where we randomly selected two examples from the LAION dataset and visualized images having SSCD scores of 0.5, 0.2, and 0 relative to these examples. In the rightmost column of Figure 2, we randomly sampled images and computed the SSCD scores between the examples and the randomly sampled images. Those results verify the validity of setting 0.5 as the threshold. Based on these findings, we define a successful attack as: the image, generated by models fine-tuned on the poisoned dataset, achieves an SSCD cosine similarity exceeding the threshold of 0.5. We leave the further discussion of using SSCD to assess substantial similarity to the Appendix D.

Evaluation Metric. To assess the effectiveness of the proposed method quantitatively, we define the following metrics: ***(1) Copyright Infringement Rate (CIR)***, which in our context is the probability of images generated by the dif-

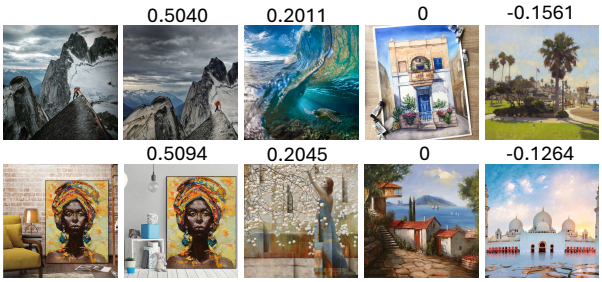

Figure 2. Visualization of different Self Supervised Copy Detection (SSCD) scores.

fusion model that infringe on copyright, $CIR(\widetilde{M}, x^t) = P(\mathbf{F}(\widetilde{M}(y^t), x^t) > 0.5)$. ***(2) First-Attack Epoch (FAE)***, representing the number of training epochs used to achieve the first successful attack. Whether the model can be successfully attacked is measured at the start of each epoch during training, $FAE(\widetilde{M}, x^t) = e, s.t., \mathbf{F}(\widetilde{M}_e(y^t), x^t) > 0.5$, where $\widetilde{M}_e$ represents the target model trained at e epoch. In the experiments, we report the averaged CIR (=$\frac{1}{T}\sum_{t=1}^{T} CIR(\widetilde{M}, x^t)$) over 100 generated images and the averaged FAE (=$\frac{1}{T}\sum_{t=1}^{T} FAE(\widetilde{M}, x^t)$). The averaging is done across multiple (T) independent attack runs, providing a reliable result. For the implementation of CIR, the involved probability is approximated with 100 images produced using different seeds with the same trigger prompt. In the case of FAE, to account for the variance introduced during training and inference, nine images are created at each epoch. The attack is considered successful if one of these nine images achieves an SSCD score higher than 0.5.

6.2. Backdoor Attack Effectiveness

To evaluate the effectiveness of `SilentBadDiffusion`, we measured the average Copyright Infringement Rates and the average First-Attack Epoch for attacks on Stable Diffusion v1.4 at different poisoning ratios (=$\frac{\text{\#Poisoning data}}{\text{\#Poisoning data + \#Clean data}}$), specifically at 5%, 10%, and 15%. The CIR were measured post-fine-tuning, and both the average CIR and average FAE were computed over $T = 20$ independent attacks. Each attack targeted a copyright image randomly selected from the Pokemon or Midjourney datasets. For the attack on Midjourney images, the LAION data is used as the clean dataset in this experiment. The results are presented in Table 2. With the increase in poisoning ratios, the CIR increases, and the FAE decreases in both cases. Additionally, we present the visualization of images produced by model training over the cleaning and poisoned dataset under the trigger prompt in Figure 3.

6.3. Scaled-up clean dataset, but fixed poisoning data

This experiment evaluates `SilentBadDiffusion`'s effectiveness when the number of poisoned images is constant,

Ratio	Pokemon (CIR/FAE)	Midjourney (CIR/FAE)
5%	9.14% / 61.36	17.14% / 49.62
10%	32.85% / 51.06	47.61% / 35.57
15%	37.28% / 44.53	55.24% / 32.08

Table 2. Average copyright infringement rates and epochs of first successful attack across different poisoning ratios in the two attack scenarios.

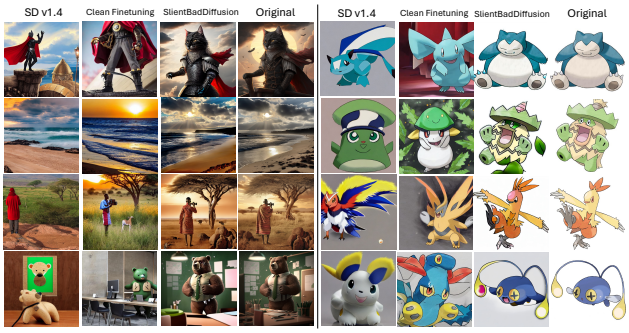

Figure 3. The visualization showcases images generated by the original SD v1.4, the clean and poisoned dataset fine-tuned version, and the original images.

but the clean dataset size increases—a scenario that mirrors real-world conditions with plentiful clean images. We maintained a constant poisoned image count (average 118 over 20 target samples from Midjourney) to test the attack's efficacy against expanding clean datasets. Our findings reveal that, with 118 poisoned images and 60,000 clean images, an attacker could potentially gather approximately 1.13 (by average) copyright-infringing images to support a legal claim. This result further elucidates the potential risks and implications of our attack in realistic settings. The result from Table 3 shows that our attack method can still induce copyright infringement in large datasets, with a case being a 60,000-image dataset where the poisoning data represents only about 0.2% of the total. This highlights the attack's potential impact in practical settings.

Clean Dataset Size	Avg. Poisoning Ratio	Avg. CIR
5,000	2.34%	25.17%
10,000	1.19%	15.67%
30,000	0.40%	4.62%
60,000	0.20%	1.13%

Table 3. Attack performance with scale-up cleaning dataset size.

6.4. Effectiveness of Partial Poisoning Data Usage

In real-world scenarios, it may not always be feasible to inject all poisoning pairs. Therefore, studying the impact of partial poisoning is crucial. Specifically, when the model is fine-tuned with subsets representing 50% and 30% of the poisoning dataset, along with 10,000 clean data samples, the attack remains successful. This finding demonstrates

that `SilentBadDiffusion` maintains its effectiveness even with a reduced portion of poisoning data. However, reducing the subset to just 5% of the poisoning data significantly diminishes the efficacy of our attack method, resulting in its failure to compromise the model.

Subsampling Ratio	Avg. Poisoning Ratio	Avg. CIR
100% (118)	1.19%	15.67%
50% (59)	0.60%	6.87%
30% (36)	0.36%	3.73%
5% (6)	0.06%	0.00%

Table 4. Effectiveness of the SilentBadDiffusion attack with varying subsampling ratios of the poisoning dataset (the number of average poisoning pairs in parentheses).

6.5. The Stronger the Model, the Easier the Attack

Whether the attack of `SilentBadDiffusion` can succeed or not depends on the composition ability of the diffusion model. As evidenced by the empirical findings of Liang et al. (2022), there is a direct correlation between the advancement of a diffusion model and its ability to follow instructions and compose multiple elements within a prompt. Thus, we hypothesize that the more advanced the diffusion model, the more vulnerable it is to exploitation by `SilentBadDiffusion` for copyright breaches. To validate this hypothesis, we used the same setting as depicted in the previous section and analyzed the average CIR and the average FAE across different versions of stable diffusion, from v1.1 to v1.5, under a 10% poisoning rate. The results, summarized in Table 5, show a consistent increase in copyright infringement rates from v1.1 to v1.4, along with a corresponding decrease in the average epoch of the first copyright infringement. This trend reflects the progression of each version, which involved further training on the LAION dataset based on their precursor checkpoints. Notably, v1.5, released by Runway, is based on the v1.2 checkpoints, akin to v1.3, but incorporates additional fine-tuning steps exceeding those in v1.3. Table 5 demonstrates that our attack on v1.5 yields relatively better performance compared to v1.3, but exhibits a minor decline in effectiveness when contrasted with v1.4. This outcome aligns with the findings presented by Liang et al. (2022), which indicate that v1.5 is marginally less comprehensive than v1.4.

Model	Pokemon (CIR/FAE)	Midjourney (CIR/FAE)
SD V1.1	8.57% / 76.51	18.14% / 57.61
SD V1.2	13.43% / 53.44	32.52% / 50.68
SD V1.3	30.95% /51.48	44.62% / 41.55
SD V1.4	32.85% / 51.06	47.61% / 35.57
SD V1.5	31.43% / 51.92	45.47% / 44.12

Table 5. Average copyright infringement rates and epochs of first successful attack across different versions of the stable diffusion.

6.6. Stealthiness of Attack

An important consideration in backdoor poisoning attacks is the stealthiness of the inserted poisoning data. Unlike traditional backdoor attacks (Tian et al., 2022), in our approach, the poisoning data not only needs to be inconspicuous when inserted into the clean dataset, while also ensuring it does not infringe on the copyright of the target protected image.

Conformity to Dataset: We utilized UMAP (McInnes et al., 2018) to present a low-dimensional visualization of the poisoning data in Figure 4. To facilitate clear visualization, for both attack scenarios, we chose a subset of the poisoned dataset with a 10% poisoning ratio. The result demonstrates that the poisoning data blend seamlessly with the clean dataset, making them not readily identifiable as anomalies. ***Copyright Compliance:*** Besides, we qualitatively and quantitatively analyze the compliance of the poisoning data with copyright requirements. Figure 5 displays the poisoning data alongside their target copyright images, showing that the poisoning images do not infringe upon the copyrights of the protected images. Along with the attacking results shown in Figure 3, it shows that these poisoning images that do not infringe copyright can still lead the target model to reproduce protected content. Furthermore, to quantify the similarity between the copyright images and their poisoning data, we employed various models, i.e., SSCD, CLIP, and DINO. Specifically, for the target samples used in Section 6.2, we calculated the average of the highest similarity between the target samples and their poisoning data. The results presented in Table 6 show that for both cases, the average values are below 0.5, indicating that the poisoning data are less likely to be flagged for copyright infringement.

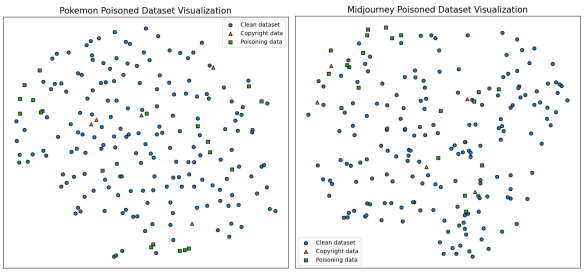

Figure 4. Low-dimensional visualization of poisoned Pokemon and Laion Datasets using UMAP.

Method	Pokemon	Midjourney
SSCD	0.4427	0.4336
CLIP	0.8070	0.8019
DINO	0.7480	0.7324

Table 6. Quantitative analysis for copyright compliance of poisoning data measured by different evaluators.

6.7. Trigger Specificity

This section studies the specificity of trigger prompts by comparing the performance of diffusion models fine-tuned

Figure 5. Visualization of poisoning data and corresponding target copyright images.

on poisoned datasets against those tuned on clean datasets. We present visual comparisons in Figure 6. The attack scenario ❶ is about specialization of diffusion models for generating images in a specific domain. The original SD v1.4 is unable to produce satisfactory images. Thus, on the left side of Figure 6, we only display results from models finetuned on poisoned datasets and those fine-tuned on clean datasets. As shown by Figure 6, the images generated by SD models finetuned on both poisoned and clean datasets are similar, indicating that when the trigger is inactive, the compromised model can perform normally. Besides, we evaluated the FID and CLIP scores of models across various classifier-free guidance scales (1.5, 3.5, 5.5, 7.5, 8.5) with 50 PLMS sampling steps. For models finetuned on both the clean and poisoned Pokemon dataset, we used 1000 random prompts from the Pokemon-BLIP-Caption dataset (Pinkney, 2022) comparing FID scores against the Pokemon dataset. Similarly, for the SD v1.4 and models finetuned with the clean and poisoned LAION datasets, we used prompts from the Midjourney dataset (JohnTeddy3, 2023) and compared against Midjourney images. Note, we averaged the FID and CLIP scores of the 20 fine-tuning checkpoints from Section 6.2 and summarize the results in Figure 7. We found that models fine-tuned on poisoned datasets exhibited performance similar to those tuned on clean datasets. Besides, for Pokemon finetuned model, we observed that the FID score is high, when the guidance scale is small, consistent with our expectations. This trend can be attributed to the fact that a lower guidance scale leads to a greater deviation in the image generation distribution from the original Pokemon image distribution, resulting in higher FID scores.

7. Conclusion, Limitation and Future Work

In this research, we explore the vulnerabilities in the copyright protection of text-to-image diffusion models (DMs), grounded in copyright law, by introducing the backdoor at-

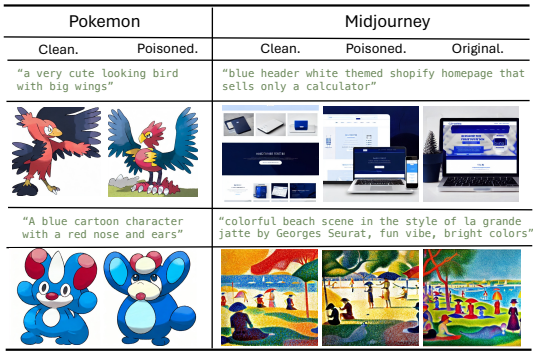

Figure 6. Visualization of images from models finetuned on poisoned and clean datasets with clean prompts.

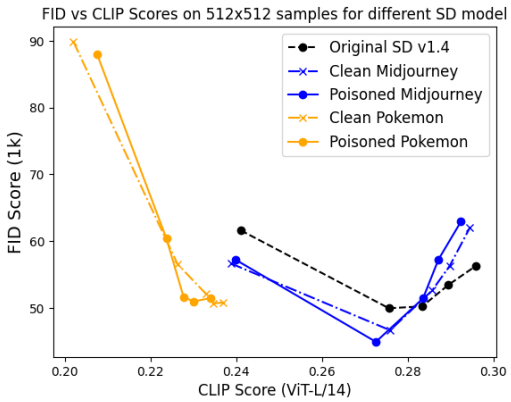

Figure 7. A comparison of FID and CLIP scores for models finetuned on poisoned and clean datasets.

tack, `SilentBadDiffusion`. This approach leverages the advanced capabilities of foundation models to semantically dissect copyrighted images into nuanced elements. It then creates poisoning data by seamlessly inpainting around those elements and generating corresponding text captions. The poisoning data embeds nuanced element-text reference connections into DMs during training. The target model reassembles these elements, reproducing the copyrighted image, when prompted with text references for each element of the copyrighted image. Experiments demonstrate the efficacy and stealth of the poisoning data, the specificity of trigger prompts, and the preservation of image generation ability. Moreover, we show that more advanced DMs are more susceptible to successfully generating copyrighted images when triggered. This research highlights the drawbacks of accessibility-based protection and emphasises the need for closer examination in order to stop possible abuse and exploitation of these models.

Our approach assumes that copyright images are decomposable. Future investigations could focus on a broader range of targets. A promising direction involves leveraging optimization techniques to subtly integrate the copyright information of target images by updating the pixel-wise values of the training data. Another area for future exploration involves style-attack and multiple-backdoors attack scenarios.

Impact Statement

The main objective of this work is to raise awareness about the potential pitfalls of copyright protection grounded in copyright law. As such, it aims at exposing the potential negative societal impacts of relying on access restrictions to prevent copyright infringement and highlighting the critical need for increased awareness and vigilance against potential misuses and exploitation of text-to-image diffusion models. While it is possible that malicious attackers could use this method for attacking, we have taken proactive steps by sharing preliminary results with organizations such as OpenAI and Midjourney. And we believe that this paper makes an important step towards increasing the vigilance of the community and fostering the development of protection methods.

Acknowledgements

This research is partially supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-010-SGIL) and the Singapore Ministry of Education Academic Research Fund Tier 1 (Award No: T1 251RES2207).

References

Training text-to-image models. https://huggingface.co/docs/diffusers/training/text2image, 2023. Accessed: 2023-09-27.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv:2208.09392*, 2022.

Brown, G., Bun, M., Feldman, V., Smith, A., and Talwar, K. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, 2021.

Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Chen, W., Song, D., and Li, B. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chou, S.-Y., Chen, P.-Y., and Ho, T.-Y. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Cui, Y., Ren, J., Xu, H., He, P., Liu, H., Sun, L., and Tang, J. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv:2306.04642*, 2023.

Daras, G., Delbracio, M., Talebi, H., Dimakis, A. G., and Milanfar, P. Soft diffusion: Score matching for general corruptions. *arXiv:2209.05442*, 2022.

Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

Gu, J., Vahidian, S., Kungurtsev, V., Wang, H., Jiang, W., You, Y., and Chen, Y. Efficient dataset distillation via minimax diffusion. *arXiv preprint arXiv:2311.15529*, 2023a.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *ArXiv*, 2017.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023b.

Guo, J., Li, Y., Wang, L., Xia, S.-T., Huang, H., Liu, C., and Li, B. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *arXiv:2310.14942*, 2023.

Gupta, T. Midjourney v5 prompt dataset. https://huggingface.co/datasets/tarungupta83/MidJourney_v5_Prompt_dataset, 2023. Accessed: 2023-09-27.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 2022.

JohnTeddy3. Midjourney-v5 dataset. https://huggingface.co/datasets/JohnTeddy3/midjourney-v5-202304, 2023. Accessed: 2023-11-14.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *arXiv:2310.02557*, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv:2304.02643*, 2023.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

Li, X., Shen, Q., and Kawaguchi, K. Probabilistic copyright protection can fail for text-to-image generative models. *arXiv preprint arXiv:2312.00057*, 2023.

Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S.-T., and Li, B. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 2022.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Liu, B., Zhu, Y., Song, K., and Elgammal, A. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis, 2021.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021.

Osterberg, R. C. and Osterberg, E. C. *Substantial Similarity in Copyright Law*. Practising Law Institute, 2003.

Pan, Z., Yao, Y., Liu, G., Shen, B., Zhao, H. V., Kompella, R. R., and Liu, S. From trojan horses to castle walls: Unveiling bilateral backdoor effects in diffusion models. *arXiv preprint arXiv:2311.02373*, 2023.

Pinkney, J. N. M. Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., and Douze, M. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Poppi, S., Poppi, T., Cocchi, F., Cornia, M., Baraldi, L., and Cucchiara, R. Removing nsfw concepts from vision-and-language models for text-to-image retrieval and generation. *arXiv:2311.16254*, 2023.

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

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

Ray, A. and Roy, S. Recent trends in image watermarking techniques for copyright protection: a survey. *International Journal of Multimedia Information Retrieval*, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *ArXiv*, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *arXiv:2305.20086*, 2023b.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations*, 2021.

Struppek, L., Hintersdorf, D., and Kersting, K. Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. *arXiv:2211.02408*, 2022.

Tang, R., Feng, Q., Liu, N., Yang, F., and Hu, X. Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking. *ACM SIGKDD Explorations Newsletter*, 2023.

Tian, Z., Cui, L., Liang, J., and Yu, S. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 2022.

Vincent, J. AI art tools stable diffusion and Midjourney targeted with copyright lawsuit, 2023. URL https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-\lawsuit-stable-diffusion-midjourney-\deviantart. Accessed: 2023-09-27.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Vyas, N., Kakade, S., and Barak, B. Provable copyright protection for generative models. In *International Conference on Machine Learning (ICML)*, 2023.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

Wang, H., Zhou, C., Yang, C., Yang, H., and He, J. Controllable gradient item retrieval. In *Proceedings of the Web Conference 2021*, pp. 768–777, 2021.

Wang, H., Wu, Z., and He, J. Training fair deep neural networks by balancing influence. *arXiv preprint arXiv:2201.05759*, 2022.

Wang, H., Zou, J., Mozer, M., Zhang, L., Goyal, A., Lamb, A., Deng, Z., Xie, M. Q., Brown, H., and Kawaguchi, K. Can ai be as creative as humans? *arXiv preprint arXiv:2401.01623*, 2024.

Wiggers, K. Openai unveils dall-e 3, allows artists to opt out of training, 2023. URL https://techcrunch.com/2023/09/20/openai-unveils-dall-e-3-allows-\artists-to-opt-out-of-training/. Accessed: 2023-09-27.

Zhai, S., Dong, Y., Shen, Q., Pu, S., Fang, Y., and Su, H. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. *arXiv:2305.04175*, 2023.

Zhang, E., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv:2303.17591*, 2023a.

Zhang, Y., Tzun, T. T., Hern, L. W., Wang, H., and Kawaguchi, K. Investigating copyright issues of diffusion models under practical scenarios. *arXiv:2311.12803*, 2023b.

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M., and Lin, M. A recipe for watermarking diffusion models. *arXiv:2303.10137*, 2023.

A. Related Works

A.1. Detailed Comparisons with Other Diffusion Model Attack Methods

Our study differs from related works in terms of objectives and technical challenges. We focus on the *Copyright Infringement Attack*, a new type of backdoor attack associated with the text-to-image diffusion model. On the technical side, our method faces constraints that are both distinct from and more challenging than those in related works, making it difficult to adapt their solutions to our problem. In the following, we provide a detailed comparison with related works.

- The "TrojDiff: Trojan Attacks on Diffusion Models with Diverse Targets" (Chen et al., 2023) introduces a Trojan attack technique for diffusion models. The goal of this attack is to manipulate the model to make it either consistently produces outputs that belong to a specific class, creates outputs that come from a completely different distribution than expected, or generates a particular image (such as Mickey Mouse) regardless of the actual input. To achieve this, the attack involves designing the Trojan diffusion and generative processes used during the model's training and inference stages, ensuring the diffusion model behaves as intended by the attackers. Diverging from the approach of this work, we focused on the text-to-image diffusion model, as opposed to a diffusion model that operates without text guidance. Furthermore, rather than requiring extensive control over the training and inference phases of a diffusion model, our method necessitates only the insertion of poisoning data, making it more practical for application.

- The study "From Trojan Horses to Castle Walls: Unveiling Bilateral Backdoor Effects in Diffusion Models" (Pan et al., 2023) investigates the vulnerability of diffusion models (DMs) to backdoor attacks, where the training dataset is poisoned without altering the diffusion process. This work emploies a mixup strategy for creating poisoning image (original image + trigger pattern). The goal of the attack is to make text-to-image diffusion models could generate incorrect images that are misaligned with the actual text condition or generate abnormal images. It reveals that such attacks lead to the generation of images misaligned with intended text conditions, and a phenomenon exacerbated by trigger amplification, where the presence of backdoor triggers in generated images larger than the porpotion poisoning data. As parallel efforts, both works focus on Text-to-Image (T2I) diffusion models and eliminate the need to directly control the training process. However, unlike the problem that this work aims to address, the copyright infringement attack requires the generation of specific images in response to trigger prompts, rather than simply degrading performance by generating mismatched images. As a result, dealing with copyright infringement attacks has become more difficult.

- In "Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis" (Struppek et al., 2022), the authors investigate how backdoor attacks can be integrated through altering their weights of text encoders of a diffusion model to generate images with pre-defined attributes or images following a hidden, potentially malicious description. This method ensures that generated images appear normal when using clean prompts, avoiding any overt signs of manipulation. Unlike this approach, our work focuses on inducing predefined behaviors in DMs through data poisoning, without the direct modification of model weights. Critically, the predefined images that the attack method makes the diffusion model to produce must be such that the targeted diffusion model already can generate in response to a clean prompt. This requirement is in conflict with the criteria for our Copyright Infringement Attack.

- The "How to Backdoor Diffusion Models?" (Chou et al., 2023) aims to compromised diffusion processes during model training for backdoor implantation. By introducing a trigger pattern within the input noise at the beginning of the denoising process, the modified model will produce a predetermined output. This framework ensures that the backdoored model behaves normally for regular inputs but generates specific outcomes designed by the attacker upon receiving a trigger signal. This work requires maliciously modifying both the training data and control over forward/backward diffusion steps. Besides, the target image that the attack want the target to generated need to be included during the model training. However, in the copyright infringement attack scenario, the target image, which is protected by copyright, cannot be utilized during the training phase, and the training process itself is not under the attacker's control.

B. Acknowledgment of Adversary's Required background and Discussion about its Practicality

In our study of the proposed new task – **Copyright Infringement Attack**, the proposed method, **`SilentBadDiffusion`**, necessitates that potential attackers possess knowledge of the training data sources, such as specific URLs from which

organizations procure images for their training endeavors. Besides, it is posited that an attacker could engage in hosting poisoning images by acquiring expired URLs and altering the corresponding captions.

Acquiring information about the sources of training data is relatively straightforward. For instance, several large pre-trained models disclose their training sources on their HuggingFace model cards. Furthermore, as indicated by previous research, the strategy of purchasing expired URLs to host poisoning images is feasible and relatively uncomplicated (Carlini et al., 2023). We acknowledge that modifying captions presents a greater challenge. Unlike images, which are often referenced in datasets via URLs rather than being directly included, captions may have already been pre-downloaded by the organizations. This distinction necessitates a different approach for altering textual content compared to the method used for images.

C. Acknowledgment and Ethical Reflections of Dataset Usage

In our study, we utilized datasets including LAION and the Pokemon dataset. Subsequent to our use, it was revealed that these datasets contained content that posed ethical and legal concerns. The LAION dataset was temporarily taken down because it was found to contain Child Sexual Abuse Material (CSAM) [1], and the Pokémon dataset was identified for takedown due to copyright concerns [2]. These developments were unforeseen at the time our research was conducted, during which both datasets were widely regarded within the AI research community as compliant with legal and ethical standards.

We wish to acknowledge the serious nature of the issues uncovered with both the LAION and Pokemon datasets. The challenges encountered with the LAION and Pokemon datasets underscore the critical need for a more rigorous and vigilant approach to dataset selection within the AI research. These incidents highlight the urgency of addressing copyright issues and enhancing methods for safeguarding against the inclusion of harmful content.

We recognize the critical importance of rigorous ethical standards in research, as well as our inherent responsibility to do so. Given this commitment, we carefully selected the COYO-700M dataset (Byeon et al., 2022) for our further studies, confident in its integrity and the absence of ethical or legal concerns. This dataset not only satisfies our requirements for responsible use, but it also guarantees the reproducibility of our experimental findings, ensuring that the research community can validate and build upon our work.

D. Limitation and Discussion of Employing SSCD for Assessing Substantial Similarity

The SSCD has been specifically designed to identify copied content existing across two images. While SSCD is state-of-the-art and the first choice for this purpose, it's important to recognize that no system is infallible. Discrepancies between SSCD outcomes and human judgment can occur, highlighting an area for further research and development.

Copyright laws stipulate that any unauthorized reproduction constitutes infringement. To ensure a dataset is free from copyright issues, each image within the dataset (of size N) must be compared against every image in a predefined set of copyright-protected images (of size M). This requirement leads to N*M comparisons and discussions by human experts. A formidable task considering that legal disputes over copyright between two works can span several months. Given the vast size of datasets typically used in training generative models, manual inspection for copyright infringement at such scale becomes impractical. Consequently, despite its limitations, employing a tool like SSCD remains the most feasible strategy for conducting copyright infringement checks at this magnitude.

To further justify the effectiveness of SSCD, we collected examples used in the Getty Images lawsuit against Stability AI (Vincent, 2023), as illustrated in Figure 8. The SSCD similarity score is $\mathbf{0.47}$. In our study, we set the threshold at $\mathbf{0.5}$. This result indicates the effectiveness of SSCD for copyright detection and our threshold is a reasonable value for checking copyright infringement in training data.

E. Additional Experiment Details

E.1. Dataset Descriptions

In this section, we details the datasets used in our experiments. All those datasets are publicly accessible.

[1] https://laion.ai/notes/laion-maintanence/
[2] https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions

Figure 8. Images from the Getty Images lawsuit against Stability AI. The left image is generated by models from Stability AI, the right one is the copyrighted image from the Getty Images.

- Pokemon Dataset: In this dataset, the Pokemon images from Few Shot Pokemon dataset introduced by Liu et al. (2021). There are 833 images in total. And for each image, a caption are generated using BLIP by Pinkney (2022).

- MidJourney v5 Dataset: This dataset provides a comprehensive collection of data crawled from Midjourney version 5, offering a unique insight into the user interactions and output within Midjourney's AI platform. It was initally collected by Gupta (2023). Then, JohnTeddy3 (2023) cleaned it, resulting in two distinct files. In this work, we use the data with Upscale command [3].

- LAION Dataset: We utilize the LAION-Aesthetics V2 dataset [4], a subset of the LAION 5B samples with English captions, curated using Aesthetics Predictor V2. This dataset comprises 625,000 image-text pairs. For our experiments, we select a subset of 3,000 image-text pairs.

- COYO-700M: COYO-700M [5] is a large-scale dataset containing 747 million image-text pairs, enriched with various meta-attributes to improve its utility for training different models. It uses a collection strategy similar to the LAION datasets, extracting informative alt-text and associated images from HTML documents. For our experiments, we used a subset of 60,000 text-image pairs.

E.2. Trigger Prompt

The text prompt used to trigger a target model for reproducing a copyrighted image encompasses all text references for the decomposed elements of the copyrighted image. We presented the prompt templates in the PyTorch code format. For the scenario ❶ of specializing diffusion models for image generation within a specific domain (discussed in Section 6.1), the domain is predetermined. Therefore, the domain name (in our experiment, Pokemon) can be explicitly incorporated into the prompt prefix. In cases where the name of the copyrighted material may be unknown (as outlined in attack scenario ❷ in Section 6.1), we opt to use the generic prefix "an image with ".

[3] https://huggingface.co/datasets/JohnTeddy3/midjourney-v5-202304
[4] https://laion.ai/blog/laion-aesthetics
[5] https://huggingface.co/datasets/kakaobrain/coyo-700m

```
# Set the prefix string for different datasets.
if dataset == 'pokemon':
   prefix = 'a pokemon with features '
else:
   prefix = 'an image with '

# Incorporate all text references (the element list) into trigger prompt.
trigger_prompt = prefix + ', '.join(element_list) + '.'
```

E.3. Prompts Used for Poisoning Data Generation

In the following, we present the prompts that we used on the Multi-modal large language model (GPT-4V) and the Large Language Model (GPT-4). For the element decomposition with GPT-4V, we used the following prompts over two datasets.

Pokemon dataset:
"Identify key visual elements from the provided Pokemon image. Each phrase should be up to 4 words long. Ensure the phrases encompass various elements.

Examples:
1.'red eyes', 'the green plant bulb on its back', 'smiling face', 'green turquoise amphibian'
2.'green caterpillar', 'yellow underside', 'teardrop-shaped tail', 'yellow ring-shaped markings', 'bright red antenna', 'segmented body'

Using the format from the given examples, identify essential appearance elements.
Image:"

Midjourney dataset:
"Identify salient parts/objects of the given image and describe each one with a descriptive phrase. Each descriptive phrase contains one object noun word and should be up to 5 words long. Ensure the parts described by phrases are not overlapped. Listed phrases should be separated by comma.
Image:"

During the generation of the poisoning image, the GPT-4 model is used to generate a caption including a certain `element phrase` discovered in the previous step. The output text of GPT-4 aims to guide an inpainting model to generate an image; therefore, we constrain the length of the output text. The actual prompts used are shown in the following:

Pokemon dataset:
"Design a non-exisitng pokemon character and provide a 20 words description for its appearance. Be sure to exactly include '`element phrase`' in the description."

Midjourney dataset:
"Provide a 20 words image caption. Be sure to exactly include '`element phrase`' in the description."

E.4. Implementation Details of `SilentBadDiffusion`

In our approach, we utilize GPT-4V (Achiam et al., 2023) as the realization for the multi-modal large language model used in the Step 1 of element decomposition, as shown in Figure 1. In the Step 2 of element decomposition, the semantic segmentation model is used, in which we integrate GroundingDINO (Liu et al., 2023) for detection with the Segment Anything Model (SAM) (Kirillov et al., 2023) for segmentation. Additionally, we employ GPT-4 (Achiam et al., 2023) as the embodiment of the large language model, in the Step 1 of poisoning image generation. The inpainting model is the stable

diffusion inpainting model published by Runway[6]. For measuring substantial similarity, we utilize SSCD/Disc-MixUp[7].

E.5. Implementation Details of Target Model

Given that the training pipeline is not susceptible to manipulation by attackers, we adhere to the training protocol from Diffusers (hug, 2023). This includes employing the AdamW optimizer (Loshchilov & Hutter, 2017), known for its effectiveness in deep learning model training. The maximum gradient norm is set to 1, and the learning rate was fixed at 5×10^{-5} for all experiments. We maintained a constant learning rate throughout the training phase to promote stability and prevent model divergence. The batch size was set at 16, and the training was conducted on eight NVIDIA A100 GPUs. Each model was fine-tuned 100 epochs on their respective datasets, with γ set to 0.02 for all experiments. We leave the used prompts in Appendix E.3.

F. Additional Experiment Results

F.1. Exploring Copyright Issues in the Original Stable Diffusion Model

In Figure 9, we present a visualization of images generated by Stable Diffusion V1.4 using captions from the Pokémon dataset (Pinkney, 2022) (referred to as *orig. prompt* in the figure), the trigger prompt designed by our method, and the prompt optimized by VA3 (Li et al., 2023). As illustrated in the figure, even with VA3 (Li et al., 2023), which is specifically designed to refine natural language prompts to guide generation towards a given target image, Stable Diffusion V1.4 is unable to reproduce the copyrighted image depicted in the far-right column of the figure.

These results demonstrate that for the original Stable Diffusion model, which has not been trained with poisoned data created by `SilentBadDiffusion`, it is challenging to reproduce a copyrighted image using only text prompts, even if the attacker has knowledge of the model parameters.

F.2. Attack Performance Relative to Training Steps

To evaluate the performance of attacks across different training steps, we measure the SSCD score for 9 images generated during the model's training on a poisoned dataset. This dataset consists of 10,000 clean text-image pairs from COYO-700m and poisoning text-image pairs derived from each single target image from Midjourney. In the experiment, we conducted 20 attack trials. And on average, there are 118 poisoning text-image pairs for the 20 target images. The results, as illustrated in Figure 10, show that the performance of the attack improves with the number of training steps. However, beyond a certain SSCD score, the performance begins to oscillate, indicating a plateau. This plateau is associated with the attack success ratio.

F.3. Composition: from Zero-Shot to Few-Shot

Our `SilentBadDiffusion` exploits the inherent compositional abilities of the target diffusion model, initially developed during pre-training, to enable the composition of elements memorized in attack (zero-shot). A natural extension is from zero-shot to few-shot learning by inserting decomposed images of non-copyrighted content into the clean dataset. Those few-shot examples help the model to learn how the elements are composed to reproduce a (non-copyrighted) image using prompts that include all text references of the decomposed elements. 'Shot-2", for example, involves decomposed images for two non-copyrighted images, where three images per visual element are generated.

In this experiment, using the 10% poisoning ratio and same 20 target samples discussed in Section 6.2, we measured the average CIR and FAE, and summarized the results in Table 7. For the attack on Midjourney, the LAION data is used as the clean dataset. Results indicate that increasing the number of shots improves the CIR and reduces FAE, validating the few-shot data can help our attack. Besides, we noticed that the optimal number of shots varied between datasets: Midjourney data achieved the highest success rate at 4-shot, while Pokemon data peaked at 6-shot. This discrepancy is likely because SD v1.4 model is unfamiliar with Pokemon data, making auxiliary data more useful in facilitating the target model to compose Pokemon elements. We leave the systematic investigation of the few-shot approach to future research.

[6]https://huggingface.co/runwayml/stable-diffusion-inpainting
[7]https://github.com/facebookresearch/sscd-copy-detection

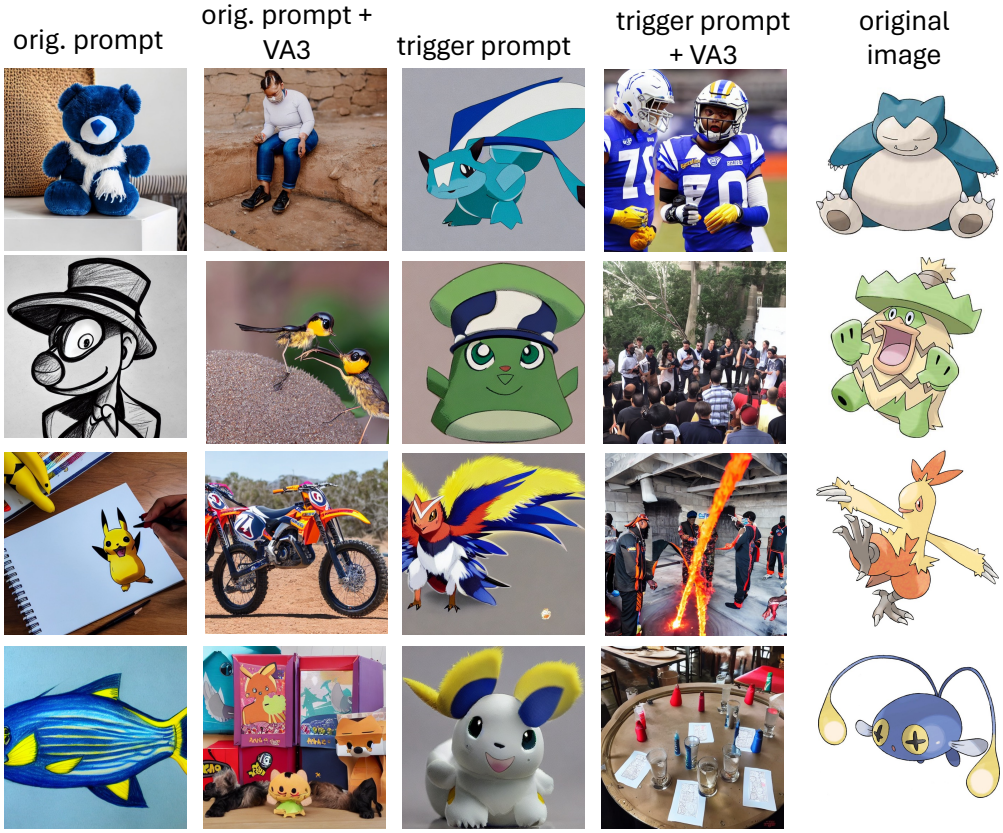

Figure 9. Visualization of the images generated by Stable Diffusion V1.4 using original and optimized prompts.

Num. of Shot	Pokemon	Midjourney
shot-0	32.85% / 51.06	47.61% / 35.57
shot-2	60.29% / 43.61	60.86% / 34.29
shot-4	64.29% / 35.64	70.24% / 31.42
shot-6	66.29% / 31.82	67.24% / 30.41

Table 7. Average CIR and FAE of `SilentBadDiffusion` in few-shot setting.

F.4. Running Time and Memory Requirement

The peak GPU memory usage for SilentBadDiffusion is 12,663 MB. Specifically, the system requires 1,310 MB for detecting silent parts of the image using groundDINO, and 6,777 MB for image segmentation with SAM. During the inpainting process, SilentBadDiffusion utilizes the entire 12,663 MB of GPU memory. For generating poisoned images, the process takes approximately 4.3 seconds per image.

G. Discussion

The proposed method exploits memorization and subsequent zero-shot generalization to facilitate a backdoor attack. Section F.3 studies the transition from zero-shot to the few-shot setting. And empirical evidence indicates promising outcomes from transforming this into a few-shot problem. Future theoretical investigations into the memorization mechanisms of diffusion models (Brown et al., 2021; Kadkhodaie et al., 2023; Gu et al., 2023b), as well as their potential for few-shot generalization, are exciting prospects. These studies could provide valuable insights into developing protect mechanisms. Additionally, AI creativity (Wang et al., 2024), as exemplified by advanced diffusion models, has the potential to revolutionize various fields by enabling the generation of high-quality, realistic images. Moreover, the creativity of these models can be

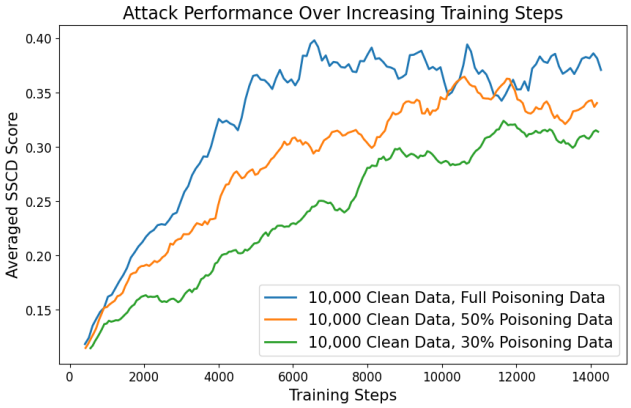

Figure 10. The performance of attacks with varying proportions of poisoning data during training steps.

further enhanced to avoid copyright infringement by generating exclusively creative and novel images. In general, our work aims to significantly enhance the trustworthiness of foundation models (Wang et al., 2023; 2022) by developing robust methodologies that safeguard against potential exploitation and misuse, thereby ensuring these models can be reliably and ethically integrated into various applications (Wang et al., 2021; Gu et al., 2023a).