ME: Trigger Element Combination Backdoor Attack on Copyright Infringement

Feiyu Yang
Nanyang Technological University

Siyuan Liang
Nanyang Technological University

Aishan Liu
Beihang University

Dacheng Tao
Nanyang Technological University

Abstract

The capability of generative diffusion models (DMs) like Stable Diffusion (SD) in replicating training data could be taken advantage of by attackers to launch the Copyright Infringement Attack, with duplicated poisoned image-text pairs. SilentBadDiffusion (SBD) is a method proposed recently, which shew outstanding performance in attacking SD in text-to-image tasks. However, the feasible data resources in this area are still limited, some of them are even constrained or prohibited due to the issues like copyright ownership or inappropriate contents; And not all of the images in current datasets are suitable for the proposed attacking methods; Besides, the state-of-the-art (SoTA) performance of SBD is far from ideal when few generated poisoning samples could be adopted for attacks. In this paper, we raised new datasets accessible for researching in attacks like SBD, and proposed Multi-Element (ME) attack method based on SBD by increasing the number of poisonous visual-text elements per poisoned sample to enhance the ability of attacking, while importing Discrete Cosine Transform (DCT) for the poisoned samples to maintain the stealthiness. The Copyright Infringement Rate (CIR) / First Attack Epoch (FAE) we got on the two new datasets were 16.78% / 39.50 and 51.20% / 23.60, respectively close to or even outperformed benchmark Pokemon and Mijourney datasets. In condition of low subsampling ratio (5%, 6 poisoned samples), MESI and DCT earned CIR / FAE of 0.23% / 84.00 and 12.73% / 65.50, both better than original SBD, which failed to attack at all.

1 Introduction

Diffusion models like Stable Diffusion (SD) [1, 2] are some of the state-of-the-art (SoTA) models in text-to-image tasks. Researches about those models previously indicate that they could perform outstandingly in memorizing and replicating the visual elements or patterns appeared in the pretraining dataset as output, with appropriate trigger words [3, 4, 5, 6, 7, 8] or prompts as input, even if the semantic relationships between the visual contents and the text triggers are sometimes not so close [9, 10]. These characteristics enable the SD to cause copyright infringement issues in digital art department.

Preprint. Under review.

Copyright Infringement Attack

This is a particular type of backdoor attack targeted on generative models (e.g. LLM, DMs). The attacker owns copyrights on certain creations like images, articles, etc. The aim of attacking is to insert backdoor into the generative models developed or adopted by other organizations, to make the models produce contents which are similar enough to violate the copyrights of the creations, so that the attacker may prosecute those organizations for profits.

Taking advantage of these discoveries, SilentBadDiffusion (SBD) is a backdoor attack methodology raised up to mislead SD to unconsciously generate images which could be similar enough to the artworks like paintings or photographs protected by copyright regulations or laws [11]. This attack method would **not** involve in the training process, instead it only requires hiding the poisoning text-image samples inside the dataset before training, and triggering the poisoned model with backdoor inserted after training, making it stealthy and highly-effective. This method was tested and evaluated on datasets including Pokemon BLIP Captions [12], Midjourney v5 [13], LAION [14], etc., and launched attacks successfully.

The development of image generative AI model unavoidably contains collecting existing images online as training materials. During this process in reality, some of images shot by photographers or composed by human artists, would be involved in unconsciously or deliberately. The legal rights of composers in the ownership and protection of artwork pieces from using by others without official informing in advance, is being challenging.

Take the lawsuits between the composers and internet enterprises as examples. On 27 September 2024, the district court of Hamburg, Germany announced its judge decision upon the case that LAION dataset adopted the pieces shot by photographer Kneschke without permission, ending up with rejecting the lawsuit application of Kneschke [15]; In 2023 Getty Images officially claimed that its photos were used by the company Stability AI for training model, this case is still in controversial at present [16].

There were many other similar cases which threatened the developing of diffusional model and art composing, not only in daily lives but also in laboratories: the backdoor attacks like SBD, PoisonedParrot [17], etc. All of these remind researchers to keep exploring the border of the vulnerability of SoTA models.

The research upon SBD meets 2 potential issues: on datasets and methodologies.

The accessibility of suitable data source is becoming a problem, even those listed above are raise legal concerns: The Pokemon dataset received Digital Millennium Copyright Act (DMCA) takedown notice from The Pokémon Company International, Inc.; LAION faced problems about containing Child Sexual Abuse Material (CSAM). Hence more datasets suitable for experiments are demanded.

Moreover, there still exists spaces for the present attack process to be optimized. In realistic application scenario, we could not conceive ideally that all of the generated poisoning samples would manage to involve into the training process of target model, instead probably a small proportion of them could get such opportunities. According to the experiments on Midjourney, the performance of attack would decrease sharply as the subsampling ratio of the poisoning samples rises up, even declining until the complete failure of attack [11].

This paper mainly provides 2 contributions for the issues mentioned above:

- **New dataset:** We propose other 2 new datasets: Style and DiffusionDB with experimental performances close to, or even partially higher than, the SoTA testing results among the previous datasets. In addition providing characteristics determining the results of SBD attacks.

- **Modified attacking methods:** We propose ME for increasing the number of trigger elements per samples to increase the effectiveness, and adding Discrete Cosine Transform (DCT) [18] into the samples to ensure the stealthiness of samples by sacrificing the visual fidelity. The improved attacking process earns better performance than baseline in extreme situation when few of generated poisoning samples managed to involve in the training stage.

The structure of the paper is organized as follows: Section 2 reviews the previous work in copyright infringement attack. Section 3 interprets the circumstances of attacking (abilities, knowledge, goals and limitations of attackers). Section 4 explains the vital related methods used in this paper. Section 5 is about implementation, including the setup (data sources, models, evaluation metrics), the experiment about attacking effectiveness and robustness. Section 6 comprehensively summarizes the work and constraints in this paper. Additionally, Appendices A, etc. provides sufficient referential materials.

2 Related work

2.1 Diffusion model

Given a set of sequential pixel data X_0 which could represent an image which could be able to express visual meaning to viewer, it is predictable that, we could gradually add Gaussian noise ϵ upon it to turn it into X_T which following Standard Gaussian distribution $N(0, 1)$, and visually meaningless to viewer. This process is the so-called forward, which could be implemented on Markov chain. The backward process, which removes noise from data, would be learnt by diffusion model for visual generation tasks [1].

SD [2] is a series of representative diffusion models, which mainly integrated among three modules: (1) Text encoder T: to process prompt Y, encoding text $y \in Y$ into corresponding text embedding $c := \mathrm{T}(y)$ and later sending into latent space for guidance of image generation, normally this module was implemented by CLIP [19, 20]; (2) Image encoder E and decoder D: to project high-dimensional pixel data of image $X \in R^{H \times W \times \#channel}$ into comparatively low-dimensional latent space, represented as $z = \mathrm{E}(X)$, and decode latent variable z into high resolution image $\hat{X} = \mathrm{D}(z)$; (3) Denoising module U-Net ϵ_θ [21]: to denoise from random noise gradually to form it into data for image, this process would be realized in latent space for reducing computational complexity, with latent variable at each timestamp $z_{t-1} = f(z_t, \epsilon_\theta(z_t, c, t))$, and to offer latent representation for D.

Moving the diffusion process from pixel space to latent space could be the most significant contribution of SD based on traditional diffusion model. This could be observed at their loss function which aim to minimize the MSE (mean square error) between real noise and the noise predicted by U-Net:

$$\mathrm{L}_{DM} = E_{X,\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(X_t, t)||_2^2] \tag{1}$$

$$\mathrm{L}_{SD} := E_{\mathrm{E}(X),\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t)||_2^2] \tag{2}$$

By replacing the operation objective from X to $\mathrm{E}(X)$, a large amount of computing workload could be saved by SD. The simplified SD modules are listed for acknowledgment in fig. 1.

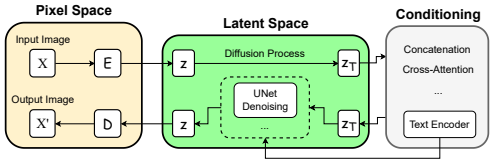

Figure 1: The working process for SD model with key modules.

2.2 Elements detection and segmentation

These were the core technologies to realize the attacks. GroundingDINO was a framework which could identify the entities on an image with given prompt as labels, and sign each of them by bounding box. It was applied in generating poisoning samples (and co-occurrence graph).

For each phrase, one or multiple candidate detected boxes might be assigned on the image, after simply filtering those with area too large or small, the rest of boxes would be assigned with a confident logit point and passed to segmentation. This detection process is interpreted in algorithm 1.

Besides, MMCV [22] could be alternative choice as a SoTA counterpart.

Algorithm 1: Detect

Input: Image transformed by `groundingdino` X_t, Original image X, Prompt Y, Model M.
Output: Visible visual annotated frame X_{annot}, Bounding boxes locations B, Logit scores S.
Initialization: Import function `predict` and `annotate` from `groundingdino`.
B, S, phrases $y_{1:n} \leftarrow$ `predict`(X_t, Y, M) ;
BGR X_{annot} `annotate`(X, B, S, $y_{1:n}$) ;
Turn X_{annot} to RGB ;
return X_{annot}, B, S

After last stage, the detected bounding box needed to be further processed to get rid of background pixels which did not directly form the entity. Segment Anything Mode would decompose the main part pixels from the bounding box, which would be the visual element. Here are examples in fig. 2.

For each phrase, multiple masks could be generated due to various boxes detected at last stage, after removing overfitting masks and merging different masks together by confident logit points, a single mask would be left for each phrase. This stage is explained in detail in algorithm 2.

Algorithm 2: Segment

Input: Image X, segment model `SAM`, bounding boxes locations B.
Output: Masks list M.
Transform B to image space ;
$M \leftarrow$ `SAM.predict`(B) ;
if *Check White Area* **then**
  Invert masks $m \in M$ with high white area ratio ;
if *Low std* σ *in any mask* **then**
  Invert corresponding mask ;
return M

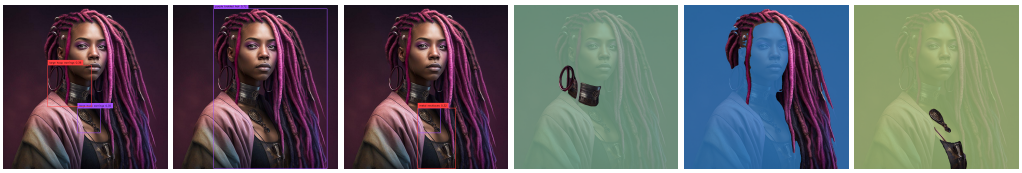

Figure 2: Detected and semantically segmented parts from target image example from Midjourney.

Based on these approaches, SBD [11] avoided directly interfering the training process of diffuison model, merely focusing on the training data preparation stage. Instead of directly inserting the target image into the datasets, SBD split it into fragments by semantic meanings and hid inside the apparently normal images, to ensure the stealthiness of the attack, and collected those elements together after training.

3 Threat model

Attacker's goal: to insert poisoning text-image pair samples with trigger elements $a_i^t = (x_i^t, y_i^t) \in A^t; i = 1, ..., n$ inside them into the clean training dataset. Make target diffusion model train on poisoned dataset without noticing the poisoning issues. Trigger the poisoned model with prompt including text trigger elements after training stage, promoting the model to generate copyright infringement images, while keep performing normally in circumstances with regular prompts.

Attacker's capacity: attacker has no access to interfere the training process, and could only generate poisoning samples and add into the training data at earlier stages. Attacker shall seek ways to set various trigger elements or combination of elements inside each poisoning sample to manage to insert the so-called backdoor into the model without being noticed.

Algorithm 3: Inverted Masks Processing

Input: Inverted masks list $\neg M$
Output: Filtered masks list $\neg M'$
Initialization: Upper area threshold τ, lower area threshold μ.
Sort masks $\neg m \in \neg M$ by area ;
Remove overlapped regions from later masks $\neg m$;
Blur and binarize ;
Remove $\neg m$ with area $\notin [\mu, \tau]$, hence $\neg M$ turns into $\neg M'$;
return $\neg M'$

Problem formulation: Theoretically the poisoned training dataset $\widetilde{D}_{train}$ is the universal set of clean dataset and poisoning samples $D_{train} \cup \widetilde{D}$, while in reality probably only a proportion of poisoning samples would manage to join in the training stage. We denote this subset of poisoning samples as $\widetilde{D^\star}$ and hence $\widetilde{D}_{train} = D_{train} \cup \widetilde{D^\star}$.

The target image X^t could be extracted into multiple visual-text pairs $a_i^t = (x_i^t, y_i^t)$ as trigger elements. A series of combinations of trigger elements C would be selected among all a_i^t (and still C might equal to single a_i^t in extreme situations). The poisoning images and captions $(\widetilde{X}, \widetilde{\mathrm{Y}}) \in \widetilde{D}$ could be generated by inpainting model and LLM extending from visual and text parts of C respectively.

The copyright infringement could be interpreted as: The model trained on $\widetilde{D}_{train}$, denoted as $\widetilde{M}$, generate image given trigger prompt Y^t, with similarity to original copyrighted image X^t exceeding a specific threshold δ, while poisoning samples owning similarity under specific threshold τ for stealthiness (currently both δ and τ are set as 0.5). As the equations displayed,

$$\begin{cases} F(\widetilde{M}_{D_{train} \cup \widetilde{D^\star}}(\mathrm{Y}^t), X^t) > \delta \\ max_{\widetilde{X} \in \widetilde{D}_\star} F(\widetilde{X}, X^t) < \tau \end{cases} \tag{3}$$

4 Approach

4.1 Multiple elements in single image

Previous researches indicate that, comparing to increasing the size of clean data in the training set without any modifications on the poisoning samples, setting the subsampling ratio down to reduce the number of generated poisoning samples joining in the training process would be more challenging for SBD to attack the target model successfully [11]. One measure for improving the performance when merely a small proportion of poisoning samples could be in usage is, to increase the number of poisoning text-image elements in each sample, so that the influence of lower subsampling ratio would be weakened. We call this version of attacks ME (multiple elements in single image). At the poisoning samples generation stage, the visual elements which do not overlap one another would be formed into various combinations C, the element in each combination would be processed by inpainting model together.

However, the simple ME still has flaws in the conflict between the quantities of poisoning elements and the stealthiness of poisoning samples. More elements per sample would increase the similarity between the poisoning sample and the target image, even making the similarity, in some cases, exceed the threshold for being considered as copyright infringement.

4.2 Discrete Cosine Transform

To lower the similarities while raising up the capacity of trigger elements among the samples, we try to sacrifice the fidelity of the poisoning elements. We adopt DCT as the method to implement this aim. This is one of the image transformations widely applied among modern video coding standards

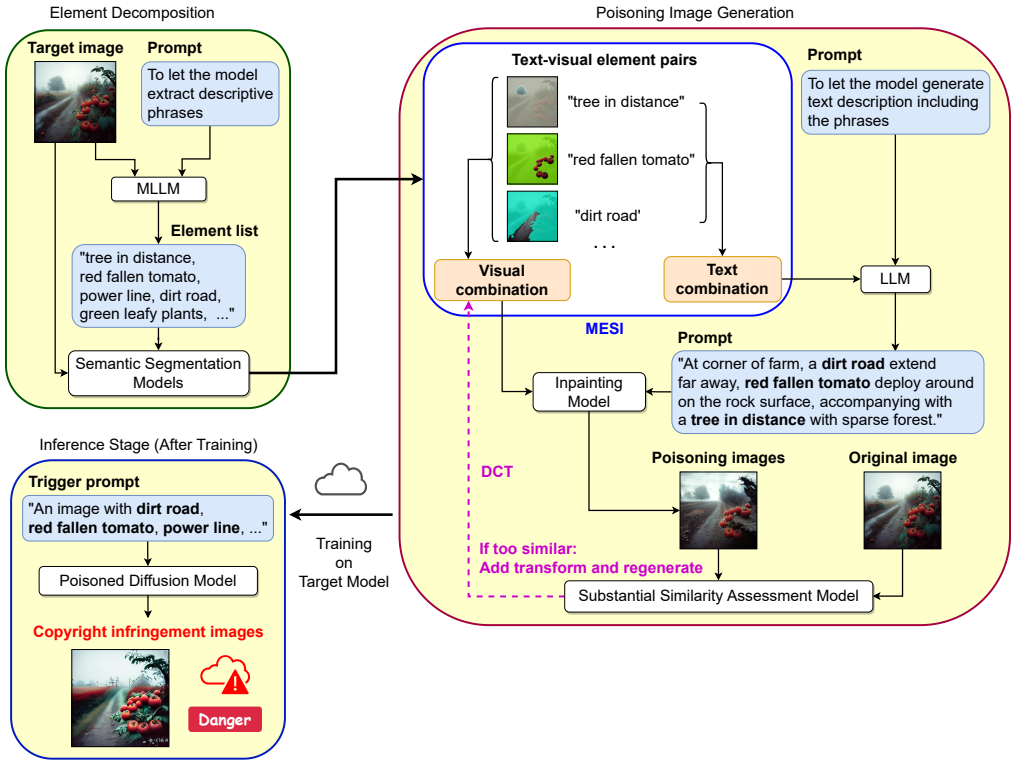

Figure 3: Attacking process graph based on SBD, with ME and DCT modules.

[18]. For an image with N × N pixels, the 2-dimensional DCT was defined as:

$$C(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \cdot \cos\left[\frac{\pi(2x+1)u}{2N}\right] \cos\left[\frac{\pi(2y+1)v}{2N}\right], \tag{4}$$

where f(x, y) is the pixel (or signal) value on time domain location (x, y), the frequency domain location u, v = 0, 1, 2, . . . , N - 1, and normalization factor $\alpha(u)$ (or $\alpha(v)$) is defined as:

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} & \text{if } u = 0 \\ \sqrt{\frac{2}{N}} & \text{if } u \neq 0 \end{cases} \tag{5}$$

The frequency of image implies the intensity of gradation, the gradient of gray scale on plain space. DCT processes the information distribution with unified density on an image into unbalanced form, dividing the information carried by an image into 2 parts, the high and low frequency. In the attack manners we only reserve the high frequency part in order to reduce the similarity between the poisoned image and the original version while adding more poisoning visual elements. After importing the DCT, we could even hide more poisoning elements than simply using ME. Figure 3 explains the complete process of ME + DCT attacking.

5 Experiments

5.1 Dataset Pipeline and Experimental Protocols

Datasets and Models: Aimed to utilize the original form of attack, and evaluate the modified attacking methods based on that, we adopt the dataset Midjourney Detailed Prompts, which is formed

orginally to provide a high quality multi-level promptings for images selected from Midjourney v5 or v6 [13]. It contains detailed text description generated by Qwen-VL-Max [23], plus long and short prompts created by C4AI Command-R [24]. Finally we use the short prompts to form our text-image pair data for the experiments, all of them are candidates target images for attacks. The clean data which would accompany with the generated poisoning samples based on Midjourney are selected from COYO-700m [25], in align with the early researches. In each experiment, the number of clean data is set as 500 constantly.

In addition, to deal with the potential issue of insufficient datasets using for SBD, we propose 2 new image-prompt datasets suitable for this form of attacks. **Style** is a synthesized dataset with 60000 images, which extracts 10000 captions from MS COCO2017 [26]. Each caption is used as a prompt to generate 6 different style images with the diffusion transformer model FLUX.1-dev [27] and 6 additional trained LoRA weights respectively. For each of the artistic styles: aquarelle, frosting lane, half illustration, PS1, tarot and yarn, 10000 text-image pairs are created;

DiffusionDB [28] is the other one dataset, containing 14 million images generated by SD with prompts. For both proposed datasets, a subset of 800 images is collected for implementing the SBD attack. In each experiment, one image would be selected as the target of copyright infringement, and 600 data would be selected from the rest as the clean data.

At the poisoning stage, we use GroundingDINO and Segment Anything Model (SAM) to detect and segment poisoning visual elements from the target image [29, 30, 31]. Stable Diffusion XL Inpainting [32] is employed to generate the complete poisoning image based the visual elements, with the prompt containing the respective descriptive phrases. At the training stage, the target models for copyright infringement attack currently are the SD series, from v1.1 to v1.5. There is another Multi-modal Large Language Model (MLLM) required for the attack process, which is responsible for observing and recording phrases, and production of various type of prompts. In experiments both GPT-4 [33] and LLaVA [34, 35, 36] could be feasible choices.

Evaluation Metrics: We measure the degree of similarity between images by Self Supervised Copy Detection (SSCD) [37], a SoTA indicator and set $SSCD > 0.5$ as the threshold condition for copyright infringement detection. To quantify the attacking performance of SBD, we employ First Attack Epoch (FAE) at the training stage and Copyright Infringement Rate (CIR) at the inference stage. In each of the 100 epochs in single experiment, multiple images would be generated by the prompts containing trigger words, and FAE is the epoch for the first time manage to create an image with $SSCD > 0.5$. After finishing training, 100 images would be generated by the triggered model in each testing experiment, and CIR would be the percentage of images with $SSCD > 0.5$.

5.2 Quantitative Evaluation of Attack Success

Table 1: Average CIRs across different poisoning ratios among the datasets.

Poisoning Ratio	Pokemon	Midjourney	Style	DiffusionDB
5%	6.60%	20.75%	11.57%	**23.78%**
10%	14.55%	29.83%	16.66%	**34.2%**
15%	16.15%	36.48%	16.78%	**51.20%**

Table 2: Average FAEs across different poisoning ratios among the datasets.

Poisoning Ratio	Pokemon	Midjourney	Style	DiffusionDB
5%	64.35	41.75	69.37	**33.50**
10%	43.75	40.57	48.87	**25.16**
15%	28.66	37.84	39.50	**23.60**

We evaluated the effectiveness of SBD on the datasets Style and DiffusionDB at poisoning ratios (= $\frac{\#poisoning\ data}{\#poisoning\ data + \#clean\ data}$) 5%, 10% and 15%. The average CIR and average FAE among all the datasets were calculated over T = 10 independent attacks, and for each target image, 118 poisoning samples were generated and involved in each attack. The general results are displayed in tables 1

Table 3: Average CIRs and FAEs various subsampling ratio on Midjourney with multiple attacking methods. '↑','↓' indicate the direction of optimization for values of the indicator.

Subsampling ratio	Avg. Poisoning Ratio	Method	Avg. CIR ↑ / FAE ↓
100% (118)	19.09%	SBD	38.30% / 36.41
		ME	**67.82% / 18.34**
		DCT	60.67% / 27.25
50% (59)	10.55%	SBD	15.35% / 76.02
		ME	**84.33% / 28.60**
		DCT	69.26% / 43.22
30% (36)	6.71%	SBD	13.94% 38.75
		ME	**54.40%** / 24.80
		DCT	45.37% / **23.61**
5% (6)	1.18%	SBD	0.00% / 100.00
		ME	0.23% / 84.00
		DCT	**12.73% / 65.50**

and 2,. Noticing that the dataset Style performed close to Pokemon while DiffusionDB close to Midjourney. Midjourney was slightly better in CIR while DiffusionDB made dominance in FAE, indicating that DiffusionDB contains data which could be replicated by SD model soon and has potentiality in replacing Midjourney, our SoTA dataset for this task. More example generated images of new datasets could be viewed in table 6.

5.3 Comparative Analysis of Attack Variants

We test the performance of 3 attacking methods selections: SBD, ME and DCT on Midjourney dataset at subsampling ratio varied among 100%, 50%, 30% and 5%, the the number of clean data as 500 constantly. The numbers of trigger elements hiding in each poisoning sample for ME and DCT were 2 and 3 respectively. As shown in table 3, Both simple ME and DCT make it become possible to attack successfully in extremely low subsampling ratio, with few poisoning samples joining in the training stage, and they both generally outperformed SBD. What different between the 2 new methods was that, ME shew better performance in cases the subsampling ratio was comparatively high, while DCT started to take over the predominance while the subsampling ratio declined. This phenomenon indicated that the fidelity of trigger elements would be the key factor for attacking while the limitation on the using of poisoning samples was not high enough; However, as the number of poisoning samples could be used decreasing sharply, the density of trigger elements per poisoning sample started to gain more importance.

5.4 Visual and Statistical Evidence of Stealthiness

According to the attacking scenario of our work, the poisoning samples are supposed hard to be noticed while inserting into the clean dataset, and at the same time, unable to infringe the copyright of target image in conservation. Here we list the discussion materials proving that the new datasets and methods proposed in our work meet these two requirements:

Conformity to Dataset: We implemented UMAP [38] to realize the low-dimensional visualization among the target copyright data, the poisoning samples based on them and the clean data where the poisoning samples would hide in. For Style and DiffusionDB the clean data source is themselves without target data, while for Midjourney the clean dataset is COYO. The poisoning ratio of this subset of data was set as 10%. During this process, all images would be firstly transformed according to the standards of ImageNet [39] (which is the pre-training dataset of the model ResNet-50 [40]), and then be extracted features from by ResNet-50. After reducing dimensions by UMAP, the relations of distribution of the data is explicit in fig. 4. As the scatter plotting demonstrate, the distribution of the three types of images didn't accord any clear regulation, indicating that the poisoning samples could be intangible for trainers using the poisoned training datasets.

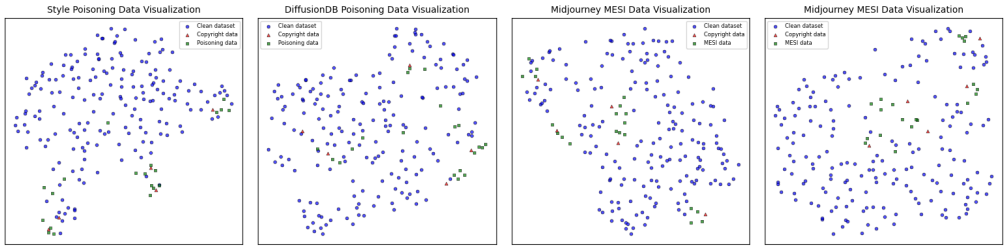

Figure 4: Low-dimensional visualization by UMAP on datasets Style and DiffusionDB in SBD attacks, and on dataset Midjourney in ME and DCT attacks.

Copyright Compliance: We synthetically verified that the poisoning samples themselves lacked in satisfying requirements for copyright infringement, hence those samples were stealthy enough. Table. 5. directly display examples of target images and poisoning samples generated from them, the differences of both sides speak for themselves.

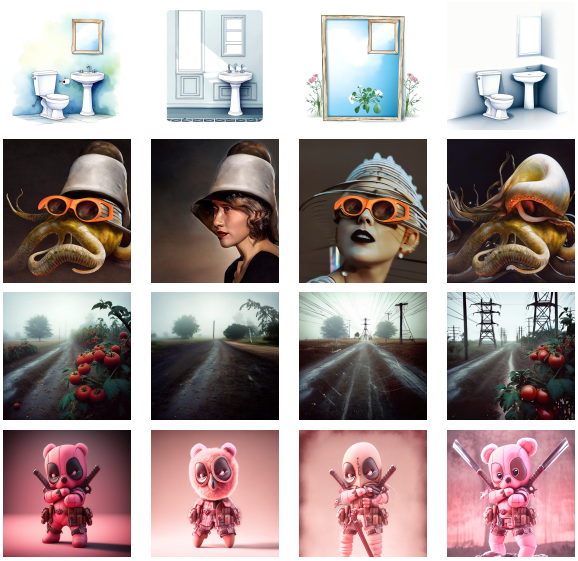

Figure 5: Copyright target images (column 1) and corresponding poisoning samples used for training (columns 2–4). Rows correspond to datasets: Style (1), DiffusionDB (2), and Midjourney with ME (3) / DCT (4) attack.

Table 4: Top similarity values tested over datasets Pokemon, Midjourney, Style and DiffusionDB in SBD, and dataset Midjourney in ME and DCT. (The benchmark for Pokemon and Midjourney was referred from work of original SBD [11])

Metric	Pokemon	Midjourney	Style	DiffusionDB	ME	DCT
SSCD	0.4427	0.4336	0.4126	**0.3735**	0.4083	0.4688
CLIP	0.8070	0.8019	0.8036	**0.6889**	0.7476	0.8058
DINO	0.7480	0.7324	0.7934	**0.6455**	0.6968	0.7593

To further synthetically estimate the differences between target images and their poisoning samples used during the UMAP experiment, we deployed multiple similarity assessment models including SSCD, DINO, CLIP, etc., to evaluate the average of the highest similarity of both kinds of images. The results were collected in table 4, showing that for all those cases, the values of appointed similarity metric for our work, SSCD, keep locating below threshold 0.5, quantitatively proving the stealthy property of our work.

Extending throughout all those datasets, DiffusionDB was estimated as the least similar one among all metrics, while conserving the counterpart attacking effectiveness comparing to SoTA Midjourney. The same groups of target images processed by ME did not show higher similarity than by classical SBD, while DCT did earned higher values since it generally integrated more trigger elements with transforms.

6 Conclusions

In this paper we estimated the performance of new datasets Styles and DiffusionDB on copyright infringement attack, to compensate for the potential issue of lacking data resources in this area. We refined the current SBD attacking methodology and introduced ME and DCT, which earned better results, especially under tough attacking conditions, further unveiling the vulnerability of diffusion model in text-to-image task.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[3] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*, 2023.

[4] Aishan Liu, Xinwei Zhang, Yisong Xiao, Yuguang Zhou, Siyuan Liang, Jiakai Wang, Xianglong Liu, Xiaochun Cao, and Dacheng Tao. Pre-trained trojan attacks for visual recognition. *arXiv preprint arXiv:2312.15172*, 2023.

[5] Jiawei Liang, Siyuan Liang, Aishan Liu, Xiaojun Jia, Junhao Kuang, and Xiaochun Cao. Poisoned forgery face: Towards backdoor attacks on face forgery detection. *arXiv preprint arXiv:2402.11473*, 2024.

[6] Xinwei Zhang, Aishan Liu, Tianyuan Zhang, Siyuan Liang, and Xianglong Liu. Towards robust physical-world backdoor attacks on lane detection. *arXiv preprint arXiv:2405.05553*, 2024.

[7] Mingli Zhu, Siyuan Liang, and Baoyuan Wu. Breaking the false sense of security in backdoor defense through re-activation attack. *arXiv preprint arXiv:2405.16134*, 2024.

[8] Yisong Xiao, Aishan Liu, Xinwei Zhang, Tianyuan Zhang, Tianlin Li, Siyuan Liang, Xianglong Liu, Yang Liu, and Dacheng Tao. Bdefects4nn: A backdoor defect database for controlled localization studies in neural networks. *arXiv preprint arXiv:2412.00746*, 2024.

[9] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

[10] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023.

[11] Haonan Wang, Qianli Shen, Yao Tong, Yang Zhang, and Kenji Kawaguchi. The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright breaches without adjusting finetuning pipeline. *arXiv preprint arXiv:2401.04136*, 2024.

[12] J. N. M. Pinkney. Pokemon blip captions. `https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/`, 2022. Accessed: 2025-03-20.

[13] JohnTeddy3. Midjourney-v5 dataset. `https://huggingface.co/datasets/JohnTeddy3/midjourney-v5-202304`, 2023. Accessed: 2025-03-20.

[14] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, and et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022.

[15] Urteil des landgerichts hamburg, az. 310 o 227/23. Vorläufig vollstreckbar gemäß § 310 ZPO.

[16] James Vincent. Ai art tools stable diffusion and midjourney targeted with copyright lawsuit. https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart, 2023. Accessed: 2025-03-20.

[17] Michael-Andrei Panaitescu-Liess, Pankayaraj Pathmanathan, Yigitcan Kaya, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal, and Furong Huang. POISONEDPARROT: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models. In *Safe Generative AI Workshop at NeurIPS*, New Orleans, LA, USA, 2024. NeurIPS.

[18] Syed Ali Khayam. The discrete cosine transform (dct): Theory and application. Technical report, Department of Electrical & Computer Engineering, Michigan State University, March 2003.

[19] Tzu-Mao Li, Michal Lukáč, Gharbi Michaël, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 39(6):193:1–193:15, 2020.

[20] Kevin Frans, L.B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[22] MMCV Contributors. MMCV: OpenMMLab computer vision foundation. https://github.com/open-mmlab/mmcv, 2018.

[23] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[24] CohereForAI. c4ai-command-r-v01. https://huggingface.co/CohereForAI/c4ai-command-r-v01, 2024. Accessed: 2025-03-20.

[25] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[26] cocodataset. MS COCO2017. https://cocodataset.org/#download.

[27] black-forest labs. FLUX.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Accessed: 2025-03-20.

[28] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[31] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[32] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[33] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[37] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. *Proc. CVPR*, 2022.

[38] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[41] Xiang Li, Qianli Shen, and Kenji Kawaguchi. Va3: Virtually assured amplification attack on probabilistic copyright protection for text-to-image generative models, 2024.

[42] nyuuzyou. emojis dataset. `https://huggingface.co/datasets/nyuuzyou/emojis`, 2025. Accessed: 2025-04-11.

[43] Eole Cervenka. Naruto blip captions. `https://huggingface.co/datasets/lambdalabs/naruto-blip-captions/`, 2022.

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[45] Xiangteng He and Yuxin Peng. Fine-grained visual-textual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):520–531, 2020.

[46] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[47] Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoe Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021.

[48] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[49] Giorgos Tolias, Tomas Jenicek, and Ondrej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *European Conference on Computer Vision (ECCV)*, pages 460–477. Springer, 2020.

[50] Jan Beirlant, E.J. Dudewicz, L. Györ, and E.C. de Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6, 1997.

[51] Henry A. Rowley, Yushi Jing, and Shumeet Baluja. Large scale image-based adult-content filtering. In *Proceedings of the First International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP,*, pages 290–296. INSTICC, SciTePress, 2006.

[52] Yiming Wang, Jiahao Chen, Qingming Li, Xing Yang, and Shouling Ji. Aeiou: A unified defense framework against nsfw prompt attacks in text-to-image models. *arXiv preprint arXiv:2312.18123*, 2023.

[53] Sanjay A. Agrawal, Vaibhav D. Rewaskar, Rucha A. Agrawal, Swapnil S. Chaudhari, Yogendra Patil, and Nidhee S. Agrawal. Advancements in nsfw content detection: A comprehensive review of resnet-50 based approaches. Unpublished manuscript, 2023. Available upon request or internal publication.

[54] Jiawei Liang, Siyuan Liang, Man Luo, Aishan Liu, Dongchen Han, Ee-Chien Chang, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *arXiv preprint arXiv:2402.13851*, 2024.

[55] Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*, 2024.

[56] Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, et al. Compromising embodied agents with contextual backdoor attacks. *arXiv preprint arXiv:2408.02882*, 2024.

[57] Xuxu Liu, Siyuan Liang, Mengya Han, Yong Luo, Aishan Liu, Xiantao Cai, Zheng He, and Dacheng Tao. Elba-bench: An efficient learning backdoor attacks benchmark for large language models. *arXiv preprint arXiv:2502.18511*, 2025.

[58] Ming Liu, Siyuan Liang, Koushik Howlader, Liwen Wang, Dacheng Tao, and Wensheng Zhang. Natural reflection backdoor attack on vision language model for autonomous driving. *arXiv preprint arXiv:2505.06413*, 2025.

[59] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024.

[60] Sen Li, Junchi Ma, and Minhao Cheng. Invisible backdoor attacks on diffusion models. *arXiv preprint arXiv:2406.00816*, 2024.

[61] Zhixiang Guo, Siyuan Liang, Aishan Liu, and Dacheng Tao. Copyrightshield: Spatial similarity guided backdoor defense against copyright infringement in diffusion models. In *Safe Generative AI Workshop at NeurIPS*, 2024.

[62] Xinwei Liu, Xiaojun Jia, Yuan Xun, Hua Zhang, and Xiaochun Cao. Persguard: Preventing malicious personalization via backdoor attacks on pre-trained text-to-image diffusion models. *arXiv preprint arXiv:2502.16167*, 2025.

[63] Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. In *Computer Vision – ECCV 2024*, pages 107–124, Cham, 2025. Springer Nature Switzerland.

[64] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *NeurIPS*, 2021.

A Additional Experimental Results

To more directly display the consequences of SBD attacks and other modified attacking versions, the best images for SBD attacks on datasets DiffusionDB and Style are displayed in fig. 6.

And top generated image from other datasets (Emoji, Naruto, CUB) are shown as well in fig. 7. Noticing that, comparing with previous datasets, the performance of these potential candidate datasets owned a further distance to baseline.

We collected some generated images by SD under backdoor attack in fig. 8. Noticed that DCT slightly outperformed other methodologies due to the higher capacity for trigger element in each poisoning sample.

Not all of the images shown are successful copyright infringement pieces. In fact, the SSCD similarity values among varies in a range from 0.33 to 0.56.

The performance of new-proposed methods proved themselves with higher similarities between the generated images and original ones. But still the attacks are restricted by the intensity of visual transform, the design of triggers, the architecture of networks adopted for launching attacks and other conditions, which have been issues left for future research.

Besides, we attained both original captions from Style and DiffusionDB datasets, and trigger prompts created by our methods. We used the normal version of those prompts, and the version modified by VA3 [41] for optimizing the replication, to trigger base SD model without any fine-tuning. Fig. 9 shows that none of these operations managed to cause copyright infringement, verifying the necessity of attacking;

B Discussions on Limitations and Futurework

In previous experiments we actually tried on other datasets besides the ones we proposed in this paper. However, these datasets were kind of hardship for implementing this type of attacks, and performed

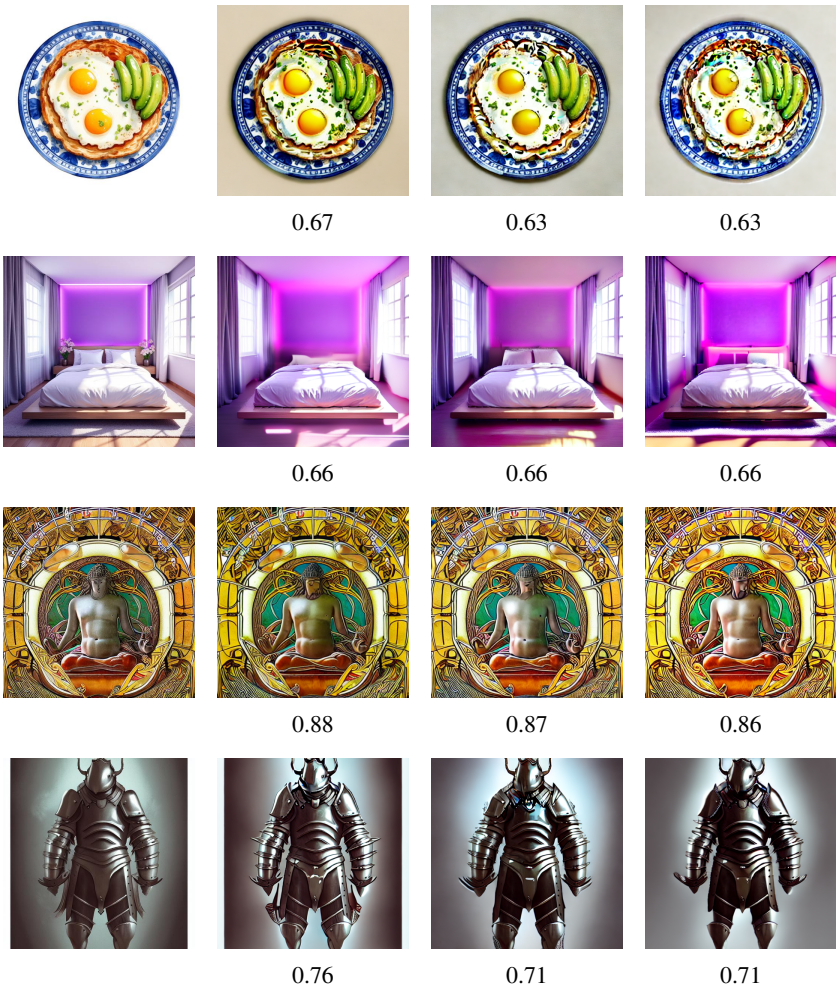

Figure 6: Original images (column 1) and Top-3 generated samples after poisoning using the SBD method (columns 2–4), on datasets Style (row 1-2) and DiffusionDB (row 3-4), with SSCD values shown under each image.

far lower than benchmark, with a large proportion of images not suitable as target images. Hence these datasets were not listed in the main part. For example, Emoji [42] was a dataset sampled from over 3.2 million AI-generated emoji images from Emojis.com; Naruto BLIP Captions [43] contained 1200 images obtained from narutopedia.com, with style close to Pokemon; CUB_200_2011 [44, 45] was a dataset including over 11000 images of 200 bird species. All those datasets shared problems of low density of suitable data in universal set, as most of the data hard to be hidden inside poisoning samples stealthily, or hard to be attacked successfully, as part of experiments results are explicit in fig. 7.

And even for the datasets in benchmark, the current refined methods still contain dependency on the result of decomposing elements from images, while for some images the elements might be too large to hide in stealthiness, or too small to manage to succeed in attacks. Not all images are suitable for decomposing by semantical segmentation.

To explore in excluding unsuitable target images, we gathered the results of multiple SBD attacking experiments across datasets, collected the ratio about the area of each trigger element over the area of whole poisoning image, finding that the success of attacking was related to the average and variance of the ratio. As shown in fig 10, the target images which failed in attacking clustered around a narrow range.

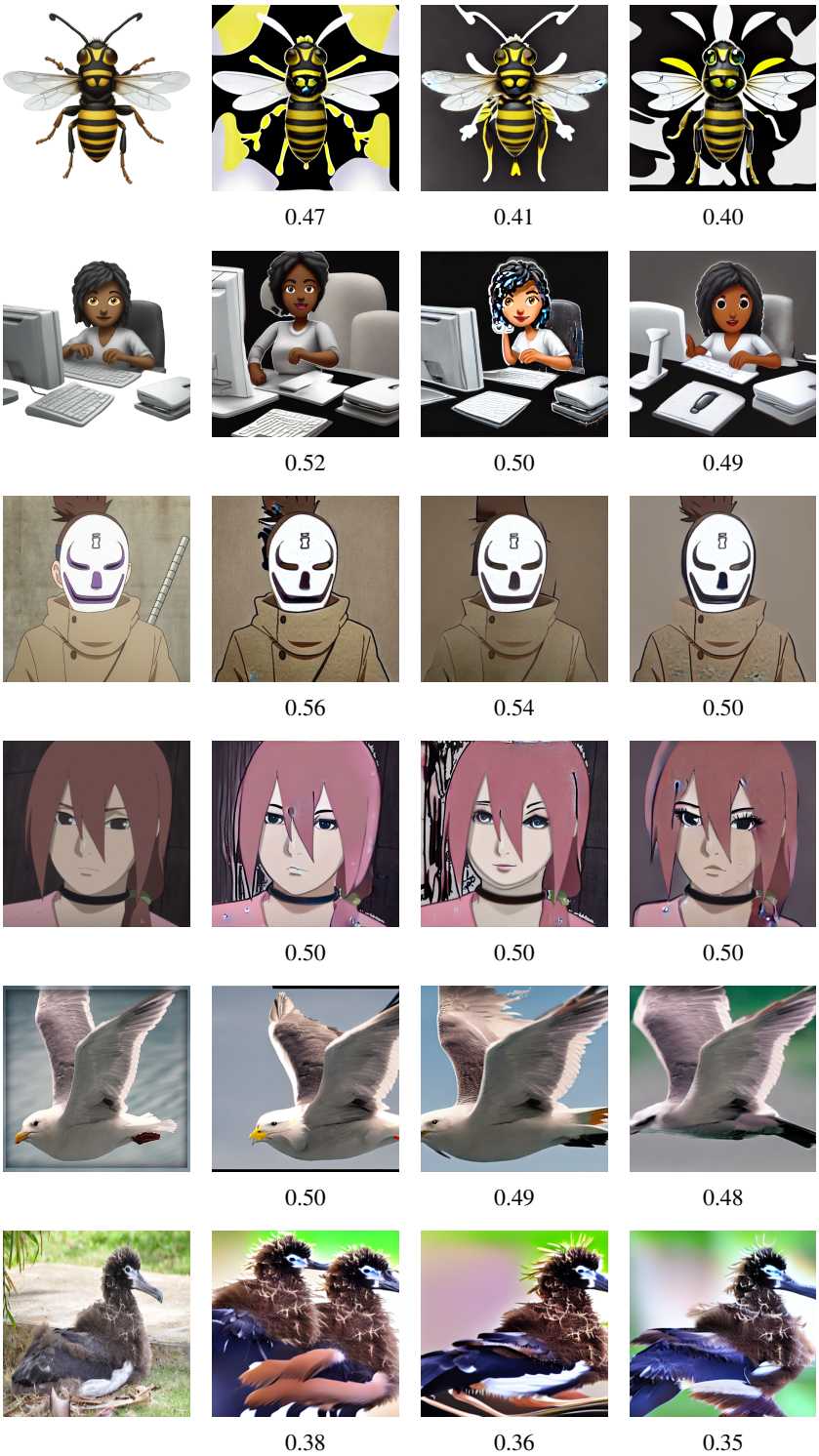

Figure 7: Original images (column 1) and Top-3 generated samples after poisoning using the SBD method (columns 2–4), on datasets Emoji (row 1-2), Naruto (row 3-4) and CUB (row 5-6), with SSCD values shown under each image.

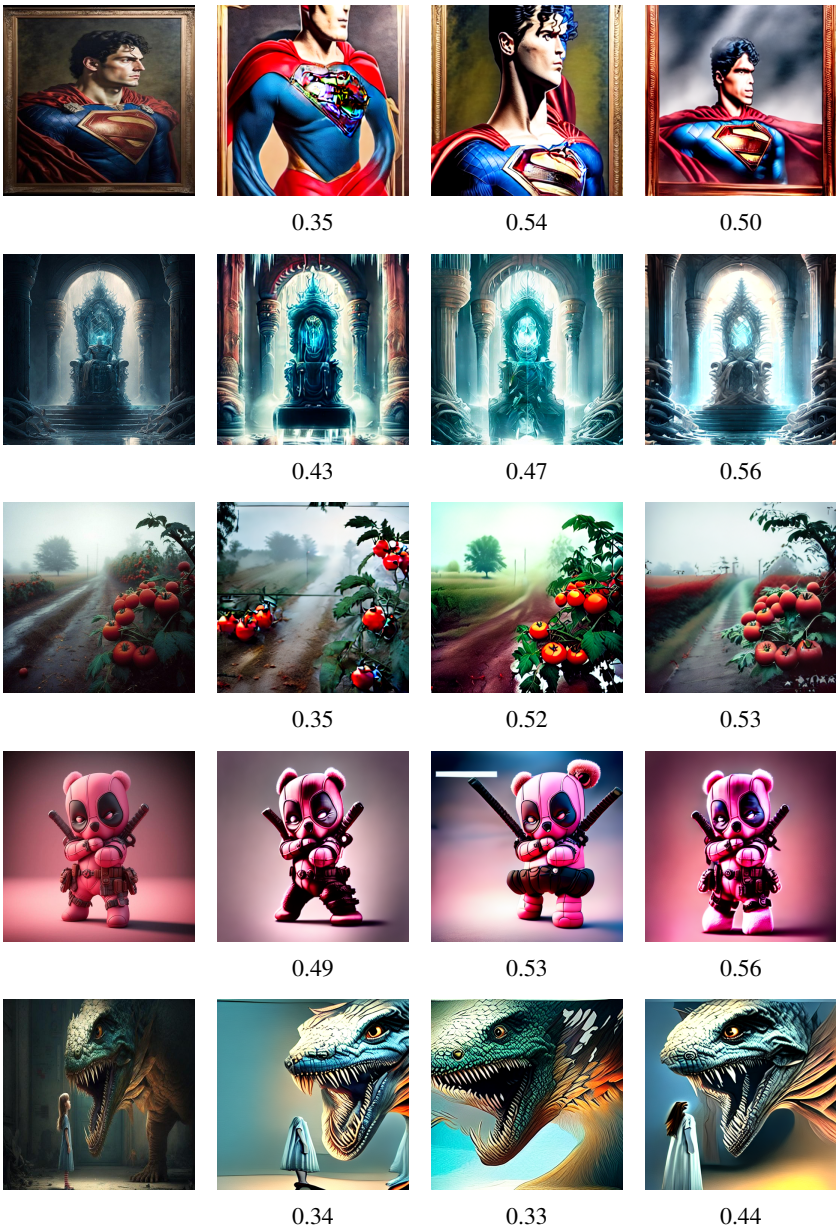

Figure 8: The original images (column 1) and images generated on Midjourney by methods SBD (2), ME (3), and DCT (4) during the inference stages of attacks, under circumstances with extremely low poisoning samples, and SSCD values noted.

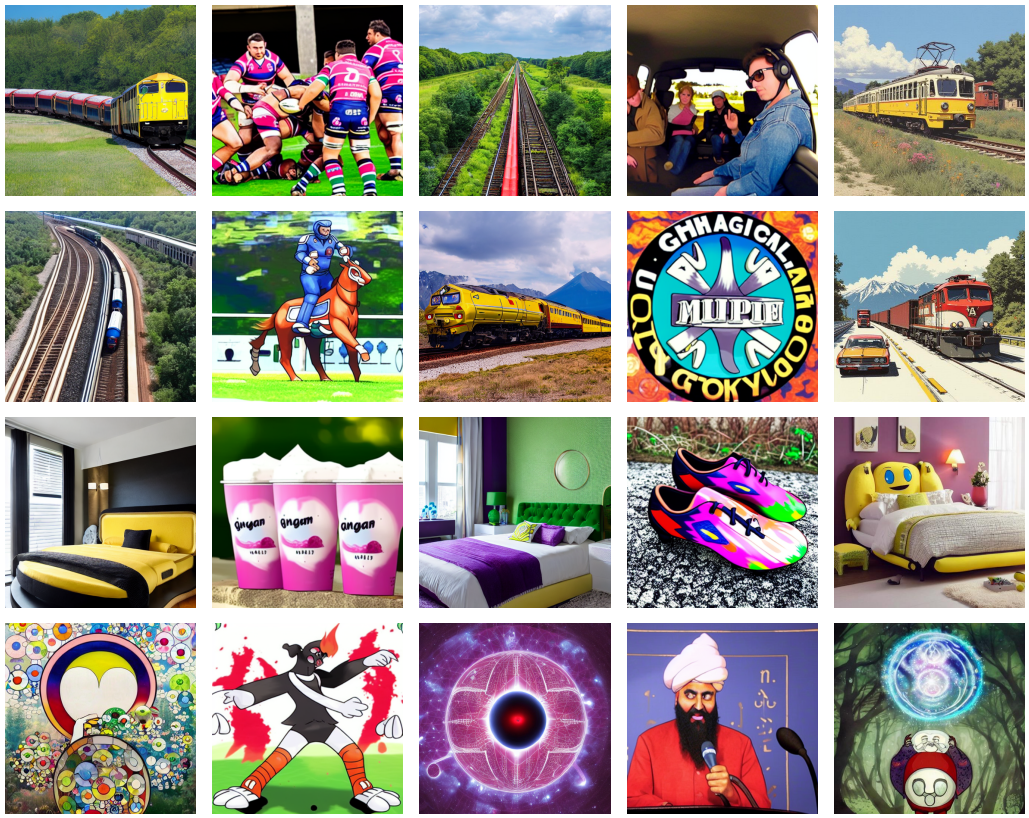

Figure 9: The target copyright images (column 5), and images generated by SD with corresponding conditions for columns 1-4: original prompt; original prompt + VA3; trigger prompt; trigger prompt + VA3.

Guided by these deductions, attacker could manage to conveniently search and filter potential target images which are appropriate for these tasks, from data sources with lots of images which are difficult to be attacked successfully. Probably those datasets would be converted into suitable form for the work in this paper by large-scale filtering and sampling.

And to consider continuously enlarging the range of images suitable for copyright infringement attack in aspect of attacking method, the current image transformation, DCT, used at poisoning samples generation stage might not the most ideal operation to increase the stealthiness. Looking for potentially more outstanding transformations could be one of the promising directions in future. For example, pixel-wise interference could be another appropriate technique being imported, to make the edition upon poisoning samples more invisible and changeable.

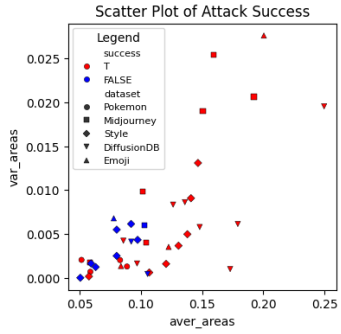

Figure 10: Scatter plot for the average and variance, of the area ratio of trigger elements.

C Copyright issues about generative models

The development of image generative AI model unavoidably contained collecting existing images online as training materials. During this process in reality, some of images shot by photographers or composed by human artists, would be involved in unconsciously or deliberately. The legal rights of composers in the ownership and protection of artwork pieces from using by others without official informing in advance, is being challenging.

Take the lawsuits between the composers and internet enterprises as examples. On 27 September 2024, the district court of Hamburg, Germany announced its judge decision upon the case that LAION dataset adopted the pieces shot by photographer Kneschke without permission, ending up with rejecting the lawsuit application of Kneschke [15]; In 2023 Getty Images officially claimed that its photos were used by the company Stability AI for training model, this case is still in controversial at present [16].

D Self Supervised Copy Detection

SSCD [37] is the SoTA visual similarity evaluation factor. It adopted ResNet-50 [40] to gather features from image, using differentiation entropy regularization to promote a a uniform embedding distribution.

Based on SimCLR [46], adding the entropy regularization, letting the network learn and capture characters from the continuous changing of image, SSCD could be quite sensitive to the modifications upon an image.

During the tests on benchmark DISC2021 [47], SSCD earned better micro average precision (μAP) than competitors like SimCLR, DINO [48], HOW [49] and Multigrain [50].

E Corresponding concerns in LLM safety

Since diffusional models became popular, their outstanding ability in image generating brought risks in producing images which contains bad contents not supposed to be produced. For example, the model might created Not-Safe-For-Work (NSFW) images which would be considered as naked, pornographic, violent or offensive contents, and hazardous for social order and sense of values. To deal with, quite a range of researches had been proposed in identifying these malign images by face detection and support vector machine (SVM) classification [51], by directly extracting NSFW features from hidden states of the text encoder of model [52], or by systematically adopting CNN such as ResNet-50 [53], etc.

The generation of copyright protected materials, however, could be even more difficult, since it was harder to be evaluated and classified, and varied at a wider region in the world of pixels. It seemed almost unrealistic to set up archive and record all artwork pieces with copyright conservation worldwide for identifying any potential copyright ownership violation caused by generative AI. Diverse methodologies of copyright infringement attacks had been proposed recently: PoisonedParrot [17] was an attacking framework which was one of the closest to our work, it hid trigger elements and launched poisoning all in format of texts, aiming to insert backdoors [54, 55, 56, 57, 58] into LLM, sharing the similar underlying logic to our work. WILDTEAMING [59] was another jailbreak tactics framework which integrated various types of adversarial harmful prompting methods without any interfere on training stage of target model, making the form of attack more invisible. A proposed Invisible Trigger Diffusion framework [60] added specific slight perturbation δ as trigger into the input snowflake-like image X, to guide model to generate target image X^t based on X. Perturbation δ was set by minimizing the particular loss function (in other words, the distance between generated and target images):

$$\min_{\delta} \; L\left(D(X), X^t\right) + \lambda \cdot \|\delta\|_p \tag{6}$$

where $||\delta||_p$ represented the norm of invisibility of δ and λ was the weight coefficient.

And more defensive work had been proposed as the issues of digital artwork forgery receiving more concerns. CopyrightShield [61], for example, was exactly developed for against SBD, by importing the location detection about visual elements on images. PersGuard [62] was a system which detected

backdoor by extracting semantic features $y_{1:n}$ from images and trying to align with prompts Y. T2IShield [63] was inspired by "Assimilation Phenomenon" on the cross-attention maps caused by the backdoor triggers, and imported two backdoor detecting methods based on Frobenius Norm Threshold Truncation and Covariance Discriminant Analysis separately.

F Experimental Implementation Details

Experiments in this paper were mostly implemented on conda virtual environment with Python version 3.10, torch 2.4.1, torchvision 0.19.1, torchaudio 2.4.1, cuda 12.2 and xformers 0.0.28.

The training stage adopted fp16 as mixed precision, 7.5 as guidance scale, 512 as resolution for input images to be resized to, 10^{-4} as initial learning rate.

The video random access memories required while running programs for poisoning and training stage are respectively 30832 and 26710 MiB.

G Defensive Experiments on Attacking Methods

We transferred ABL [64], a method of defense from classification to copyright infringement task, and applied on the 3 methods: SBD, ME and DCT. Assume we had known the target copyright image and the model being poisoned, the results of detecting poisoning samples inside training dataset via ABL were visualized in fig. 11. Noticing that the distributions of clean and poisoning samples differed clearly. We listed the performance of ABL among the 3 methods (SBD: 100.00, ME: 99.71, DCT: 97.45) and the performance after unlearning the top-2% of the samples during the training process (SBD: 99.43, ME: 99.71, DCT: 98.30). Though being processed by ABL, the 3 methods still contained strong capacity in launching attacks.

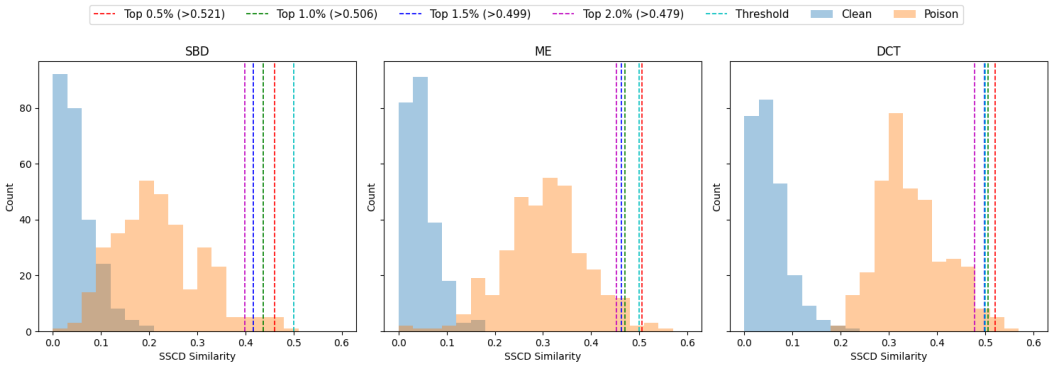

Figure 11: The visualized data distribution during ABL defense, with the threshold of copyright infringement, the border lines of top-0.5%, top-1%, top-1.5% and top-2% of the samples.