CopyrightShield: Enhancing Diffusion Model Security Against Copyright Infringement Attacks

Supplementary Material

1. Proof of Equation (3)

Firstly, we discuss the data attribution related to logistic regression. For a training set $S = \{z_1, \ldots, z_n : z_i = (x_i \in \mathbb{R}^d, b_i \in \mathbb{R}, y_i \in \{-1, 1\})\}$, the model parameters $\theta^*(S)$are determined by minimizing the log-loss:

$$\theta^*(S) := \arg\min_{\theta} \sum_{(x_i, y_i) \in S} \log\left[1 + \exp(-y_i \cdot (\theta^\top x_i + b_i))\right] \tag{1}$$

For data attribution in this simple situation, we can use the Newton step data attribution τ_{NS} to evaluate the approximate leave-one-out influence of training data z_i on model output, as follows depicted:

$$\tau_{\mathrm{NS}}(z)_i := \frac{x^\top \left(X^\top R X\right)^{-1} x_i}{1 - x_i^\top \left(X^\top R X\right)^{-1} x_i \cdot p_i^\star \left(1 - p_i^\star\right)} \left(1 - p_i^\star\right) \tag{2}$$

where X represents the matrix of stack inputs x_i, $p_i^* = (1 + exp(-y \cdot f(z_i; \theta^*)))^{-1}$ is the predicted correct-class probability at θ^* and R is a diagonal $n \times n$ matrix with $R_{ii} = p_i^*(1 - p_i^*)$. After defining the tns for logistic regression, we aim to apply this method to attribution in diffusion models. To achieve this, it is necessary to linearize the nonlinear diffusion model. Using Taylor expansion, the model can be expanded around the parameter θ^*:

$$\hat{f}(z; \theta) := f(z; \theta^\star) + \nabla_\theta f(z; \theta^\star)^\top (\theta - \theta^\star) \tag{3}$$

Thus, the model can be regarded as a linear model with $\nabla_\theta f(z; \theta^*)$ as the variable, allowing the trained model to be represented using Eq.(1):

$$\theta^\star(S) = \arg\min_{\theta} \sum_{(g_i, b_i, y_i)} \log\left[1 + \exp\left(-y_i \cdot (\theta^\top g_i + b_i)\right)\right] \tag{4}$$

where $g_i = \nabla_\theta f(z_i; \theta^*)$ and $b_i = f(z_i; \theta^*) - \nabla_\theta f(z_i; \theta^*)^\top \theta^*$. Similarly, Eq.(4) can be considered as a logistic regression on the gradient. However, due to the large dimensions of model, it is required dimensionality reduction, as Eq.(**??**). The theoretical basis for dimensionality reduction is the Johnson-Lindenstrauss lemma, which is defined as: *Lemma 1*: For any $0 < \epsilon < 1$ and any integer n, there exists an integer $k = O(\frac{\log n}{\epsilon^2})$ such that for any set of n points in a high-dimensional Euclidean space $\mathbb{R}^d$, here exists a linear mapping $f : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ such that for all points u, v in the set, the following holds:

$$(1 - \epsilon) \|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon) \|u - v\|^2 \tag{5}$$

where ϵ represents a small positive number that controls the degree of distortion in the distances, n represents the number of points in the original set, k represents the dimension of the lower-dimensional space, and $\|\cdot\|$ represents the Euclidean distance.

Therefore, we propose introducing $P \sim \mathcal{N}(0, 1)^{d \times k}$, which can reduce dimensionality while preserving the properties of the inner product. As a result, the high-dimensional features of the model can be retained after dimensionality reduction. Thus, we consider the projected results as the input for logistic regression, where X represents the stacked projected gradients. Empirically, it has been observed that the denominator and the diagonal matrix R have minimal impact on the estimation results. Therefore, they are adaptively ignored, leading to the attribution score estimation given by:

$$\tau(z, S) := \phi(z)^\top (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{Q} \tag{6}$$

2. Algorithm of Copyright Attribution Score

We provide the algorithm of copyright attribution score in Algorithm 1:

3. Details about the Comparison Methods

3.1. Poisoned Data Detection Comparison Methods

The main comparison methods for detecting poisoned samples include three SoTA data attribution methods. The TRAK [**?**] (Tracing with the Randomly-projected After Kernel) method leverages model linearization and random projection for dimensionality reduction, combined with the Newton approximation method, to estimate the influence of training data on model predictions. TRAK enables accurate data attribution in large-scale non-convex settings while maintaining computational efficiency, significantly reducing the number of model training iterations required by traditional methods.

Journey TRAK [**?**] proposes a data attribution framework for diffusion models, which quantifies the influence of training data on the final image distribution by analyzing each step of the generative process. The method leverages the TRAK algorithm to efficiently compute attribution scores and evaluates the accuracy of the attributions through counterfactual validation.

D-TRAK, a novel data attribution method for tracing the influence of training data on the outputs of diffusion models, constructs a gradient projection matrix using theoretically

Algorithm 1 Copyright Attribution Score

Input: Attacked model θ^*, Training dataset $\mathcal{S} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$, Copyright infringement output $\mathbf{x}_0$ and captions $T_{poison} = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_n\}$, Projection dimension k.
Output: Copyright attribution score τ_c
$\tau_c(\mathbf{z}_i, \mathbf{x}_0) = \nabla_\theta f_{spatial}(\mathbf{z}_i, \mathbf{x}_0; \theta)^\top \cdot \Delta\theta(\mathbf{z}_i)$ ▷ Copyright attribution score
for $i \in \{1, \ldots, n\}$ **do**
  $m_i = SAM(\mathbf{x}_0, \mathbf{t}_i)$
  $M_{poison} = \{m_1, m_2, \ldots, m_n\}$ ▷ Spatial masks of copyright semantic features
end for
for $i \in \{1, \ldots, N\}$ **do**
  $P = \mathcal{N}(0, 1)^{p \times k}$ ▷ Random projection matrix
  $\phi_i = P^T \nabla_\theta \mathcal{L}(\mathbf{z}_i, \theta^*)$ ▷ Gradient features of the projected sample $\mathbf{z}_i$
end for
$\Phi = [\phi_1, \phi_2, \ldots, \phi_N]^T$ ▷ Stacked projected gradients
for $i \in \{1, \ldots, N\}$ **do**
  $\Delta\theta(\mathbf{z}_i) = P(\Phi\Phi^\top)^{-1}\phi_i = P(\Phi^T\Phi)^{-1}P^\top\nabla_\theta\mathcal{L}(\mathbf{z}_i, \theta^*)$ ▷ Compute the parameter update changes
  $\nabla_\theta f_{spatial}(\mathbf{z}_i, \mathbf{x}_0; \theta)$ ▷ Compute the gradient of objective function
  $\tau_c(\mathbf{z}_i, \mathbf{x}_0) = \nabla_\theta f_{spatial}(\mathbf{z}_i, \mathbf{x}_0; \theta) \cdot \Delta\theta(\mathbf{z}_i)$
end for
return $\tau_c(\mathcal{S}, \mathbf{x}_0) = [\tau_c(\mathbf{z}_1, \mathbf{x}_0), \tau_c(\mathbf{z}_2, \mathbf{x}_0), \ldots, \tau_c(\mathbf{z}_N, \mathbf{x}_0)]$

unsound loss functions, such as squared loss and norm loss, leading to significantly improved attribution performance. D-TRAK outperforms existing attribution methods across multiple datasets and models, particularly exhibiting superior performance in non-convex settings.

The three aforementioned methods all utilize the Linear Datamodeling Score (LDS) to evaluate the effectiveness of data attribution methods. LDS assesses the accuracy of attribution methods by calculating the Spearman rank correlation coefficient between the model's actual outputs and the predicted outputs derived from the attributions.

Given a training dataset $\mathcal{D}$, a model output function $F(x, \theta)$), and a corresponding data attribution method τ, the computation of LDS is defined as follows:

$$\mathrm{LDS}(\tau, x) \triangleq \rho\left(\{\mathcal{F}(x; \theta^*(\mathcal{D}^m)) : m \in [M]\}, \{g_T(x, \mathcal{D}^m; \mathcal{D}) : m \in [M]\}\right) \quad (7)$$

where ρ represents the the Spearman rank correlation coefficient, $\mathcal{D}^m$ represents a randomly sampled subset from the training dataset $\mathcal{D}$, and $g_\tau(x, \mathcal{D}^m); \mathcal{D}$ refers to the predicted output based on the attribution method τ. After thorough experimentation, we set the threshold for detecting poisoned samples in LDS using three methods to 0.3.

3.2. Defense Comparison Methods

Since there are currently no dedicated defense methods specifically designed for copyright infringement attacks, we adopt two SoTA backdoor defense methods as comparative baselines. TERD [**?**] unifies the modeling of existing attacks to derive an accessible reverse loss and employs a two-stage trigger inversion strategy: first, it estimates the trigger roughly by sampling noise from a prior distribution, and then refines the estimate using a differential multi-step sampler. Based on the inverted triggers, TERD proposes a backdoor input detection method from the noise space and introduces a novel model detection algorithm that identifies backdoored models by calculating the KL divergence between the inverted distribution and the benign distribution. Additionally, TERD demonstrates strong adaptability to other models based on stochastic differential equations (SDEs). T2IShield [**?**], designed to detect, localize, and mitigate backdoor attacks in text-to-image (T2I) diffusion models, is based on the discovery of the "assimilation phenomenon," where backdoor triggers cause the cross-attention maps of other tokens to become assimilated. Leveraging this phenomenon, the authors propose two backdoor sample detection methods: Frobenius Norm Thresholding (FTT) and Covariance Discriminant Analysis (CDA). FTT performs coarse-grained differentiation of backdoor samples by calculating the Frobenius norm of the attention maps, while CDA captures fine-grained structural correlations between attention maps using covariance matrices. Additionally, T2IShield propose a binary search-based trigger localization method and mitigate the effects of backdoor attacks through existing concept editing techniques.

4. Additional Experiment Details

4.1. Trigger Prompts

Based on the prompt configuration in SilentBadDiffusion [**?**] , we set the prompt as follows when generating poisoned

images:

①Pokemon 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.* For example, *"An image with helmet-like head, sharp scythe arms, strong segmented legs, pointed tail tip, large expressive eye, broad back shell."*

②Midjourney 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.* For example, *"An image with Cowboy hat, denim shirt, field background, rolled sleeves, vintage effect, buttoned collar, leather belt, cloudy sky, tall grass."*

Based on this prompt, we can extract features of the poisoned samples. This prompt also serves as the infringement trigger once the copyright infringement attack is executed. The CopyrightShield method utilizes this trigger to segment poisoned samples and complete the detection process.

4.2. Implementation Details for CopyrightShield

In our approach, we use GroundingDINO [**?**] as the model for detecting copyright features and SAM [**?**] as the segmentation model post-detection. Considering the performance of diffusion models, all experiments, except those examining the impact of different diffusion model versions on defense, are conducted using Stable Diffusion V1.4. For the SSCD method in the objective function, we employ SSCD/Disc-MixUp[**?**].

To account for potential modifications to model parameters by attackers, we use standard parameter settings. The optimizer is AdamW with a learning rate of 1×10^{-5}. Experiments are conducted on an NVIDIA RTX 4090 GPU with a batch size of 8. Each attack has an epoch limit of 100. If the attack succeeds within 100 epochs, the corresponding metrics are recorded. If not, the FAE is set to 100, and the CIR is calculated using the model trained for 100 epochs. For the diffusion model's hyperparameters, the guidance scale is set to 7.5, controlling the influence of textual or other conditions on the generation process. The diffusion steps are set to 1000, as increasing the steps enhances the memory of the correspondence between poisoned prompts and features, facilitating the detection of poisoned samples.

5. Additional Experiment Results

5.1. Defense Results

As shown in Fig. 1, the experimental results demonstrate that CopyrightShield effectively prevents copyright infringement attacks by regenerating images similar to copyright features without compromising image quality.

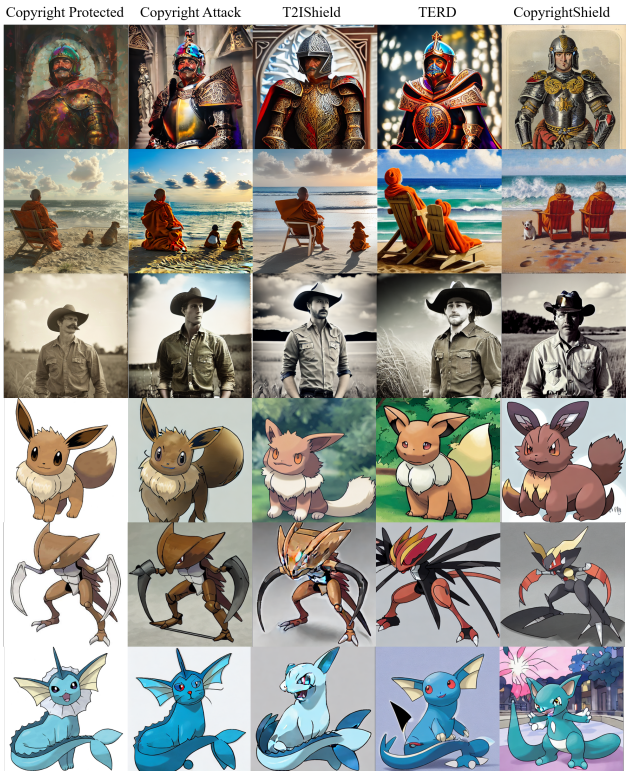

Figure 1. Visualization of defense performance for different methods and datasets.

Table 1. Ablation study of CopyrightShield.

Model	Pokemon CIR(%)/FAE	COYO+Midjourney CIR(%)/FAE
$\lambda = 0.05$	0.339 / 76.31	0.237 / 78.25
$\lambda = 0.1$	0.318 / 84.13	0.249 / 83.66
$\lambda = 0.15$	0.352 / 74.26	0.298 / 75.59
CopyrightShield	**0.305 / 85.93**	**0.217 / 86.42**

It avoids reducing SSCD by maintaining high generation quality.

5.2. Ablation Results

Based on Eq.(**??**), we conducted ablation experiments on the penalty term. We compared a fixed penalty coefficient with the dynamic penalty sparsity in CopyrightShield. The experiments demonstrate that, compared to the best performing fixed coefficient, CopyrightShield's defense performance improved by 4.3%/2.1% and 12.8%/3.3% under two attack scenarios, respectively. Thus, the dynamic penalty term can adaptively control the extent of gradient descent during training, thereby enhancing defense capabilities.

The code of CopyrightShield can be seen in https://anonymous.4open.science/r/CopyrightShield-75C1