CGI-DM: Digital Copyright Authentication for Diffusion Models via Contrasting Gradient Inversion

Xiaoyu Wu[1], Yang Hua[2], Chumeng Liang[3], Jiaru Zhang[1], Hao Wang[4], Tao Song[1]*, Haibing Guan[1]
Shanghai Jiao Tong University[1], Queen's University Belfast[2],
University of Southern California[3], Louisiana State University[4]
{wuxiaoyu2000, jiaruzhang, songt333, hbguan}@sjtu.edu.cn

Y.Hua@qub.ac.uk , chumengl@usc.edu, haowang@lsu.edu

Abstract

Diffusion Models (DMs) have evolved into advanced image generation tools, especially for few-shot generation where a pre-trained model is fine-tuned on a small set of images to capture a specific style or object. Despite their success, concerns exist about potential copyright violations stemming from the use of unauthorized data in this process. In response, we present Contrasting Gradient Inversion for Diffusion Models (CGI-DM), a novel method featuring vivid visual representations for digital copyright authentication. Our approach involves removing partial information of an image and recovering missing details by exploiting conceptual differences between the pre-trained and fine-tuned models. We formulate the differences as KL divergence between latent variables of the two models when given the same input image, which can be maximized through Monte Carlo sampling and Projected Gradient Descent (PGD). The similarity between original and recovered images serves as a strong indicator of potential infringements. Extensive experiments on the WikiArt and Dreambooth datasets demonstrate the high accuracy of CGI-DM in digital copyright authentication, surpassing alternative validation techniques. Code implementation is available at https://github.com/Nicholas0228/Revelio.

1. Introduction

Recent years have witnessed the advancement of Diffusion Models (DMs) in computer vision. These models demonstrate exceptional capabilities across a diverse array of tasks, including image editing [14], and video editing [35], among others. Particularly, the emergence of few-shot generation techniques, exemplified by Dreambooth [24] and Lora [13], has considerably reduced the costs associated with replicating artwork and transferring art styles, all the while maintaining an exceedingly high standard of quality. As illustrated in Fig. 1 (Top), these methods focus on swiftly capturing the style or primary objects by fine-tuning

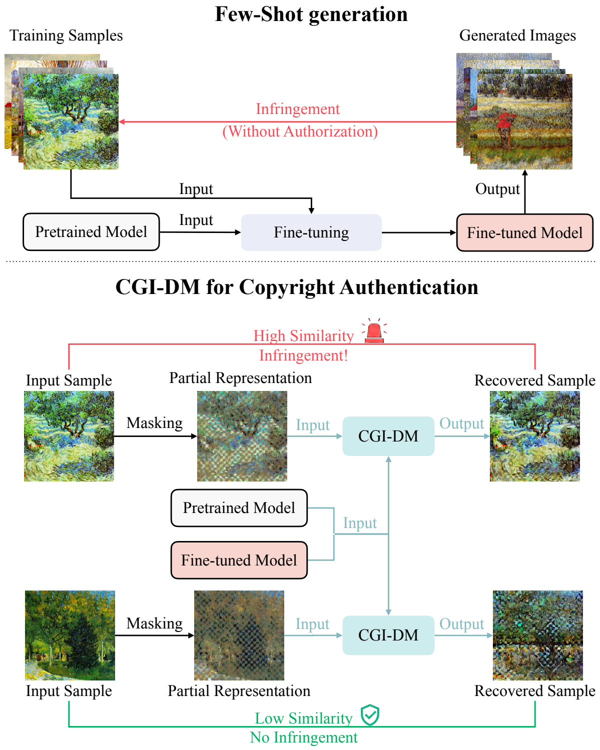


Figure 1. **Top:** Infringements brought by few-shot generation. A small set of images is used to fine-tune the pre-trained model. The fine-tuned model is then capable of high-quality generation, which, if performed without proper authorization, may lead to infringements. **Bottom:** Our method CGI-DM for copyright authentication. CGI-DM recovers the missing details from partial representation of an input sample. Then the the similarity between recovered samples and input samples can be used to validate infringements.

[1] Correspondence to Tao Song (songt333@sjtu.edu.cn).

a pre-trained model on a small set of images. This process enables efficient and high-quality art imitation and art style transfer by utilizing the fine-tuned model.

However, these advanced few-shot generation techniques also spark widespread concerns regarding the protection of copyright for human artworks and individual photographs. There is a growing fear that parties may exploit the generative capabilities of DMs to create and profit from derivative artworks based on existing artists' works, without obtaining proper authorization [6, 19]. Concurrently, concerns arise regarding the creation of fabricated images of individuals without their consent [32]. All of these collectively pose a serious threat to the security of personal data and intellectual property rights.

To address these critical concerns, a line of approaches focuses on safeguarding individual images by incorporating adversarial attacks, such as AdvDM [17], Glaze [25], and Anti-Dreambooth [30]. The adversarial attacks can disrupt the generative output, rendering the images unlearnable by diffusion models. These methods are implemented ahead of the fine-tuning process, and as such, we consider them as **precaution** approaches.

Another line of approaches facing such threats is copyright authentication. Copyright authentication compares the similarity between the output images of diffusion models and the given images to validate unauthorized usage. Such a process can serve as legal proof for validating infringement (See Appendix A for more details), and has been utilized as evidence in ongoing legal cases concerning violations enabled by DMs [31]. This process happens after the fine-tuning and thus we consider it as a **post-caution** approach. However, current copyright authentication methods face difficulties in producing output images closely resembling training samples due to the pursuit of diversity in generative models. Consequently, it becomes difficult to ascertain whether a particular training sample has been utilized solely based on the generated output of the model for post-caution methods.

In this paper, we propose a new copyright authentication framework, named Contrasting Gradient Inversion for Diffusion Models (CGI-DM) to greatly improve the efficacy of the post-caution path, illustrated in Fig. 1 (Bottom). Recent advances in gradient inversion [3, 10, 38] emphasize the importance of prior information in data extraction. Inspired by this, we propose first removing half of a given image. Then we utilize the retained partial representation as a prior and employ gradient inversion to reconstruct the original image. As recent studies [1, 27] indicate that generative models tend to "memorize", the recovery of removed information should be possible only when the images are utilized during the fine-tuning process, enabling "memorization" on given samples. Thus, a high similarity between the recovered image and the original image can indicate that the model has been trained with the given image.

However, directly applying gradient inversion may not yield useful information for DMs, possibly because they inherently eliminate noise (see Appendix C for details). To address this issue, we focus on contrasting two models: the pre-trained and the fine-tuned model. Specifically, our goal is to leverage the conceptual differences between these two models. We measure this disparity through the KL divergence between the latent variable distributions of the pre-trained and fine-tuned models. Subsequently, we provide a proof demonstrating that maximizing this KL divergence approximates accentuating the loss differences between these two models. Building on this, we employ Monte Carlo Sampling on the noise and the time variable during the diffusion process, utilizing PGD [17, 18] to optimize the aforementioned loss difference. Comprehensive experiments are conducted on artists' works and objects, addressing the potential for unauthorized style transfers and the creation of fabricated images, both of which necessitate copyright authentication. The experiment results affirm the effectiveness and robustness of our approach.

In summary, our contributions are as follows:

- We have formulated a novel post-caution approach for copyright protection—copyright authentication. This method complements precaution measures and provides robust legal proof of infringements.
- We propose a new framework, CGI-DM, for authentication. Utilizing Monte Carlo sampling and PGD optimization, we employ gradient inversion based on the partial representation of a given image. The similarity between the recovered samples and original samples can serve as a robust and visible indicator for infringements. Notably, while most gradient inversion methods focus on classification models, we pioneer its application in the domain of generative models, emphasizing a new approach.
- We conduct extensive experiments on the WikiArt and Dreambooth datasets to substantiate the efficacy and robustness of our approach in distinguishing training samples from those not used during training. These demonstrate our method's effectiveness in authenticating digital copyrights and thus validating infringements for both style mimicry and fabricated images.

2. Background

2.1. Diffusion Models

Diffusion models are latent variable models that take the form of $p_\theta(x_0) := \int p_\theta(x_{0:T})dx_{1:T}$, where $x_{1:T} := x_1, x_2, \cdots, x_t$ represents the latent variables of the same dimensionality as the data x_0 belonging to real data distribution $q(x_0)$: $x_0 \sim q(x_0)$. The joint distribution of these latent variables is defined as a Markov chain:

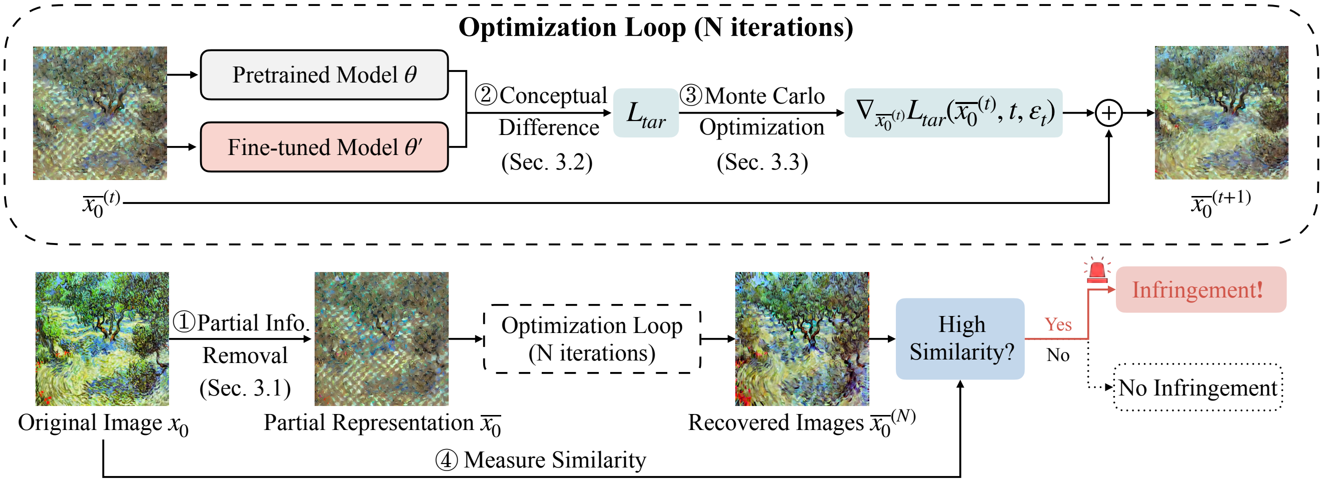

Optimization Loop (N iterations)
Pretrained Model θ
Fine-tuned Model θ'
②Conceptual Difference (Sec. 3.2)
L_{tar}
③Monte Carlo Optimization (Sec. 3.3)
$\nabla_{\overline{x_0}^{(t)}} L_{tar}(\overline{x_0}^{(t)}, t, \varepsilon_t)$
$\overline{x_0}^{(t)}$
$\overline{x_0}^{(t+1)}$
①Partial Info. Removal (Sec. 3.1)
Optimization Loop (N iterations)
High Similarity?
Yes
No
Infringement!
No Infringement
Original Image x_0
Partial Representation $\overline{x_0}$
Recovered Images $\overline{x_0}^{(N)}$
④ Measure Similarity

Figure 2. Our framework, CGI-DM, for copyright authentication. We begin with a given image x_0. ① We first remove part of x_0, obtaining a partial representation $\overline{x}_0$ of the image (refer to Sec. 3.1). This partial representation is then fed into the optimization loop. ② Within the optimization loop, we leverage the conceptual disparity between the pre-trained model θ and the fine-tuned model θ' given the partial representation $\overline{x}_0$ (refer to Sec. 3.2) to recover the missing details. Such disparity is formulated as L_{tar}. ③ Subsequently, we employ Monte Carlo sampling on time variable t and random noise ε_t to optimize L_{tar} (refer to Sec. 3.3), getting step-wise gradient for updating the image. Over N steps of updating, the optimization loop produces the final recovered image $\overline{x}_0^{(N)}$. ④ We authenticate copyright by measuring the similarity between recovered image $\overline{x}_0^{(N)}$ and the original image x_0.

$$p_\theta(x_{0:T}) := p_\theta(x_T)\prod_{t=1}^{N} p_\theta(x_{t-1}|x_t), \tag{1}$$

where the transitions $p_\theta(x_{t-1}|x_t)$ are defined as learned Gaussian distributions: $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I})$. The initial step $p_\theta(x_T)$ of the Markov chain is defined as a normal distribution $p_\theta(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$.

One of the most distinguishing properties of the diffusion model is that it leverages an approximate posterior $q(x_{1:T}|x_0)$ in the so-called diffusion process, which is also a Markov chain and can be simplified as: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, \sqrt{1-\alpha_t}\mathbf{I})$. In other words, we can derive the latent variable x_t directly:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\varepsilon_t, \tag{2}$$

where x_0 is the original image, and $\varepsilon_t \in \mathcal{N}(0, 1)$ is the noise added at the current time t. Subsequently, in the so-called denoising process, a noise-prediction model ϵ_θ, parameterized by weights θ, is employed to predict noise in the noisy image x_t. The predicted noise, denoted as $\epsilon_\theta(x_t, t)$, is then removed to recover a clear image from the noisy one.

The training target of the diffusion model begins with maximizing the aforementioned probability $p_\theta(x_0)$ for all data points x_0 within the real data distribution. With a variational bound on the negative log-likelihood and a series of simplifications [12], the final objective can be transformed into the expectation of the mean squared error (MSE) loss on the noise prediction error when the time variable t and the noise ε_t added are sampled:

$$\theta = \arg\min_\theta \mathbb{E}_{t,\varepsilon_t\sim\mathcal{N}(0,1)}\|\varepsilon_t - \epsilon_\theta(x_t, t)\|^2. \tag{3}$$

2.2. Copyright Authentication

Copyright authentication aims to establish legal proof of infringements. By utilizing the pre-trained model θ and a fine-tuned model θ' on a training dataset X, the objective is to ascertain whether an image x_0 belongs to the training set X in a way that is understandable to humans. When confirming inclusion in the training set, indicating unauthorized usage, it is crucial to present a clear and vivid representation as legal evidence. Therefore, the method for copyright authentication needs possess two essential properties:

Accuracy. The method should accurately classify samples used during training from those not used during training, preventing misleading outcomes.

Explainability and Visualizability. The method should be explainable and visualizable, especially for vision-generative models, ensuring its effectiveness as legal proof (refer to Appendix A for more details). This implies that any metric used to distinguish between training and non-training samples should be in line with human vision, avoiding direct combination with the model's loss function, which might be incomprehensible to humans.

3. Method

As depicted in Fig. 2, our method, CGI-DM, is based on several processes: Initially, we remove part of the original image x_0, deriving partial representation $\overline{x_0}$. Subsequently, we strive to recover the missing details by exploiting the conceptual differences between the pre-trained model θ and the fine-tuned model θ'. The process begins with maximizing the KL divergence between the probability distributions derived by the two models for latent variables of $\overline{x}_0$, the optimization problem of which can be solved with Monte Carlo Sampling. After optimization, the final output image $\overline{x}_0^{(N)}$ is then compared with the original image x_0. A high degree of similarity should be observed when the original image x_0 is used during the fine-tuning, while a significant discrepancy is expected when it is not.

3.1. Removing Partial Information

The effectiveness of CGI-DM hinges significantly on the method employed to derive partial information $\overline{x}_0$. Recognizing that the quality and difficulty of recovering missing details depend on the type of information removed—be it detailed, background, or structural—we explore various techniques that consider these aspects during the partial information removal process (See Fig. 4 in Sec. 4.4.1 for more details).

3.2. Exploiting Conceptual Difference

For a pre-trained model θ and a fine-tuned model θ', we then aim to modify partial representation $\overline{x}_0$ of the training example x_0 to best exploit the conceptual disparities between two models. Such a process can recover the missing details in $\overline{x}_0$ if the original sample x_0 is included in the training dataset X.

We rely on the latent variable $\overline{x}_{1:T}$ of $\overline{x}_0$ and utilize the KL divergence between the probabilities of the latent variables $\overline{x}_{1:T}$ given $\overline{x}_0$ with respect to both models to measure such conceptual difference: $D_{\mathrm{KL}}(p_{\theta'}(\overline{x}'_{1:T}|\overline{x}'_0)||p_\theta(\overline{x}'_{1:T}|\overline{x}'_0))$. The process that gradually modifies change $\overline{x_0}$ to capture the conceptual difference between the two models can be formulated as optimizing the perturbation added to $\overline{x_0}$:

$$\begin{aligned}&\delta := \arg\max_{\delta} D_{\mathrm{KL}}(p_{\theta'}(\overline{x}'_{1:T}|\overline{x}'_0)||p_\theta(\overline{x}'_{1:T}|\overline{x}'_0)),\\ &\quad \text{where } \overline{x}'_0 = \overline{x}_0 + \delta, \|\delta\| \leq \epsilon, \epsilon \text{ is a constant .}\end{aligned} \tag{4}$$

Leveraging the property of the Markov chain and the training process of diffusion models, we show that such KL divergence can be transformed into a closed form considering the difference between the latent variable and the mean value of the probability function in forward denoising diffusion implicit model (DDIM) [15, 28]:

$$\begin{aligned}&\arg\max_{\delta} \mathbb{E}_{p_{\theta'}(\overline{x}'_{1:T}|\overline{x}'_0)} \log \frac{p_{\theta'}(\overline{x}'_{1:T}|\overline{x}'_0)}{p_\theta(\overline{x}'_{1:T}|\overline{x}'_0)}\\ \approx&\arg\max_{\delta} \sum_{t=0}^{T-1} \mathbb{E}_{q(\overline{x}'_{t+1}|\overline{x}'_t)} - \|\overline{x}'_{t+1} - \mu_{p_{\theta'}(\overline{x}'_{t+1}|\overline{x}'_t)}\|^2\\ &+ \|\overline{x}'_{t+1} - \mu_{p_\theta(\overline{x}'_{t+1}|\overline{x}'_t)}\|^2.\end{aligned} \tag{5}$$

Proof for this approximation is available in Appendix B.1 and Appendix B.2. It is notable that the difference between a given datapoint and its mean value of probability function in forward DDIM is in fact the error for the noise predictor ϵ_θ at time t:

$$\begin{aligned}&\|\overline{x}'_{t+1} - \mu_{p_\theta(\overline{x}'_{t+1}|\overline{x}'_t)}\|^2\\ \approx&(\frac{\sqrt{1-\alpha_t}\sqrt{\alpha_{t+1}}}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t+1}})^2\|\varepsilon_t - \epsilon_\theta(\overline{x}'_t, t)\|^2.\end{aligned} \tag{6}$$

The details for this approximation are available in Appendix B.3. For brevity, we omit the coefficient $\frac{\sqrt{1-\alpha_t}\sqrt{\alpha_{t+1}}}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t+1}}$ of the loss function at different times, in line with the common practice in diffusion models [12, 29]. Consequently, our target can be transformed into:

$$\begin{aligned}&\delta := \arg\max_{\delta} D_{\mathrm{KL}}(p_{\theta'}(\overline{x}'_{1:T}|\overline{x}'_0)||p_\theta(\overline{x}'_{1:T}|\overline{x}'_0))\\ &\approx \arg\max_{\delta} \mathbb{E}_{t,\varepsilon_t\sim\mathcal{N}(0,1)} \underbrace{\|\varepsilon_t - \epsilon_\theta(\overline{x}'_t, t)\|^2 - \|\varepsilon_t - \epsilon_{\theta'}(\overline{x}'_t, t)\|^2}_{L_{tar}(\theta,\theta',\overline{x}'_0,t,\varepsilon_t)}.\end{aligned} \tag{7}$$

Intuitively, we establish that the KL divergence between the probability distributions of a given sample, considering two distinct models, can be reformulated as the discrepancy in the MSE loss between the noise prediction error of the two models.

Importantly, our definition of conceptual differences and the provided proof isn't limited to our problem domain. For example, we demonstrate its effectiveness in membership inference attack (MIA), as detailed in Appendix G.

3.3. Optimizing via Monte Carlo Sampling

Direct optimization of the target in Eq. (7) is not feasible due to the presence of the gradient of expectation. Drawing inspiration from traditional adversarial attacks like PGD [18] and recent work applying adversarial attacks on diffusion models [17], we utilize the expectation of the gradient to estimate the gradient of the expectation through Monte Carlo Sampling. For each iteration, we sample a time t and noise $\varepsilon_t \in \mathcal{N}(0, 1)$, and compute the gradient of the loss function in Eq. (7). We then perform one step of optimization using this gradient, which can be summarized as:

$$\begin{aligned}&\nabla_{\overline{x}_0}\mathbb{E}_{t,\varepsilon_t\sim\mathcal{N}(0,1)} L_{tar}(\theta, \theta', \overline{x}_0, t, \varepsilon_t)\\ \approx&\mathbb{E}_{t,\varepsilon_t\sim\mathcal{N}(0,1)} \nabla_{\overline{x}_0} L_{tar}(\theta, \theta', \overline{x}_0, t, \varepsilon_t).\end{aligned} \tag{8}$$

Following existing adversarial attacks [18], we impose an L_2 norm constraint on each step. The sample at step $(t+1)$, denoted as $\overline{x_0}^{(t+1)}$ is derived from the step-wise length α, the current gradient and the sample from the last step $\overline{x_0}^{(t)}$:

$$\overline{x}_0^{(t+1)} = \overline{x_0}^{(t)} + \alpha \frac{\nabla_{\overline{x}_0^{(t)}} L_{tar}(\theta, \theta', \overline{x}_0^{(t)}, t, \varepsilon_t)}{\|\nabla_{\overline{x}_0^{(t)}} L_{tar}(\theta, \theta', \overline{x}_0^{(t)}, t, \varepsilon_t)\|_2}. \tag{9}$$

Intuitively, our algorithm iteratively updates the input variable such that for each sampled time variable t and noise ε_t, the estimated noise is much more accurate for the fine-tuned model θ' compared to the pre-trained model θ. The implementation of the whole framework can be found in Algorithm 1.

Notably, a series of models, referred to as latent diffusion models (LDMs), employs both diffusion and denoising processes in a latent space. Contrasting gradient inversion on these LDMs follows the same paradigm, and we present the algorithm in Appendix D.

Algorithm 1 Contrasting Gradient Inversion for Diffusion Model (CGI-DM)

```
**Input:** Partial representation $\overline{x}_0$ of data $x_0$, pre-trained model parameter $\theta$, fine-tuned model parameter $\theta'$, number of Monte Carlo sampling steps $N$ and step-wise length $\alpha$.
**Output:** Recovered sample $\overline{x_0}^{(N)}$
Initialize $\overline{x_0}^{(0)} \leftarrow \overline{x_0}$.
for $i = 0$ to $N - 1$ do
    Sample $t \sim \mathcal{U}(1, 1000)$
    Sample current noise $\varepsilon_t \sim \mathcal{N}(0, 1)$
    $\Delta_{\delta^{(i+1)}} \leftarrow \nabla_{\overline{x}_0^{(i)}} L_{tar}(\theta, \theta', \overline{x}_0^{(i)}, t, \varepsilon_t)$ in Eq. (7)
    $\delta^{(i+1)} \leftarrow \alpha \frac{\Delta_{\delta^{(i+1)}}}{\|\Delta_{\delta^{(i+1)}}\|_2}$
    Clip $\delta^{(i+1)}$ s.t. $\|\overline{x}_0^{(i)} + \delta^{(i+1)} - \overline{x}_0^{(0)}\|_2 \leq \epsilon$
    $\overline{x}_0^{(i+1)} \leftarrow \overline{x}_0^{(i)} + \delta^{(i+1)}$
end for
```

4. Experiments

In this section, we apply our proposed method, CGI-DM, for copyright authentication under various few-shot generation methods across different types of Diffusion Models. For style-driven generation, which focuses on capturing the key style of a set of images, we randomly select 20 artists each with 20 images from the WikiArt dataset [21]. For subject-driven generation, which emphasizes details of a given object, we randomly choose 30 objects from the Dreambooth dataset [24], each consisting of more than 4 images. Half of these images are utilized for training, while the other half remains untrained. This results in 10 images used for training a style and 2-6 images used for training an object, aligning with the recommended number for training in the mentioned fine-tuning methods [13, 24]. We adopt the aforementioned image selection process as the default setting.

The default model used for training is Stable-Diffusion-Model V1.4[2]. Additionally, we demonstrate the adaptability of our method to various types and versions of diffusion models, different training steps, and a larger number of training images (refer to Sec. 4.3 for more details).

As noted in previous studies [1], the presence of near-duplicate examples across different datasets complicates the differentiation between seen and unseen datasets. To address this, we define near-duplicate examples as those with Clip-similarity [11] exceeding 0.90, and we take precautions to prevent their inclusion within the WikiArt datasets (refer to Appendix F for further details).

Our objective is to apply our method CGI-DM to distinctly differentiate between images used for training (membership) and those left untrained (holdout) [8, 16]. We conduct tests across various fine-tuning scenarios, including direct Dreambooth (with prior loss), Dreambooth (without prior loss) [24], lora [13]. Specifics of the training process are outlined in Appendix H. It is noteworthy that the Dreambooth (without prior loss) scenario holds particular significance, given that the process closely mirrors direct fine-tuning.

We set Monte Carlo sampling steps (N) to 1000 and the step-wise length (α) to 2, with an overall updating of 70 within the L_2 norm by default for our method. This value (70) approximates the average distance between the partial representation of the images and the original images. The default block size for deriving partial representation is 4. For all experiments in Sec. 4.2, where our method is compared with other approaches, we utilize all 20 classes from the WikiArt Dataset and all 30 classes from the Dreambooth Dataset. Additionally, for ablation studies and experiments validating the generalization ability of our method in Sec. 4.3, Sec. 4.4 and Sec. 4.5, we randomly select five classes of images from the WikiArt Dataset.

4.1. Evaluation Metrics

As mentioned in Sec. 2.2, it is crucial for the metric to be visualizable and understandable by human beings. Therefore, upon deriving the extracted images and original input images, we calculate the visual similarity between them. Specifically, following the approach in a prior study [24], we employ Clip similarity and Dino similarity, utilizing the feature space of the Clip [11][3] and Dino models [2]. These measures have been demonstrated to align closely with human vision [11, 24].

Upon obtaining the similarity scores, we determine an optimal threshold to distinguish between membership and holdout data. Subsequently, we calculate the Accuracy (**Acc.**) and Area Under the ROC Curve (**AUC**) [8, 16] to assess the effectiveness of a method in discriminating between

[2] https://huggingface.co/CompVis/stable-diffusion

[3] We use image embeddings of Clip ViT-B/32.

Style-Driven: WikiArt Dataset												
	Dreambooth (w. prior loss)				Dreambooth (w/o. prior loss)				Lora			
	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑
Text2img	0.64	0.70	0.68	0.74	0.67	0.72	0.69	0.76	0.60	0.66	0.61	0.68
inpainting	0.67	0.71	0.71	0.74	0.71	0.75	0.77	0.82	0.59	0.59	0.60	0.59
Img2img	0.82	0.86	0.89	0.92	0.83	0.90	0.90	0.94	0.68	0.73	0.74	0.79
CGI-DM	**0.90**	**0.95**	**0.96**	**0.98**	**0.86**	**0.95**	**0.94**	**0.99**	**0.74**	**0.80**	**0.81**	**0.87**
Subject-Driven Generation: Dreambooth Dataset												
	Dreambooth (w. prior loss)				Dreambooth (w/o. prior loss)				Lora			
	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑
Text2img	0.65	0.71	0.68	0.75	0.69	0.75	0.70	0.79	0.64	0.69	0.66	0.74
inpainting	0.63	0.71	0.67	0.76	0.67	0.79	0.71	0.82	0.62	0.64	0.64	0.68
Img2img	0.67	0.77	0.73	0.82	0.75	0.81	0.81	0.87	0.65	0.75	0.68	0.76
CGI-DM	**0.84**	**0.85**	**0.90**	**0.91**	**0.89**	**0.88**	**0.94**	**0.94**	**0.76**	**0.79**	**0.82**	**0.86**

Table 1. Comparison of CGI-DM and other existing pipelines in copyright authentication for style-driven generation using WikiArt dataset [21] and for object-driven generation under Dreambooth dataset [24] under different fine-tuning methods. The Monte Carlo sampling steps (N) for CGI-DM is fixed to 1000 and the step-wise length (α) is fixed to 2 in L_2 norm. The overall budget is fixed to 70 in L_2 norm. We employ block-wise masking with a block size of 4 for removing partial information. The experimental results demonstrate that CGI-DM exhibits superior performance under all scenarios.

Diffusion Model Structures	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑
SD(v1.4)	0.93	0.96	0.97	0.98
SD(v1.5)	0.97	0.96	0.97	0.99
SD(v2.0)	0.88	0.94	0.91	0.98
AltDiffusion	0.92	0.94	0.98	0.99

Table 2. Robustness of CGI-DM to different model structures and parameters. Experiments are conducted under Dreambooth (with prior loss) fine-tuning using 5 classes of images from the WikiArt dataset. All parameters for CGI-DM are set as the default value in Tab. 1.

# of Training Images	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑
5	0.90	0.90	0.93	0.97
10	0.93	0.96	0.97	0.98
20	0.86	0.94	0.92	0.97
50	0.86	0.96	0.95	0.99

Table 3. Impact of training image number on the performance of CGI-DM on WikiArt dataset. All other parameters are set the same as those in Tab. 2.

the two, thereby evaluating its performance in authenticating copyright. We abbreviate the Acc. and AUC under similarity measured by Clip as Acc. (C) and AUC (C), and the ones under similarity measured by Dino as Acc. (D) and AUC (D).

Notably, for a given fine-tuning method, we utilize a fixed threshold to distinguish membership and holdout among different objects and styles—an approach that accounts for real-world scenarios. Generally, obtaining membership and holdout datasets beforehand, and determining an optimal threshold for each style or object, is impractical. Therefore, our threshold is set independently of different styles or objects, representing a more practical and meaningful scenario. Results for a specific threshold for each style or object are also presented in Appendix J to ensure a comprehensive discussion.

4.2. Comparison with Existing Methods

While no methods explicitly claim to be used for authenticating copyright on few-shot generation, both existing image generation [1, 31] and inpainting [33] pipelines exhibit the potential to do so. Particularly, we compare our method with three types of pipelines that could be employed for identifying infringements:

Text-to-Image Generation. Using the official Text-to-Image pipeline in diffusers[4], we generate 100 $\times$ Num images for each fine-tuned model using the prompt employed during training. Here, 'Num' represents the total number of images in the membership and holdout datasets. This process is applied individually to each image in both datasets.

Image-to-Image Generation. Leveraging the official Image-to-Image pipeline in diffusers[5], we use the training prompt to generate 100 $\times$ Num images for each fine-tuned model. The inference step is set to 50 and the img2img strength is set to 0.7 for each image in the membership or holdout dataset.

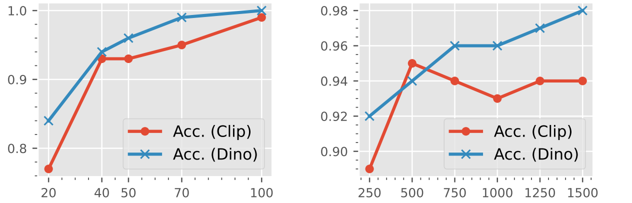
Inpainting. Employing the state-of-the-art inpainting pipeline DDNM [33], we generate 100 $\times$ Num images by masking the right half of each image. We use the prompt applied during training for inpainting, and the inference step is fixed at 50.

We set the number of generated images per input image to 100. This ensures that the time cost for both the baseline method and our method remains similar, facilitating a fair comparison (see Appendix I for more details). We define the similarity between the output images and a given image as the highest similarity achieved among all the generated images corresponding to one target image.

The comparison between CGI-DM and other pipelines is presented in Tab. 1. It is evident that CGI-DM outperforms others significantly in various few-shot generation scenarios across different datasets.

[4] https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/text2img

[5] https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/img2img



(a) CGI-DM under different training steps

(b) CGI-DM under different extraction steps

Figure 3. CGI-DM under different training steps and extraction steps. All other parameters are set the same as those in Tab. 2.

4.3. Generalization

In this section, we take a step further to test whether our method can be applied to a broader range of scenarios, including different Diffusion model structures, varying numbers of training images, and different numbers of training steps.

Different DMs. We select different versions of two distinguishable Diffusion Models: Stable Diffusion Model [23] and AltDiffusion [36], which are representative of latent-space DMs and multilingual DMs, respectively. We conduct experiments using the following versions of the two models: SDv1.4[6], SDv1.5[7], SDv2.0[8], and AltDiffusion[9].

As shown in Tab. 2, our method consistently maintains high Acc. (above 88%) and AUC (above 90%) scores across different DMs.

Number of Training Images. As the number of training images increases, the learned concept during fine-tuning becomes more intricate, thereby enhancing the difficulty of copyright authentication. To thoroughly examine how this influences performance, we conduct experiments with varying numbers of training images, the results of which are presented in Tab. 3. It is evident that our method remains almost stable despite changes in the number of training images. Specifically, CGI-DM consistently achieves Acc. around 90% in all cases, which demonstrates the robustness of CGI-DM to scenarios with various training data sizes during fine-tuning.

Training Steps. Increasing training steps is known to lead to stronger memorization of the training images, making the extraction process easier [1, 27]. Therefore, we also analyze the performance of our method under different training steps. As depicted in Fig. 3a, our method maintains a high authentication success rate (above 90% Acc.) under the Dreambooth scenario (with prior loss), with 40 training steps per image. The 40 training steps per image are notably fewer than the typically recommended 100 steps per image for training, as reported in [24]. As the number of training steps increases, the model becomes more adept at memorizing the concepts present in the given images, leading to a rise in copyright authentication accuracy for CGI-DM.

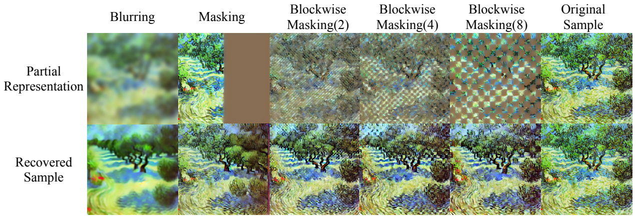


Figure 4. Visualization of different methods for removing partial information and corresponding recovered samples for CGI-DM on Van Gogh's paintings from the WikiArt dataset.

4.4. Ablation Study

4.4.1 Methods for Removing Partial Information

As mentioned earlier, our approach commences with a partial representation of the images. In this section, we explore the following techniques to eliminate partial information from the provided images:

Blurring. We employ Gaussian blurring[10], with a fixed kernel size of 16 and a blurring rate of 7.

Masking. We utilize a large mask which covers right half of the images, following previous work [1].

Block-wise Masking. We use square with different sizes (2, 4 and 8) to mask half of the region in the images.

The representation is illustrated in Fig. 4. As indicated in both Fig. 4 and Tab. 6, we observe that to achieve a match between the recovered images and the desired training samples, the masked information must be fine-grained, ensuring retention of local information in the masked image. Otherwise, the recovered content may be irrelevant to the given image. Consequently, the block-wise masking emerges as the superior choice. We also find that the block size of the mask has a negligible impact on performance.

4.4.2 Extraction Steps

One critical hyper-parameter in our algorithm is the number of Monte Carlo Sampling steps, which we denote as "Extraction steps". Notably, the time required by CGI-DM increases linearly with the increment in extraction steps. We conduct experiments with different extraction steps and present the results in Fig. 3b. The findings suggest that excessively small extraction steps fail to yield the optimal solution, leading to inadequate exploitation of the conceptual differences between the fine-tuned models and the pre-trained models. Consequently, this inadequacy results in significantly poorer extraction performance and copyright authentication. Notably, performance stabilizes when the number of steps reaches approximately 1000. Further

[6] https://huggingface.co/CompVis/stable-diffusion

[7] https://huggingface.co/runwayml/stable-diffusion-v1-5

[8] https://huggingface.co/stabilityai/stable-diffusion-2

[9] https://huggingface.co/docs/diffusers/api/pipelines/alt_diffusion

[10] https://pillow.readthedocs.io/en/stable/reference/ImageFilter.html

Defense Methods	Acc. (C)↑	Acc. (D)↑	AUC (D)↑	AUC (D)↑
No Defense	0.93	0.96	0.97	0.98
HorizationalFlip [8]	0.93	0.94	0.98	0.98
Cutout [7]	0.87	0.93	0.92	0.97
RandAugment [4]	0.83	0.85	0.86	0.92

Table 4. Performance of CGI-DM under possible defenses. All parameters are set the same as those in Tab. 2.

increasing the number of steps substantially escalates the computational cost while yielding only marginal improvements.

4.5. Performance of CGI-DM under Defenses

As highlighted in prior research [8, 16, 22], it is possible to defend against methods that leverage the loss function of a model. In line with this, we conduct experiments on CGI-DM under various defenses, including RandomHorizontalFlip [8], Cutout [7], and RandAugment [4]. The block size for Cutout is consistently fixed at 64 × 64. Notably, RandAugment serves as a strong privacy-preserving defense, at the price of a decline in generation quality [8]. The results presented in Tab. 4 underscore the consistent effectiveness of our method under various defense mechanisms. Even RandAugment, recognized for its strong defense capabilities, demonstrates limited effectiveness in reducing the efficacy of copyright authentication for CGI-DM, with reductions of approximately 10% in Acc. and 8% in AUC.

5. Related Work

Model Inversion (MI). Model Inversion (MI) Attack was first introduced by Matthew *et al.* [9]. The attack focuses on deriving the training dataset of a network based on its parameters, mostly aiming at white-box scenarios [3, 9, 10, 20, 37, 38]. Recent approaches on MI [3, 10, 38] highlight the importance of prior information during inversion, emphasizing that the proper prior of a generative model could contribute to the inversion of a classification model. Our inversion from partial representation is also inspired by such findings.

Although model inversion has been extensively researched for classification models over the years, there is no systematic research on model inversion for generative models as far as we know.

Membership Inference Attack (MIA). Membership Inference Attack [8, 16, 22, 26] centers on identifying which subset of a larger data pool is used for training. These methods generally leverage the loss function of the model in white-box scenarios where the model parameter is available, or simulate the loss in gray-box or black-box scenarios where the model parameter is not as clear.

Limited research has been conducted on the use of MIA for copyright authentication [22], possibly due to its intrinsic limitations. While MIA efficiently assesses the origin of a sample, its reliance on the model's loss function as the metric for classification restricts its visualization capabilities. Consequently, MIA-based techniques typically provide only binary outcomes, hindering their use as strong legal evidence. Further discussion is available in Appendix A.

Data Watermark on Diffusion Models. Data watermarking [5, 34, 40] entails embedding watermarks into training images, enabling post-training extraction from the generated output. This process holds potential for copyright authentication by identification of training images via extracting the embedded watermarks in the generated images. However, prior studies [5, 34, 40] show that most methods necessitate a single designated watermark across the entire diffusion model dataset. The efficacy of such an approach in few-shot generation scenarios, where only several input images correspond to a single watermark, remains uncertain. Moreover, the introduction of watermarks may suffer from image degradation [5, 40], content changes [5], and potential removal [39].

While recognizing these possible issues, it is crucial to note that watermarks and our proposed method are separate ways to protect copyright, and they do not clash. It is conceivable that in the future, these approaches could converge, combining to create a more robust copyright authentication system.

6. Conclusion

In this paper, we introduce a new method for copyright protection called copyright authentication. Our framework, CGI-DM, validates the use of training samples featuring vivid visual representation, serving as a tool for digital copyright authentication. We start by removing part of the input image. Then, using Monte Carlo sampling and PGD, we exploit the differences between the pre-trained and fine-tuned model to recover the removed information. A high similarity between the recovered samples and the original input samples suggests a potential infringement. Through experiments on WikiArt and Dreambooth datasets, we demonstrate CGI-DM's robustness and effectiveness, surpassing alternative approaches. Such experimental results show that CGI-DM is adept at providing legal evidence for art-style mimicry and unauthorized image fabrication. In conclusion, CGI-DM not only offers a robust method for infringement validation in the evolving DM landscape but also pioneers the application of gradient inversion in generative models.

7. Acknowledgements

This study benefited from partial funding provided by the Shanghai Key Laboratory of Scalable Computing and Systems, alongside support from Intel Corporation under UFunding 12679. Tao Song is the corresponding author.

References

[1] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models. In *USENIX Security*, 2023. 2, 5, 6, 7, 4

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-supervised Vision Transformers. In *ICCV*, 2021. 5

[3] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched Distributional Model Inversion Attacks. In *ICCV*, 2021. 2, 8

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. In *CVPR*, 2020. 8

[5] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models. *arXiv preprint arXiv:2306.04642*, 2023. 8, 1, 2

[6] Andrew Deck. AI-Generated Art Sparks Furious Backlash from Japan's Anime Community. https://restofworld.org/2022/ai-backlash-anime-artists/, 2022. 2

[7] Terrance DeVries and Graham W Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017. 8

[8] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are Diffusion Models Vulnerable to Membership Inference Attacks? *arXiv preprint arXiv:2302.01316*, 2023. 5, 8, 1

[9] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin dosing. In *USENIX Security*, 2014. 8

[10] Ali Hatamizadeh, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, Mona G Flores, Jan Kautz, Daguang Xu, et al. Do Gradient Inversion Attacks make Federated Learning Unsafe? *IEEE Transactions on Medical Imaging*, 42(7):2044–2056, 2023. 2, 8

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 3, 4

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 5

[14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-Based Real Image Editing With Diffusion Models. *arXiv preprint arXiv:2210.09276*, 2022. 1

[15] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided Diffusion Models for Robust Image Manipulation. In *CVPR*, 2022. 4, 2

[16] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An Efficient Membership Inference Attack for the Diffusion Model by Proximal Initialization. *arXiv preprint arXiv:2305.18355*, 2023. 5, 8, 1, 4

[17] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, XUE Zhengui, Ruhui Ma, and Haibing Guan. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. In *ICML*, 2023. 2, 4

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018. 2, 4

[19] Deborah MT. How AI Art Can Free Artists, Not Replace Them. https://medium.com/thesequence/how-ai-art-can-free-artists-not-replace-them-a23a5cb0461e, 2022. 2

[20] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking Model Inversion Attacks Against Deep Neural Networks. In *CVPR*, 2023. 8

[21] K. Nichol. Painter by Numbers, WikiArt. https://www.kaggle.com/c/painter-by-numbers, 2016. 5, 6

[22] Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box Membership Inference Attacks against Diffusion Models. *arXiv preprint arXiv:2308.06405*, 2023. 8, 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 7, 3

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, 2023. 1, 5, 6, 7

[25] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. 2

[26] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks against Machine Learning Models. In *S&P*, 2017. 8

[27] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and Mitigating Copying in Diffusion Models. *arXiv preprint arXiv:2305.20086*, 2023. 2, 7

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2020. 4, 2

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling Through Stochastic Differential Equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[30] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc Tran, and Anh Tran. Anti-DreamBooth: Protecting Users from Personalized Text-to-image Synthesis. In *ICCV*, 2023. 2

[31] James Vincent. The Scary Truth About AI Copyright Is Nobody Knows What Will Happen Next. https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data, 2022. 2, 6

[32] Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. Security and Privacy on Generative Data in AIGC: A Survey. *arXiv preprint arXiv:2309.09435*, 2023. 2

[33] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *ICLR*, 2022. 6

[34] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. *arXiv preprint arXiv:2305.20030*, 2023. 8, 1, 2

[35] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion Probabilistic Modeling for Video Generation. *arXiv preprint arXiv:2203.09481*, 2022. 1

[36] Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. AltDiffusion: A Multilingual Text-to-Image Diffusion Model. *arXiv preprint arXiv:2308.09991*, 2023. 7

[37] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See Through Gradients: Image Batch Recovery via Gradinversion. In *CVPR*, 2021. 8

[38] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The Secret Revealer: Generative Model-inversion Attacks against Deep Neural Networks. In *CVPR*, 2020. 2, 8

[39] Xuandong Zhao, Kexun Zhang, Yu-Xiang Wang, and Lei Li. Generative Autoencoders as Watermark Attackers: Analyses of Vulnerabilities and Threats. *arXiv preprint arXiv:2306.01953*, 2023. 8

[40] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A Recipe for Watermarking Diffusion Models. *arXiv preprint arXiv:2303.10137*, 2023. 8, 1, 2