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

Supplementary Material

A. Legal Issues

A.1. Copyright Infringement for Artificial Intelligence-generated Content (AIGC)

In traditional copyright infringement cases, the initial process of determining whether an image has been "referenced" in the creation of another image is relatively straightforward. However, the subsequent legal judgment, which involves quantifying the extent of similarity and deciding if it amounts to infringement, presents more significant challenges. This necessitates an examination of substantial similarity aspects, including style and composition, while also considering the originality of the work in question and any authorized use. This stage poses challenges due to the requirement for a nuanced understanding of the intricate aspects of copyright law.

In the traditional context, particularly in the field of painting, the acceptance of visual and comprehensible evidence as a key element for legal decisions has evolved over time. Once established, a high degree of similarity found in such evidence often led to findings of infringement. For instance, an artist was fined a huge amount of money for plagiarizing artworks[11]. In this case, visually similar paintings serve as evidence of infringement, underscoring the significance of visual proof in traditional copyright cases. This sets a precedent, highlighting the need for adapting similar methodologies to address the unique challenges in the AIGC realm.

In the era of AIGC, the stark absence of widely-accepted factual evidence, in contrast to traditional contexts, is evident. The intricate nature of AI algorithms complicates the task of establishing direct references. This has led to a significant shift in legal regulations, moving from focusing primarily on direct infringement judgments to tackling anti-unfair competition concerns. This transition is a response to the lack of clear, visual, and easily interpretable evidence, reflecting the adjustments current regulations are undergoing to accommodate the nuances of AIGC[12].

Addressing these complexities in AIGC demands the cultivation of a widely-recognized factual understanding of what constitutes a "reference." Here, the importance of robust copyright authentication tools becomes apparent. These tools are crucial for not only establishing facts but also for making them visually comprehensible and publicly acceptable. The recent lawsuit involving "Stable Diffusion"[13] brings to light the current reliance on visual evidence in copyright lawsuits. It also, however, points to the inadequacy of these methods due to the limited similarity between AI-generated output images and the original training data. Our work with CGI-DM focuses on overcoming this hurdle, offering a method to aid in substantiating infringement claims with higher visual similarity.

	Need Preprocessing?	Defense	Accuracy	Visualizability
MIA [8, 16]	✗	**Hard**	**High**	Poor
Data Watermark [5, 34, 40]	✓	Uncertain	Uncertain	Moderate
Copyright Authentication	✗	**Hard**	**High**	**Strong**

Table 5. Comparative Overview of Post-cautionary Copyright Protection Methods. Bold text highlights advantageous features, while plain text indicates areas where these features are less pronounced.

In the short term, these copyright authentication tools offer valuable assistance in current lawsuits against AIGC. While they may not be crucial at this juncture, their role in establishing necessary visual and comprehensible evidence is significant as we adapt our legal frameworks to the challenges posed by AIGC. In the long term, their importance grows, as establishing well-accepted facts becomes paramount for the effective adjudication of AIGC-related copyright issues. Tools like the proposed CGI-DM method are increasingly vital in this extended context, which enable a shift in focus from merely contending with anti-unfair competition to directly addressing copyright infringement. This reorientation is vital for ensuring stronger protection of the intellectual property rights of human artists, safeguarding their work against the advancing tide of AIGC.

In conclusion, the development and acceptance of robust validation tools are essential in shaping the future of copyright law enforcement in the context of AIGC. Reflecting on the historical trajectory of how visual evidence gradually became integral to legal judgment in traditional contexts, we see a clear need for a similar evolution in the AIGC landscape. This historical perspective reinforces the importance of societal acceptance in the legal adjudication process and underscores the necessity of adapting these principles to the emerging challenges of AIGC.

[11] https://www.caixinglobal.com/2023-10-25/chinese-artist-fined-5-million-yuan-in-landmark-case-for-plagiarizing-foreign-work-102120348.html

[12] https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/

[13] https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929

A.2. More Discussion of Post-Cautionary Methods for Copyright Protection

This section presents an in-depth analysis of various post-caution methods for copyright protection. The summary table in Tab. 5 illustrates that, in terms of preprocessing, defense, accuracy, and visualizability, copyright authentication emerges as a superior choice.

Membership Inference Attack(MIA). There are only a limited number of studies that explore the use of MIA for infringement validation [22], possibly due to the inherent limitations of this method. MIA primarily assesses the probability of a sample being part of the training dataset, but its dependency on the model's loss function results in outcomes that are challenging to visualize. This limitation restricts MIA's use in legal settings, as it provides only binary 'yes' or 'no' outputs, lacking the detailed nuance needed for legal evidence. Additionally, these binary outcomes, being dependent on the model's loss function, deviate from human visual perception, thus limiting their reliability and applicability in the process of legal judgment.

Data Watermark. Data watermarking typically requires preprocessing on training images [5, 34, 40]. This method's limitations include potential degradation in image quality and the feasibility of watermark removal. Furthermore, its effectiveness is predominantly aligned with current Digital Models (DMs) and may not extend to future DMs or other generative models. The intrinsic design of data watermarking implies permanent security once implemented. Besides, its performance under defense and its accuracy remain uncertain under few-shot generation scenarios. The method offers moderate visualizability since the embedded watermark can be discerned in the output, making it more comprehensible than MIA.

Copyright Authentication. At present, copyright authentication requires no preprocessing and demonstrates robustness in defense scenarios, achieving high accuracy and strong visualizability. These attributes make it an effective tool for providing legal support in cases of copyright infringement.

It is noteworthy that the methods aforementioned do not mutually exclude each other. In the long term, they could potentially converge to form a comprehensive system for validating copyright infringement.

B. Proof Details

B.1. Forward DDIM

It has recently been found that the diffusion process from x_T to x_1 can be directly obtained by leveraging x_0 [28]. The reverse of this formulation can even lead to a predicted forward process using the model weight of the diffusion model [15]:

$$\begin{aligned} p_\theta(x_{t+1}|x_t) =& \mathcal{N}(x_{t+1}; \sqrt{\alpha_{t+1}} f_\theta(x_t, t) + \\ & \sqrt{1-\alpha_{t+1}}\epsilon_\theta(x_t, t), \sigma_t^2 \mathbf{I}), \end{aligned} \tag{10}$$

where $f_\theta(x_t, t)$ is the predicted x_0 defined as:$f_\theta(x_t, t) := \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t,t)}{\sqrt{\alpha_t}}$.

B.2. Markov Chain Expanding and Expectation Term Transformation

KL divergence $D_{\mathrm{KL}}(p_{\theta'}(\overline{x}'_{1:T}|\overline{x}'_0)||p_\theta(\overline{x}'_{1:T}|\overline{x}'_0))$in Eq. (4) can be extended to the expectation of the logarithm of the division of two probability functions, given the pre-trained model and the fine-tuned model.

$$\begin{aligned} &\delta := \arg\max_\delta \mathbb{E}_{p_{\theta'}(x'_{1:T}|x'_0)} \log \frac{p_{\theta'}(x'_{1:T}|x'_0)}{p_\theta(x'_{1:T}|x'_0)}, \\ &\text{where } x_0 \sim q(x_0), x'_0 = x_0 + \delta. \end{aligned} \tag{11}$$

We isolate a single term within the logarithm to simplify the notation and defer the division of two probability terms to a later step.

$$\begin{aligned} &\max_\delta \mathbb{E}_{p_{\theta'}(x'_{1:T}|x'_0)} \log p_{\theta'}(x'_{1:T}|x'_0) \\ =& \max_\delta \mathbb{E}_{\prod_{t=0}^{T-1} p_{\theta'}(x'_{t+1}|x'_t)} \log \prod_{t=0}^{T-1} p_{\theta'}(x'_{t+1}|x'_t) \\ =& \max_\delta \sum_{t=0}^{T-1} \mathbb{E}_{\prod_{k=0}^{T-1} p_{\theta'}(x'_{k+1}|x'_k)} \log p_{\theta'}(x'_{t+1}|x'_t) \\ =& \max_\delta \sum_{t=0}^{T-1} \mathbb{E}_{p_{\theta'}(x'_{t+1}|x'_t)} \log p_{\theta'}(x'_{t+1}|x'_t). \end{aligned} \tag{12}$$

The last equation holds true as the possibility function $p(x'_{t+1}|x'_t)$ is a Markov chain and is therefore independent of any possibility that precedes time t. It is important to note that, for a well-trained model on data points $x_0 \sim q(x_0)$, it should effectively simulate the real prior distribution. In other words, $p_{\theta'}(x'_{t+1}|x'_t)$ can be approximated as $q(x'_{t+1}|x'_t)$. Notably, this approximation only holds for the fine-tuned model θ' as it is the only model trained on image x_0, and it does not hold true for the pre-trained model θ. Based on this approximation, we can modify our target to:

$$\begin{aligned} &\max_\delta \sum_{t=0}^{T-1} \mathbb{E}_{p_{\theta'}(x'_{t+1}|x'_t)} \log p_{\theta'}(x'_{t+1}|x'_t) \\ \approx& \max_\delta \sum_{t=0}^{T-1} \mathbb{E}_{q(x'_{t+1}|x'_t)} \log p_{\theta'}(x'_{t+1}|x'_t). \end{aligned} \tag{13}$$

B.3. Approximation of the loss function

Given Equation (10) in Section 2.1, we can demonstrate that the discrepancy between the initial image x_0 and its predicted counterpart $x_0 := f_\theta(x_t, t)$ is indeed linked to the disparity between the forecasted noise $\epsilon_\theta(x_t, t)$ at time t with respect to the true noise ε_t:

$$\begin{aligned} & x_0 - f_\theta(x_t, t) \\ =& x_0 - \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \\ =& x_0 - \frac{\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\varepsilon_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \\ =& \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}(\varepsilon_t - \epsilon_\theta(x_t, t)). \end{aligned} \tag{14}$$

Based on this finding, we are now able to establish the validity of the approximation presented in Equation (6):

$$\begin{aligned} & \|x'_{t+1} - \mu_{p_\theta(x'_{t+1}|x'_t)}\|^2 \\ =& \|\sqrt{\frac{\alpha_{t+1}}{\alpha_t}}x'_t + \sqrt{1-\frac{\alpha_{t+1}}{\alpha_t}}\varepsilon_{t+1} - \sqrt{\alpha_{t+1}}f_\theta(x'_t, t) - \\ & \sqrt{1-\alpha_{t+1}}\epsilon_\theta(x'_t, t)\|^2 \\ =& \|\sqrt{\alpha_{t+1}}x'_0 + \sqrt{(1-\alpha_t)\frac{\alpha_{t+1}}{\alpha_t}}\varepsilon_t + \sqrt{1\frac{\alpha_{t+1}}{\alpha_t}}\varepsilon_{t+1} - \\ & \sqrt{\alpha_{t+1}}f_\theta(x'_t, t) - \sqrt{1-\alpha_{t+1}}\epsilon_\theta(x'_t, t)\|^2 \\ =& \|\sqrt{\alpha_{t+1}}(x'_0 - f_\theta(x'_t, t)) - \sqrt{1-\alpha_{t+1}}(\epsilon_\theta(x'_t, t) - \varepsilon_t) + \\ & (\sqrt{(1-\alpha_t)\frac{\alpha_{t+1}}{\alpha_t}} - \sqrt{1-\alpha_{t+1}})\varepsilon_t + \sqrt{1-\frac{\alpha_{t+1}}{\alpha_t}}\varepsilon_{t+1}\|^2 \\ \approx& \|\sqrt{\alpha_{t+1}}(x'_0 - f_\theta(x'_t, t)) - \sqrt{1-\alpha_{t+1}}(\epsilon_\theta(x'_t, t) - \varepsilon_t)\|^2. \end{aligned} \tag{15}$$

The final equation is valid due to the approximations of both $\sqrt{1-\frac{\alpha_{t+1}}{\alpha_t}}$ and $(\sqrt{(1-\alpha_t)\frac{\alpha_{t+1}}{\alpha_t}} - \sqrt{1-\alpha_{t+1}})$ being close to 0. Hence, we obtain:

$$\begin{aligned} & \|x'_{t+1} - \mu_{p_\theta(x'_{t+1}|x'_t)}\|^2 \\ \approx& \|\sqrt{\alpha_{t+1}}(x'_0 - f_\theta(x'_t, t)) - \sqrt{1-\alpha_{t+1}}(\epsilon_\theta(x'_t, t) - \varepsilon_t)\|^2 \\ =& -(\frac{\sqrt{1-\alpha_t}\sqrt{\alpha_{t+1}}}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t+1}})^2\|\varepsilon_t - \epsilon_\theta(x_t, t)\|^2 \end{aligned} \tag{16}$$

C. Direct Gradient Inversion (GI) Leads to Zero-information Extracted Result

We observe that inverting the gradient solely based on the fine-tuned model parameterized by θ' resulted in a zero-information-extracted outcome. Specifically, for the partial representation $\overline{x_0}$ of an image x_0, we optimize it based on the following target:

$$\begin{aligned} x'_0 &= \arg\min_{x'_0} p_{\theta'}(x'_{1:T}|x'_0) \\ &\approx \arg\min_{x'_0} \mathbb{E}_{t,\epsilon_t\sim\mathcal{N}(0,1)}\|\varepsilon_t - \epsilon_{\theta'}(x'_t, t)\|^2. \end{aligned} \tag{17}$$

We omit the proof, as it closely resembles the proof mentioned in Section 3.2. We employ a similar optimization method utilizing Monte Carlo Sampling and PGD attack, as discussed in Section 3.3. As illustrated in Figure 5, the extracted images appear blurred, with no useful information discernible. This is primarily attributed to the fact that the diffusion model functions as a robust noise predictor, resulting in the most probable image according to the model being a smooth one with low information content, thus enabling the model to easily predict the introduced noise.

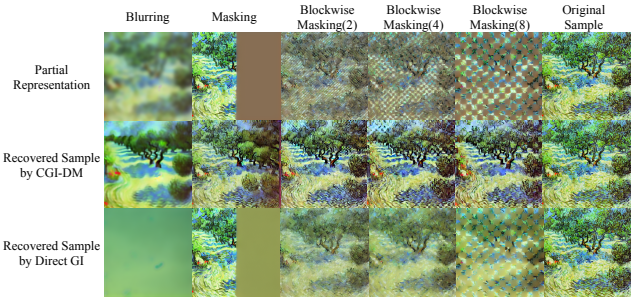


Figure 5. Comparison of CGI-DM and direct GI under different partial representations. We can find that direct GI does not recover any semantic information of the images.

Removing Methods	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑
Blurring	0.65	0.80	0.69	0.86
Masking	0.72	0.73	0.76	0.75
Block-wise Masking(2)	0.93	**0.97**	**0.97**	0.99
Block-wise Masking(4)	0.93	0.96	**0.97**	0.98
Block-wise Masking(8)	**0.94**	**0.97**	**0.97**	**1.00**

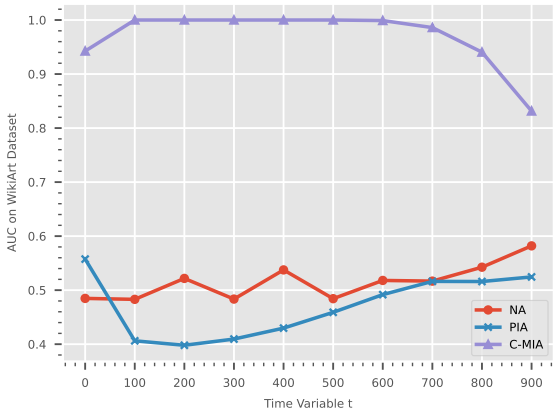
Table 6. Influence of different approaches for removing partial information on CGI-DM. All other parameters are set as the same Tab. 2.

D. CGI-DM on Latent Diffusion Model (LDM)

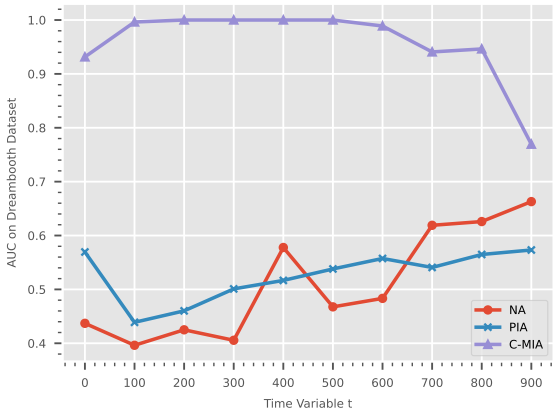
The key distinction between the latent diffusion model (LDM) model [23] and others lies in the occurrence of the diffusion process within the latent space. Consequently, our algorithm, CGI-DM, for LDM, closely resembles CGI-DM for DM, with the exception that the partial information removal and the consideration of predicted noise differences are both based on the latent space of a LDM. Refer to Algorithm 2 for more details.

E. Partial Information Removal in CGI-DM

Quantilized results are detailed in Tab. 6.



(a) MIA methods under WikiArt Dataset



(b) MIA methods under Dreambooth Dataset

Figure 6. Comparison of C-MIA and other MIA methods under two representative datasets under few-shot generation scenarios.

Algorithm 2 CGI-DM for LDMs

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

F. Removal of near-duplicate samples

Following the precedent set by previous research, we classify samples with a clip similarity exceeding 0.90 as near-duplicates [1]. Within the WikiArt dataset, we eliminate near-duplicate samples until only one copy remains. Given that the utilization of any of these near-duplicate samples already constitutes a violation of the provided data, the presence of such samples in both the training dataset and other datasets would result in an impractically low assessment of the effectiveness of the infringement validation algorithm.

G. Conceptual Difference Exploiting for MIA (C-MIA)

The definition of conceptual differences and the provided proof in Sec. 3.2 isn't limited to copyright authentication, which has strong effects in MIA under few-shot generation scenarios. In detail, for pre-trained model θ, fine-tuned mode θ' and a given sample x, we leverage the KL divergence of the probability of x between the two models:

$$D_{\mathrm{KL}}(p_{\theta'}(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0)) \approx \mathbb{E}_{t,\varepsilon_t\sim\mathcal{N}(0,1)} \underbrace{\|\varepsilon_t - \epsilon_\theta(x_t,t)\|^2 - \|\varepsilon_t - \epsilon_{\theta'}(x_t,t)\|^2}_{L_{tar}(\theta,\theta',x_0,t,\varepsilon_t)}. \tag{18}$$

We omit details for this approximation as it mainly follows the provided proof in Appendix B and Sec. 3. We then utilize such divergence under one given time t and sampled noise ε_t, denoted as $L_{tar}(\theta, \theta', x_0, t, \varepsilon_t)$, directly for MIA. We name this method as "C-MIA". We experiment under both WikiArt dataset and Dreambooth dataset, with settings aligned to the one mentioned in Tab. 1. Our method is compared to two representative MIA methods: Naive Attack (NA) [16] and Proximal Initialization Attack (PIA) [16]. The performance comparison of these methods is shown in Fig. 6, where we observe consistent outperformance by C-MIA across various time variable choices and datasets.

H. Finetuning Details

The details of the parameters in the fine-tuning methods are presented below. We use Num to represent the number of images utilized for training

- **Dreambooth (With Prior)**: We use the training script

Style Transferring: Wikiart Dataset												
	DreamBooth(w. prior loss)				DreamBooth(w/o. prior loss)				Lora			
	ASR(Clip)↑	ASR(Dino)↑	AUC(Clip)↑	AUC(Dino)↑	ASR(Clip)↑	ASR(Dino)↑	AUC(Clip)↑	AUC(Dino)↑	ASR(Clip)↑	ASR(Dino)↑	AUC(Clip)↑	AUC(Dino)↑
Text2img	0.72	0.77	0.68	0.75	0.74	0.80	0.71	0.80	0.71	0.74	0.63	0.70
inpainting	0.73	0.78	0.72	0.76	0.77	0.84	0.77	0.85	0.67	0.67	0.61	0.60
Img2img	0.87	0.94	0.88	0.95	0.91	0.95	0.92	0.97	0.77	0.80	0.76	0.80
CGI-DM	**0.96**	**0.98**	**0.97**	**0.99**	**0.96**	**0.99**	**0.97**	**1.00**	**0.82**	**0.90**	**0.83**	**0.93**
Subject-Driven Generation: DreamBooth Dataset												
	DreamBooth(w. prior loss)				DreamBooth(w/o. prior loss)				Lora			
	ASR(Clip)↑	ASR(Dino)↑	AUC(Clip)↑	AUC(Dino)↑	ASR(Clip)↑	ASR(Dino)↑	AUC(Clip)↑	AUC(Dino)↑	ASR(Clip)↑	ASR(Dino)↑	AUC(Clip)↑	AUC(Dino)↑
Text2img	0.81	0.89	0.72	0.85	0.82	0.92	0.78	0.90	0.77	0.90	0.68	0.85
inpainting	0.79	0.87	0.70	0.83	0.81	0.91	0.73	0.90	0.76	0.84	0.65	0.77
Img2img	0.92	0.95	0.78	**0.94**	0.92	0.98	0.90	0.98	0.81	**0.93**	0.73	0.90
CGI-DM	**0.97**	**0.96**	**0.96**	**0.94**	**0.96**	**0.99**	**0.96**	**1.00**	**0.92**	**0.93**	**0.90**	**0.91**

Table 7. Comparison of CGI-DM and other existing pipelines under separate thresholds. All other settings are fixed the same as in Tab. 1. The experimental results demonstrate that CGI-DM also exhibits superior performance under separate thresholds.

	Inference Steps/output img	Number of output imgs/Input Img	Overall inference steps	Need backward?	Time costs(min)/Input Img	Overall Time costs(min)
Text2img	50	100	5000	✗	10.36	207.30
Inpainting	50	100	5000	✗	9.19	183.89
Img2img	35	100	3500	✗	6.45	129.03
CGI-DM	1000	1	1000	✓	5.68	113.67

Table 8. Time costs comparison of CGI-DM with other methods. We experiment under Dreambooth (with prior loss) for one classes in WikiArt-dataset for three times and report the mean time costs here.

provided by Diffusers[14]. Both the text encoder and the U-Net are fine-tuned during the training process. By default, the number of training steps is set to $50 \times \text{Num}$, with a learning rate of 2×10^{-6}. The batch size is set to 1, and the number of class images used for computing the prior loss is $50 \times \text{Num}$ by default. The prior loss weight remains fixed at 1.0. For the WikiArt dataset, the instance prompt is "sks style", and the class prompt is "art style". For the Dreambooth dataset, the instance prompt is "sks object", and the class prompt is "an object".

- **Dreambooth (No Prior)**: We use the training script provided by Diffusers[14]. All default parameters remain the same as in the case of Dreambooth (With Prior), except for the exclusion of the prior loss. Notably, this adjustment will lead to training that closely resembles direct fine-tuning.
- **Lora**: We use the training script provided by Diffusers[15]. All default parameters remain consistent with the case in Dreambooth (No Prior), with the exception of the learning rate and training steps, which are adjusted to 1×10^{-4} and to $100 \times \text{Num}$, respectively.

I. Time costs

In this section, we present the time costs for overall validation and per-image validation for both our method and the compared pipelines. As demonstrated in Table 8, the time costs for the baseline methods are approximately equal to or greater than ours under $100 \times \text{Num}$ generated images, ensuring a fair comparison where all methods require a similar amount of time.

We further show the performance of best baseline methods (Img2img) with increasing number of generated images in Table 7, where we can observe that $100 \times \text{Num}$ is near saturated.

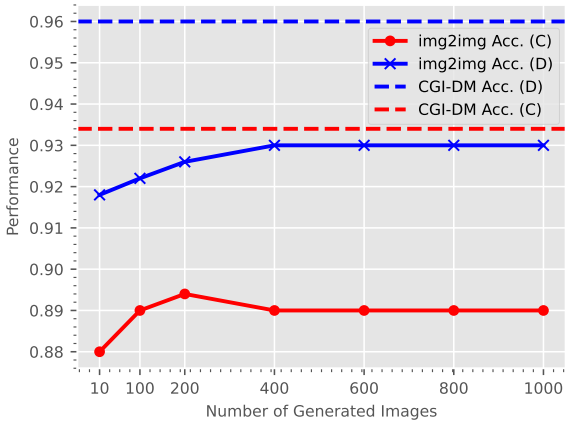


Figure 7. Trend of img2img baseline. Experimented under 5 subsets of WikiArt dataset.

J. Experiment Result under Separate Threshold

The default experiment setting employs a universal threshold across various styles or objects. In this section, we present results obtained with specific thresholds for each style or object. As depicted in Tab. 7, CGI-DM also con-

[14]https://github.com/huggingface/diffusers/blob/main/examples/dreambooth/train_dreambooth.py

[15]https://github.com/huggingface/diffusers/blob/main/examples/dreambooth/train_dreambooth_lora.py

sistently outperforms other methods in this configuration.

K. Visualization

We present visualization of CGI-DM in Fig. 8, Fig. 9 and Fig. 10. The experimental setup aligns with the one described in Tab. 1.

L. No-reference Visual Metrics

We show results[16] in Tab. 9 with BRISQUE[17] and CLIP-IQA,[18] where a similar trend can be observed. It can be seen as a stricter measurement align with human vision and can be helpful for overall evaluations.

	Acc. (C)↑	AUC (C)↑	Acc. (BRISQUE)↑	AUC (BRISQUE)↑	Acc. (CLIP_IQA)↑	AUC (CLIP_IQA)↑
CGI-DM	**0.90**	**0.96**	**0.66**	**0.69**	**0.73**	**0.79**
Img2img	0.82	0.89	0.57	0.54	0.52	0.50
inpainting	0.67	0.71	0.55	0.55	0.52	0.50
Text2img	0.64	0.68	0.50	0.50	0.50	0.50

Table 9. CGI-DM v.s. baseline under no-reference metrics.

M. More dicussion with Inpainting Baseline

The standard approach to inpainting employs half masking techniques. We also delve into inpainting methods that utilize block-wise masking, akin to the CGI-DM process. We include the results in Tab. 10. Experimental setup matches Tab. 1 in line 402.

	Acc. (C)↑	Acc. (D)↑	AUC (C)↑	AUC (D)↑
CGI-DM	**0.90**	**0.95**	**0.96**	**0.98**
inpainting with half masking	0.67	0.71	0.71	0.74
inpainting with block-wise masking (4)	0.71	0.80	0.77	0.87

Table 10. CGI-DM v.s. inpainting under block-wise masking(4).

[16] Experimental setup matches Tab. 1 in line 402.

[17] No-Reference Image Quality Assessment in the Spatial Domain (2012).

[18] Exploring CLIP for Assessing the Look and Feel of Images (2022).

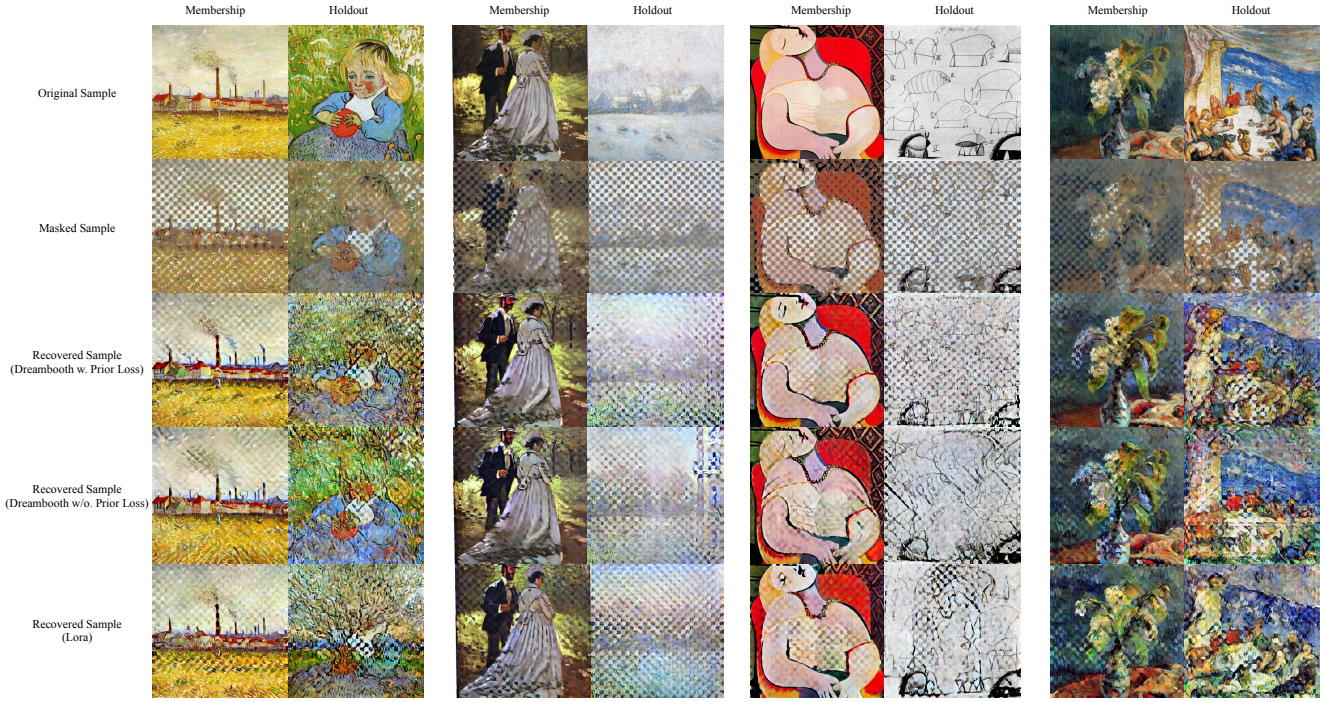


Figure 8. Visualization of CGI-DM for membership and holdout data under 4 classes of WikiArt dataset. From left to right: Vincent Willem van Gogh, Claude Monet, Pablo Ruiz Picasso and Paul Gauguin.

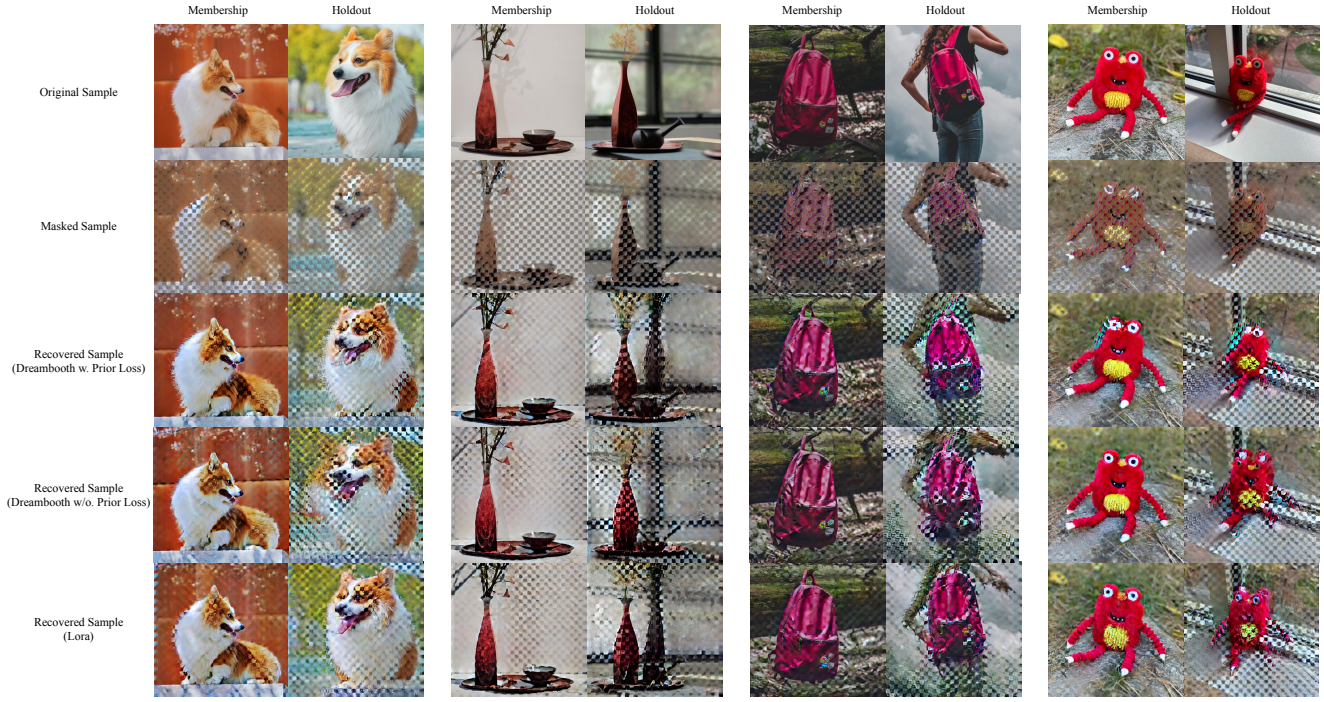


Figure 9. Visualization of CGI-DM for membership and holdout data under 4 classes of Dreambooth dataset. From left to right: dog, vase, backpack and monster_toy.

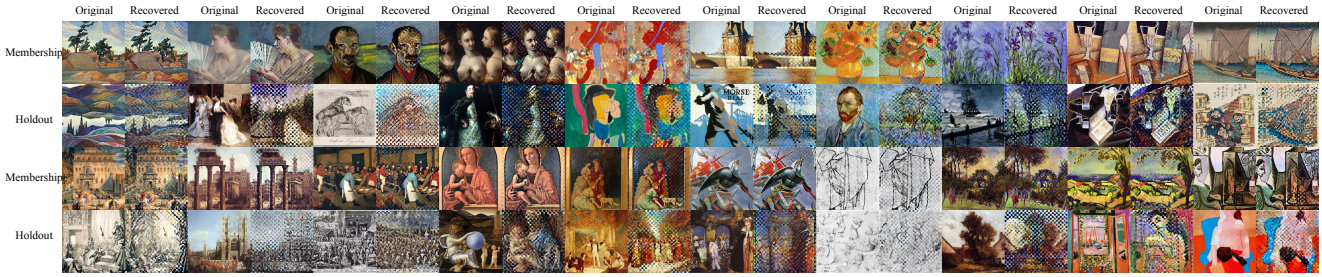


Figure 10. Visualization of CGI-DM for membership and holdout data from 20 different artists in WikiArt dataset under the default setting in against Dreambooth (with prior loss). The figure is under high resolution and can be zoomed for more details.

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