
How to Trace Latent Generative Model Generated Images without Artificial Watermark?

Zhenting Wang¹ Vikash Sehwal² Chen Chen² Lingjuan Lyu² Dimitris N. Metaxas¹ Shiqing Ma³

Abstract

Latent generative models (e.g., Stable Diffusion) have become more and more popular, but concerns have arisen regarding potential misuse related to images generated by these models. It is, therefore, necessary to analyze the origin of images by inferring if a particular image was generated by a specific latent generative model. Most existing methods (e.g., image watermark and model fingerprinting) require extra steps during training or generation. These requirements restrict their usage on the generated images without such extra operations, and the extra required operations might compromise the quality of the generated images. In this work, we ask whether it is possible to *effectively and efficiently* trace the images generated by a specific latent generative model without the aforementioned requirements. To study this problem, we design a latent inversion based method called LATENTTRACER to trace the generated images of the inspected model by checking if the examined images can be well-reconstructed with an inverted latent input. We leverage gradient based latent inversion and identify a encoder-based initialization critical to the success of our approach. Our experiments on the state-of-the-art latent generative models, such as Stable Diffusion, show that our method can distinguish the images generated by the inspected model and other images with a high accuracy and efficiency. Our findings suggest the intriguing possibility that today’s latent generative generated images are naturally watermarked by the decoder used in the source models. Code: <https://github.com/ZhentingWang/LatentTracer>.

Work partially done during Zhenting Wang’s internship at Sony AI. ¹Rutgers University ²Sony AI ³University of Massachusetts at Amherst. Correspondence to: Lingjuan Lyu <Lingjuan.Lv@sony.com>.

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

1. Introduction

Recently, latent generative models (Rombach et al., 2022) have attracted significant attention and showcased outstanding capabilities in generating a wide range of high-resolution images with surprising quality. Many state-of-the-art image generation models belong to *latent generative models*, such as DALL-E 3 (Betker et al., 2023) from OpenAI, Parti (Yu et al., 2022a) from Google, and Stable Diffusion (Rombach et al., 2022) from Stability AI. These models allow for achieving a near-optimal point between reducing computing complexity and preserving visual details, greatly boosting the efficiency in both training and generation phase. Among them, Stable Diffusion is the most widely-used, which has already gained more than 10 million users¹.

As latent generative models become more prevalent, the issues surrounding their potential for misuse are becoming increasingly important (Schramowski et al., 2023; Wang et al., 2023c; Pan et al., 2024; Liu et al., 2024; Wen et al., 2023b; Chen et al., 2023). For example, malicious users may use the latent generative models to generate and distribute images containing inappropriate concepts such as “sexual”, “drug use”, “weapons”, and “child abuse” (Schramowski et al., 2023). AI-powered plagiarism (Francke & Bennett, 2019) and IP (intellectual property) infringement problem surrounding the latent generative models are also important issues. For instance, users may synthesize high-quality images using one company’s or open-sourced latent generative models and then dishonestly present them as their own original artwork (e.g., photographs and paintings) to gain recognition and reputation, which is harmful to society and may cause a series of IP problems. Consequently, it’s crucial to be able to trace the source of images generated by latent generative models, i.e., determining if a certain image was produced by a specific model.

There are several existing methods for tracing the source of the images. Watermarking-based methods (Luo et al., 2009; Pereira & Pun, 2000; Tancik et al., 2020) typically add watermark into the images and the images from specific origins can be identified via analyzing if the particular watermark is inserted in the images or not. Classification-based

¹<https://journal.everypixel.com/ai-image-statistics>

approaches (Sha et al., 2022) train multi-classes classifiers where each class corresponds to a specific origin (source model). Another set of methods inject fingerprints into the model during training (Yu et al., 2019; 2021b) or by modifying the architectures of the models (Yu et al., 2022b), so that the images generated by the injected models will contain the fingerprinting and they can be detected by the fingerprinting decoders held by the model owner. All the above-listed methods share the limitation that requires extra steps during the training or generation phase, restricting their usage on the generated images without such operations. Also, many of the extra required operations might compromise the quality of the generated images. In addition, there is an increasing number of proposed attacks specifically targeting artificial watermarks, such as watermark stealing attacks (Jovanović et al., 2024) and watermark forgery (Wang et al., 2021). The usages of the artificial watermarks itself may also include the vulnerabilities.

In this paper, we investigate whether it is possible to trace the images generated by a specific latent generative model without the aforementioned requirements such as adding artificial watermarks during the generation (Tancik et al., 2020; Wen et al., 2023a) and injecting fingerprintings during training (Yu et al., 2021b; 2022b; Fernandez et al., 2023). Inspired by Wang et al. (2023d), we find that the input reverse-engineering based method is a promising way to achieve *alteration-free origin attribution*. Thus, we develop a latent inversion based method to trace the generated images of the inspected model by checking if the examined images can be well-reconstructed with an inverted latent input. In detail, it works by reverse-engineering the latent input of the decoder in the inspected model for each examined image, and the examined image is considered as the generated image of the inspected model if the distance between the reconstructed image and the examined image is smaller than a pre-computed threshold.

We observe that directly using the gradient-based optimization approach to invert the latent input suffers from *sub-optimal effectiveness and low efficiency* on the state-of-the-art text-to-image latent generative models. We then find that the main reason for this phenomenon is the sub-optimal initialization (i.e., the randomly sampled starting point in the optimization process typically has a large distance to the ground-truth input). To solve this problem, we propose a strategy for finding a better starting point in optimization by exploiting the invertibility of the autoencoder, which is a main component in latent generative models. In particular, we leverage the latent projection of the given image by the encoder as an initialization. Our experiments on the state-of-the-art latent generative models (e.g., Stable Diffusion (Rombach et al., 2022) and Kandinsky (Razzhigaev et al., 2023)) show that our method is *highly effective and more efficient* than existing methods for tracing the images

generated by the inspected model in the alteration-free manner. (Figure 5 shows some examples). Our results also reflect that the generated images of a specific model are naturally watermarked by the decoder module, which is essentially exploited by our method to enable alteration-free origin attribution. Our contributions are summarized as follows: ① We propose a new alteration-free inversion-based origin attribution method (called LATENTTRACER) designed for latent generative models which does not require additional operations on the model’s training and generation phase. ② We evaluate our method on the state-of-the-art latent generative models, such as Stable Diffusion (Rombach et al., 2022) and Kandinsky (Razzhigaev et al., 2023). The results show that our method is more *effective and efficient* than existing alteration-free methods. As a result, our approach can even correctly identify the source from different versions of the Stable Diffusion models. ③ We demonstrate the generalizability of our approach to both diffusion and autoregressive models, and show effectiveness of our approach against both discrete and continuous autoencoders. Our code can be found at <https://github.com/ZhentingWang/LatentTracer>.

2. Related Work

Latent Generative Models. Significant progress in the image synthesis task has been made with the advent of latent generative models (Rombach et al., 2022; Gu et al., 2022; Luo et al., 2023). The latent generative models leverage the features and visual patterns learned by the autoencoder, which can reduce the dimensionality of the samples in the generation process and help the models generate highly detailed images. Such latent-based architecture enable reaching a near-optimal balance between enhancing efficiency and preserving effectiveness. A prime example of these models is the Stable Diffusion, which utilizes a diffusion process within a latent space derived from an autoencoder.

Detection of AI-generated Images. Detecting images generated by generative models (e.g., deepfake images (Mirsky & Lee, 2021; Wu et al., 2024)) has become more and more important due to the increasing concern about the potential misuse of these generated images (Kietzmann et al., 2020; Flynn et al., 2021; 2022; Whittaker et al., 2020; Partadiredja et al., 2020). Many of the existing approaches (Frank et al., 2020; Dolhansky et al., 2020; Wang et al., 2020; Zhao et al., 2021; Corvi et al., 2023) frame this problem as a binary classification task, which aims at distinguishing between synthetic and authentic images. These methods leverage the artifacts such as frequency signals (Frank et al., 2020; Durall et al., 2019; 2020; Jeong et al., 2022) and texture patterns (Liu et al., 2020) as the key features to solve this binary classification problem. There are also existing works, such as DIRE (Wang et al., 2023b) which focus on differentiating

between real images and diffusion-generated images (i.e., images generated by all diffusion models). Although these techniques have shown promise in identifying AI-generated images, they lack the ability to determine whether a specific image was created by a particular image generation model.

Origin Attribution of Generated Images. Many techniques have been developed to trace the origins of images, including watermark-based methods (Swanson et al., 1996; Luo et al., 2009; Pereira & Pun, 2000; Tancik et al., 2020; Wen et al., 2023a), classification-based approaches (Sha et al., 2022), and model fingerprinting methods (Yu et al., 2019; 2021b; 2022b; Fernandez et al., 2023; Nie et al., 2023). However, these methods are limited by the need for additional steps either in the training or image generation stages, which are not feasible for images produced without these processes. **In contrast, our method does not have such limitations.** Recent research Wang et al. (2023d) shows that the origin attribution without the requirements on additional steps can be achieved by reverse-engineering the input for the model. However, it suffers from sub-optimal effectiveness and low efficiency on the state-of-the-art latent generative models, especially in the cases for distinguishing the generated images of the inspected large latent generative model and that of other models having similar architectures. Another set of attribution methods (Albright & McCloskey, 2019; Zhang et al., 2020) focus on finding the source model of a specific image given a set of suspicious candidate models. These methods rely on the assumptions that the inspector has the white-box access to every model in the candidate set, and the inspected image must originate from one of these models. By contrast, our method does not have such assumptions.

Input Inversion for Generative Models. Previous works on generative model inversion techniques mainly focus on image editing applications (Karras et al., 2020; Jahanian et al., 2019; Zhu et al., 2016). However, our focus is on a substantially different task - tracing the origins of generated images. Additionally, previous studies only include the studies on relatively outdated and small-scale GANs (Goodfellow et al., 2014). In contrast, we shift our attention to state-of-the-art large-scale latent generative models such as Stable Diffusion (Rombach et al., 2022), which have more advanced architectures with significantly greater numbers of parameters, different training data distributions, and much larger training data diversity.

3. Preliminary

Inversion-based method is a promising way to achieve alteration-free origin attribution method (Wang et al., 2023d). In this section, we provide more background about it. To facilitate discussion, we first introduce the definition of the belonging of image generation model.

Definition 1. (Belonging of Image Generation Model) Given an image generation model $\mathcal{M} : \mathcal{I} \mapsto \mathcal{X}_{\mathcal{M}}$, where \mathcal{I} denote the input space. $\mathcal{X}_{\mathcal{M}}$ is the output space of the model. A sample x is a **belonging** of model \mathcal{M} if and only if $x \in \mathcal{X}_{\mathcal{M}}$. It is a **non-belonging** if $x \notin \mathcal{X}_{\mathcal{M}}$.

Intuitively, the belongings of a latent generative model are the images generated by this model, and the non-belongings include the images generated by other models and the real images. Then, we have the definition of the input inversion:

Definition 2. (Input Inversion) Given an image generation model $\mathcal{M} : \mathcal{I} \mapsto \mathcal{X}$, where \mathcal{I} and \mathcal{X} are input space and pixel space of the image generation model, respectively, the latent inversion task for an image x is finding the input i^* that makes the generated image $\mathcal{M}(i^*)$ as close as possible to the image x . The reconstruction loss is defined as $\mathcal{L}(\mathcal{M}(i^*), x)$, where \mathcal{L} is a distance measurement.

Based on the definition of the input inversion, we have the definition of the inversion-based origin attribution:

Definition 3. (Inversion Based Origin Attribution) Given an inversion method $\mathcal{R} : \mathcal{X} \mapsto \mathcal{I}$, the inversion-based origin attribution method determines a given sample x is a belonging of a given model \mathcal{M} if the reconstruction loss of the reverse-engineered sample is smaller than a threshold value t , i.e., $\mathcal{L}(\mathcal{M}(i^*), x) < t$.

The inversion-based origin attribution method will be highly effective if the reconstruction losses of the belonging images and that of the non-belonging images are well-separated by the threshold value. The most straightforward way to conduct the inversion is using the gradient-based method (Wang et al., 2023d), which is formally defined as follows:

Definition 4. (Gradient-based Inversion) Given a model \mathcal{M} and an image x , the gradient-based inversion searches the inverted input i' by repeatedly updating the input via the gradient on the reconstruction loss until converge. For each step, the searched input i' is updated via the following equation: $i' = i' - lr \cdot \frac{\partial \mathcal{L}(x, \mathcal{M}(i'))}{\partial i'}$, where lr is the learning rate and \mathcal{L} is the measurement of the distance.

4. Method

In this section, we first provide the formulation of the investigated problem and then introduce our method LATENTTRACER designed for tracing the generated images of the inspected models.

4.1. Problem Formulation

Inspector’s Goal. The goal of the inspector is tracing the belongings of the inspected model \mathcal{M} in an *alteration-free manner* (i.e., without any extra requirement during the development or generation phase of the model). It can be viewed as designing a tracing algorithm \mathcal{B} whose inputs are the

examined image x and the inspected model \mathcal{M} , it then returns a flag representing whether the examined image is the belonging of the inspected model or not. Formally, it can be written as $\mathcal{B} : (\mathcal{M}, x) \mapsto \{0, 1\}$, where 0 denotes image belonging to the model, and 1 denotes non-belonging.

Inspector’s Capability. The inspector has white-box access to the inspected model \mathcal{M} , allowing for the access of intermediate outputs and the computation of the model’s gradients. This assumption is practical, especially in the scenario where the inspector is the owner of the inspected model. However, the inspector cannot necessarily control the development and the generation process of the inspected model. Since our approach exploits the implicit watermarks in autoencoders module of latent generative models, we focus on traceability in latent generative models with unique and distinct auto-encoders. Distinguishing the belonging images of the inspected model and the images generated by other models sharing the same auto-encoder with the inspected model will be our future work.

4.2. Latent Inversion

For the latent generative model \mathcal{M} , the process for synthesizing images can be written as $x = \mathcal{M}(p, n) = \mathcal{D}(\mathcal{C}(p, n))$, where p and n are the conditional input and the unconditional noise for the latent generation process, respectively. \mathcal{D} is the decoder to transform the latent to the pixel space. \mathcal{C} here is the latent generation process in the latent space. x is the image synthesized by the model \mathcal{M} .

Asymmetric challenge of data generation and origin attribution. A straightforward way to conduct the inversion-based origin attribution is by searching the input space of the whole model, including the conditional prompts and the unconditional noise, to identify whether an input would have led to generation of the given space. However, during the data generation phase, the users of the model have full control of selecting the hyper-parameters in the latent generation process (e.g., diffusion process and autoregressive process), such as the selection of different diffusion samplers in latent diffusion models (e.g., DDIM (Song et al., 2020) and DPM-Solver (Lu et al., 2022)). Therefore, inverting the input for the whole model is challenging as the inspector does not know the exact hyper-parameters used in the latent generation phase.

Leveraging deterministic decoders for origin attribution.

In latent generative models, the first step is sampling a new latent vector from a latent generation process (e.g., diffusion process and autoregressive process), which is then upsampled to pixel space using a deterministic decoder. While the stochastic and asymmetric generation process in latent space creates an asymmetric challenge in traceability, as discussed above, we bypass it by solely focusing on traceability using the deterministic decoder. As we discussed

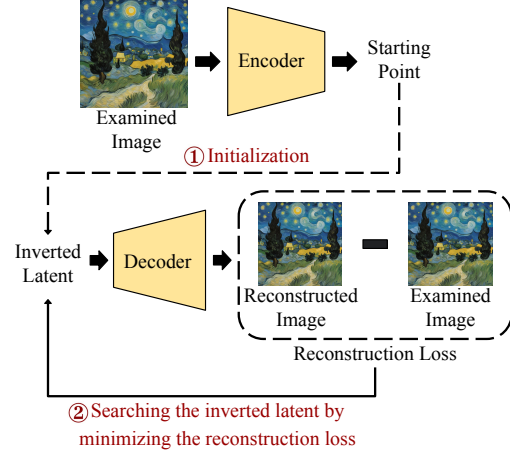
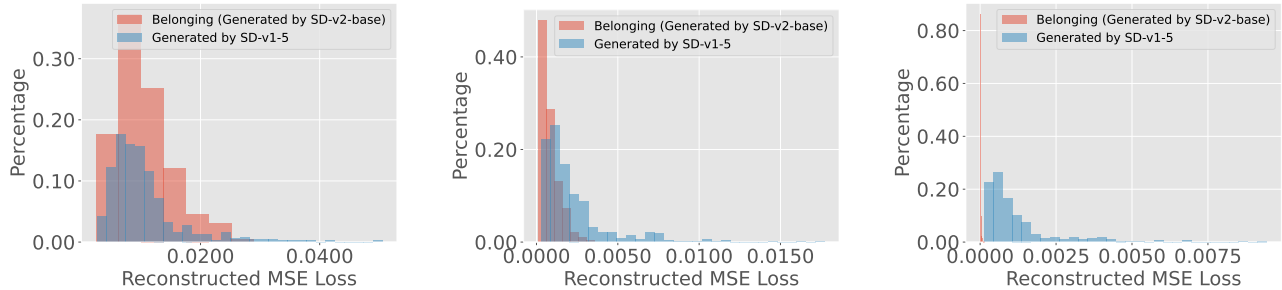


Figure 1. Pipeline of our latent inversion method. First, our method uses the corresponding encoder to get the starting point for the inversion. Then, it uses the gradient-based optimization to search the inverted latent by minimizing the reconstruction loss. The examined image is flagged as a belonging image of the inspected model if the final reconstruction loss is smaller than a threshold.

in § 4.1, we focus on the scenario where the latent generative models have unique autoencoders. In this case, if the examined image x is generated by the decoder \mathcal{D} , then it must be the belonging of the latent generative model \mathcal{M} . To design an origin attribution method that is *orthogonal to the selection of the samplers and hyper-parameters during generation*, we convert the problem of origin attribution into detecting whether the images are generated by the decoder of the inspected model.

Therefore, we focus on the *latent inversion* introduced as follows: Given the decoder $\mathcal{D} : \mathcal{A} \mapsto \mathcal{X}$ in the latent generative model and the examined image x , where \mathcal{A} is the latent space and \mathcal{X} is the pixel space, we aim at finding the latent a^* that makes the generated image $\mathcal{D}(a^*)$ as close as possible to the examined image x . We then perform origin attribution based on the reconstruction loss of the latent inversion. In detail, we consider the examined image x as a belonging if $\mathcal{L}(\mathcal{D}(a^*), x) < t$, where t is a threshold value. The pipeline of our method is demonstrated in Figure 1. Different from the inversion method in Wang et al. (2023d) that uses the random starting point for optimization, our method first uses the corresponding *encoder* to obtain the starting point for the inversion. Following this, it works by checking if the examined images can be well-reconstructed on the inspected model with an inverted latent searched by the gradient optimization. Our method is generalizable to the models equipped with different types of auto-encoders, e.g., Continuous Auto-encoder (Kingma & Welling, 2013) in Stable Diffusion (Rombach et al., 2022), Quantized Auto-encoder (Van Den Oord et al., 2017). Our method also works on Autoregressive Model (Yu et al., 2021a) used in Parti (Yu et al., 2022a) (See § 5.4). The design details of our method can be found in the following sections.



(a) Random Initialization + Gradient-based Inversion (Wang et al., 2023d) (b) Encoder-based Inversion (c) Encoder-based Initialization + Gradient-based Inversion (Ours)

Figure 2. Comparison on the reconstruction loss distributions for different inversion methods. The scenario is distinguishing the 500 images generated by the inspected model (i.e., Stable Diffusion v2-base) and the 500 images generated by other model (i.e., Stable Diffusion v1-5 here). 50 prompts sampled from PromptHero (Inc.) are used to generate these belonging images and non-belonging images (More details about the used prompts can be found in § 6.10). Our method is highly effective since the reconstruction losses for the belongings and that for non-belongings are nearly completely separated in our method.

4.3. Limitation of Canonical Gradient-based Inversion

A straightforward way for conducting latent inversion is directly using the gradient-based method to search the inverted latent (similar to Definition 4). However, we find that directly using the gradient-based inversion has sub-optimal effectiveness and low efficiency on the state-of-the-art latent generative models, especially in the case of distinguishing the belonging images and the images generated by other models having similar architectures. Figure 2a provides the empirical results indicating the sub-optimal effectiveness of the gradient-based latent inversion method since it demonstrates that the reconstruction losses of the belonging images and non-belonging images are not well-separated. The experimental settings can be found in the caption of Figure 2. We also find that it requires large number of optimization steps to convergence.

Upon investigation we find that a key factor for the efficiency and the effectiveness of the gradient-based latent inversion is the selection of the starting point of the optimization (Bertsekas, 2009). More specifically, a starting point closer to the ground-truth latent (i.e., the latent used when generating the belonging image) will lead to better efficiency and effectiveness for the origin attribution. To

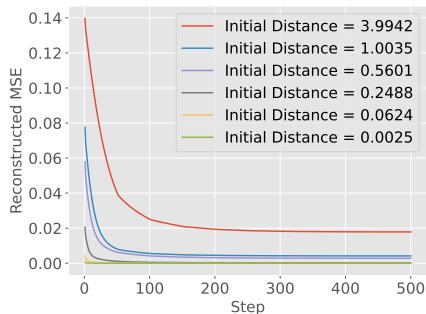


Figure 3. Effects of the initial distance to the ground-truth latent. The starting point closer to the ground-truth latent will lead to better efficiency and effectiveness for the inversion.

confirm this, we create different starting points that have different initial distances to the ground-truth latent and collect their reconstruction losses during the optimization process. Figure 3 demonstrates that a closer starting point will lead to a faster convergence speed and lower reconstruction loss after the convergence, and it confirms our analysis. **Technically, our contribution can be summarized as follows:** We observed the ineffectiveness and inefficiency for the existing inversion based origin attribution methods on the state-of-the-art large latent generative models, and identify the key reason for this phenomenon. We propose a simple yet effective approach to solve this problem, and demonstrate that images produced by the state-of-the-art latent generative models may naturally carry an implicit watermark added by the decoder when decoding the latent samples as they can be detected by our method without any artificial watermarks.

4.4. Our Approach: Exploiting the Invertibility of the Auto-encoder

In this section, we introduce our method LATENTTRACER that combines the gradient-based method and the invertibility of the auto-encoder. For the auto-encoder used in the latent generative models, the main training objective is to make the decoder have the ability to reconstruct the input of the encoder (Kingma & Welling, 2013), i.e., $x \approx \mathcal{D}(\mathcal{E}(x))$, where \mathcal{E} is the encoder and \mathcal{D} is the decoder, x denotes the image input. Existing image editing methods for the latent generative models (Mokady et al., 2023; Parmar et al., 2023) demonstrate that the encoder (i.e., \mathcal{E}) of the auto-encoder used in the latent generative models can also approximate the latent input of the decoder \mathcal{D} , i.e., $a \approx \mathcal{E}(\mathcal{D}(a))$, where a is the input in the latent space. Intuitively, we can directly use the encoder to invert the latent input of the decoder, and we define this inversion method as encoder-based inversion.

Definition 5. (Encoder-based Latent Inversion) Given a decoder \mathcal{D} and its corresponding encoder \mathcal{E} , for an image x ,

Table 1. Comparison for the initial distances from the ground-truth latent to the starting points generated by different initialization methods. The reported number is the mean and standard deviation values of 100 different belonging samples.

Model	Initial Distance	
	Random Initialization	Encoder-based Initialization
Stable Diffusion v1-5	1.930±0.021	0.019±0.007
Stable Diffusion v2-base	1.844±0.017	0.014±0.004

the encoder-based latent inversion finds the inverted latent \mathbf{a}' by directly exploiting the encoder, i.e., $\mathbf{a}' = \mathcal{E}(\mathbf{x})$.

The encoder-based latent inversion is highly efficient since it only requires one forward process of the encoder and the runtime for it is nearly negligible. The research question that we want to explore is: *Is the encoder-based inversion effective for the origin attribution problem of the latent generative models?* We show results of a preliminary evaluation in Figure 2b, where the reconstruction loss of the belonging images (i.e., the images generated by the inspected model Stable Diffusion v2-base) is represented by red color, and the reconstruction losses of the images generated by other models are shown in the blue color. As can be seen, the reconstruction losses of the belonging samples and those of the non-belonging samples are not well-separated. Thus, we have the following observation:

Observation 1. Encoder-based latent inversion method has low effectiveness for the origin attribution problem.

This observation is expected since the auto-encoder only provides an approximation of the inversion. Although directly using the encoder to invert the input in the latent space leads to low effectiveness, we still can exploit the invertibility of the auto-encoder. Note that in § 4.3, we discussed that the effectiveness and the efficiency of the gradient-based inversion are highly sensitive to the starting point before the optimization. Since the auto-encoder used in the latent generative models has potential invertibility, another question we want to explore is: *Before the optimization process of the gradient-based latent inversion, is it possible to use the encoder to get the starting point that is better than the random starting point?* To investigate this question, we compare the distance from the randomly generated starting point to the ground-truth latent and that from the encoder-generated starting point. The results are shown in Table 1. The models used here are the Stable Diffusion v1-5 and the Stable Diffusion v2-base. The results indicate that the encoder-generated starting point’s distance to the ground-truth latent is much smaller than that of the random-generated starting point. To further explore the effectiveness and the efficiency of the encoder-based latent initialization with the gradient-based inversion method, we record the reconstruction loss curve for both the random latent initialization with the gradient-

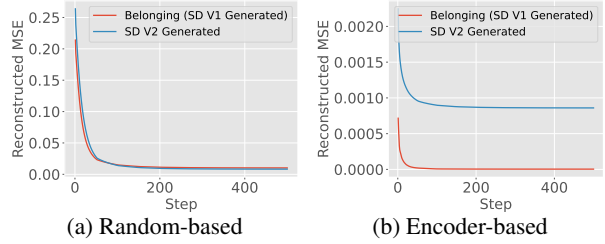


Figure 4. Comparison on the reconstruction loss curve for the random starting point and the encoder-generated starting point. The encoder-based starting point lead to a faster convergence speed and a better separation of the reconstruction losses.

based inversion and the encoder-based latent initialization with the gradient-based inversion. The inspected model here is the Stable Diffusion v1-5. The belonging image and the non-belonging image here are generated by using the same prompt. As shown in Figure 4, the convergence speed of using the encoder-generated starting point is much faster than that of using the random starting point. At the same time, the encoder-based initialization leads to an obvious separation of the reconstructed losses on belonging samples and the non-belonging samples, while the random initialization does not have such an effect. The results for the reconstruction losses of 500 belonging images and 500 non-belonging images in Figure 2a and Figure 2c also show that the gradient-based inversion with encoder-based initialization is much more effective than the inversion with random initialization since the former leads to a much higher separability for the reconstruction losses of the belonging samples and the non-belonging samples. Based on these results, we have the following observation:

Observation 2. For gradient-based latent inversion and origin attribution, the encoder-based initialization leads to better effectiveness and efficiency compared to the random initialization.

Algorithm. We design our algorithm based on our analysis and observations. Our algorithm, detailed in Algorithm 1, takes as input the image under examination, denoted as \mathbf{x} , alongside the inspected model \mathcal{M} . Its output manifests as the inference outcomes, determining whether the examined image either is the belonging of the inspected model or not. At the first step of the process, the algorithm determines the threshold for the detection. In detail, it utilizes the model \mathcal{M} to generate N (defaulting to 100 in this paper) images from randomly selected prompts. It then proceeds to calculate the mean value (μ) and standard deviation (σ) of the reconstruction loss on the belonging samples of the model. Following Wang et al. (2023d), we use Grubbs’ Hypothesis Testing (Grubbs, 1950) to determine the threshold:

$$t = \frac{(N-1)\sigma}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/N, N-2})^2}{N-2 + (t_{\alpha/N, N-2})^2}} + \mu \quad (1)$$

Algorithm 1 Origin Attribution by Gradient-based Inversion with Encoder-based Initialization

Input: Model: \mathcal{M} , Examined Data: \mathbf{x}
Output: Inference Results: Belonging or Non-belonging

```

1: function INFERENCE( $\mathcal{M}, \mathbf{x}$ )
2:    $\mathcal{E} = \mathcal{M}.$ Encoder
3:    $\mathcal{D} = \mathcal{M}.$ Decoder
4:   ▷ Obtaining Threshold (Offline)
5:    $t \leftarrow$  Calculating Threshold [Equation 1]
6:   ▷ Reverse-engineering
7:    $\mathbf{a} = \mathcal{E}(\mathbf{x})$ 
8:   for  $e \leq \text{max\_epoch}$  do
9:      $\text{cost} = \mathcal{L}(\mathcal{D}(\mathbf{a}), \mathbf{x})$ 
10:     $\Delta_{\mathbf{a}} = \frac{\partial \text{cost}}{\partial \mathbf{a}}$ 
11:     $\mathbf{a} = \mathbf{a} - lr \cdot \Delta_{\mathbf{a}}$ 
12:   ▷ Determining Belonging
13:   if  $\text{cost} \leq t$  then
14:     return Belonging
15:   else
16:     return NonBelonging

```

Here $t_{\alpha/N, N-2}$ is the critical value of the t distribution with $N - 2$ degrees of freedom and a significance level of α/N , where α is the significance level of the hypothesis testing (i.e., 0.05 by default in this paper). More details of the critical value can be found in § 6.2. This calculation, being an offline process, necessitates execution only once per model. In Line 7, we calculate the starting point of the optimization by exploiting the forward process of the encoder. Moving on to Line 8-11, the reconstruction loss is calculated and the inverted latent is optimized by gradient descent optimizer. Note that the starting point is obtained by using the encoder \mathcal{E} to encode the inspected image \mathbf{x} . Lastly, Line 13-16 involves determining the affiliation of the examined data \mathbf{x} with the model \mathcal{M} .

5. Experiments and Results

We first introduce the experiment setup (§ 5.1). We then evaluate the effectiveness (§ 5.2) and the efficiency (§ 5.3) of our method LATENTTRACER. We also study the results on different types of auto-encoders in § 5.4 and compare our method to existing methods requiring extra operations during training phase or generation phase in § 5.5.

5.1. Experiment Setup

Our method is implemented with Python 3.10 and PyTorch 2.0. We conducted all experiments on a Ubuntu 20.04 server equipped with 8 A100 GPUs (one experiment/GPU).

Models. There are six state-of-the-art latent generative models involved in the experiments: Stable Diffusion v1-4 (SD

v1-4), Stable Diffusion v1-5 (SD v1-5), Stable Diffusion v2-base (SD v2-base), Stable Diffusion v2-1 (SD v2-1), Stable Diffusion XL-1.0-base (SD XL-1.0-base) and Kandinsky 2.1. Details of the models are in § 6.1.

Evaluation Metrics. The effectiveness of the origin attribution methods is measured by calculating the accuracy of detection (Acc). For an inspected model, when dealing with a mix of images that are either belonging or non-belonging of it, Acc represents the proportion of accurately identified images to the total number of images. Additionally, we present a comprehensive count of True Positives (TP, or correctly identified belonging images), False Positives (FP, or non-belongings identified as belongings), False Negatives (FN, or belongings identified as non-belongings), and True Negatives (TN, or non-belongings correctly identified).

5.2. Effectiveness

In this section, we study the effectiveness of our method on the origin attribution task. We first investigate the effectiveness on distinguishing belonging images and images generated by other models, following with the study for the effectiveness on distinguishing belongings and real images.

Distinguishing Belonging Images and Images Generated by Other Models. In this section, we study our method’s effectiveness on distinguishing between belonging images of a particular model and the images generated by other models. To measure the effectiveness of our method, we first randomly collect 50 prompts designed for the state-of-the-art text-to-image generative models on the PromptHero (Inc.) (a main-stream website for prompt engineering of the generative models). For each prompt, we generate 10 samples on each model by using different random seeds. Thus, we have 500 generated samples for each model. For different inspected models, we use our method to distinguish the belonging images and the images generated by other models. We compare our method to the existing reverse-engineering based origin attribution method RONAN (Wang et al., 2023d). While most of the existing methods (such as image watermarking (Wen et al., 2023a) and model fingerprinting (Yu et al., 2022b)) require extra operations during the training or generation phase, RONAN is the only method that can achieve alteration-free origin attribution and *has the same threat model with our method*. The results are shown in Table 2, where Model \mathcal{M}_1 denotes the inspected model, and Model \mathcal{M}_2 represents the other model. For our method, the average detection accuracy (Acc) of our method is 93.4%, confirming its good performance for distinguishing between the belongings of the inspected model and the images generated by other models. As can be seen, the detection accuracy of RONAN is only around 50% in our setting. The results indicate our method outperforms the existing alteration-free origin attribution method by a

Table 2. Results for distinguishing belonging images and images generated by other models. Here, Model \mathcal{M}_1 is the inspected model, Model \mathcal{M}_2 is the other model. As we discussed in § 4.1, we focus on the traceability in latent generative models with unique and distinct auto-encoders. Thus, we do not include the setting where \mathcal{M}_1 and \mathcal{M}_2 share the same autoencoder, e.g., \mathcal{M}_1 =SD v1-4 and \mathcal{M}_2 =SD v1-5. To our understanding, RONAN (Wang et al., 2023d) is the only method that shares the same problem formulation as our method.

Model \mathcal{M}_1	Model \mathcal{M}_2	RONAN					LATENTTRACER (Ours)				
		TP	FP	FN	TN	Acc	TP	FP	FN	TN	Acc
SD v1-4	SD v2-base	477	485	23	15	49.2%	480	53	20	447	92.7%
	SD v2-1	480	485	20	15	49.5%	482	55	18	445	92.7%
	SD XL-1.0-base	475	470	25	30	50.5%	476	81	24	419	89.5%
	Kandinsky	484	470	16	30	51.4%	475	65	25	435	91.0%
SD v1-5	SD v2-base	477	484	23	16	49.3%	481	55	19	445	92.6%
	SD v2-1	478	482	22	18	49.6%	479	54	21	444	92.5%
	SD XL-1.0-base	477	472	23	28	50.5%	477	85	23	415	89.2%
	Kandinsky	482	469	18	31	51.3%	477	66	23	434	91.1%
SD v2-base	SD v1-4	481	425	19	75	55.6%	482	0	18	500	98.2%
	SD v1-5	480	422	20	78	55.8%	480	0	20	500	98.0%
	SD XL-1.0-base	476	480	24	20	49.6%	480	53	20	447	92.7%
	Kandinsky	473	473	27	27	50.0%	478	1	22	499	97.7%
Kandinsky	SD v1-4	481	479	19	21	50.2%	474	12	26	488	96.2%
	SD v1-5	483	480	17	20	50.3%	476	11	24	489	97.5%
	SD v2-base	482	478	18	22	50.4%	480	46	20	454	93.4%
	SD v2-1	483	475	17	25	50.8%	479	46	21	454	93.3%
	SD XL-1.0-base	478	490	22	10	48.8%	480	82	20	418	89.8%

Table 3. Results for distinguishing belongings and real images. The real images used are randomly sampled from LAION (Schuhmann et al., 2022).

Inspected Model	TP	FP	FN	TN	Acc
SD v1-4	484	10	16	490	97.4%
SD v1-5	487	9	13	491	97.8%
SD v2-base	487	6	13	494	98.1%
Kandinsky	483	0	17	500	98.3%

large margin on the origin attribution for the state-of-the-art latent generative models. In § 5.5, we also compare our method to existing state-of-the-art methods that require extra operations during the training or generation phase.

Distinguishing Belonging Images and Real Images. In this section, we investigate our approach’s effectiveness in distinguishing between belonging images of a particular model and real images. The investigated models include Stable Diffusion v1-4, Stable Diffusion v1-5, Stable Diffusion v2-base, and Kandinsky. We first randomly sample 500 images and their corresponding text captions in the LAION dataset (Schuhmann et al., 2022), which is the training dataset for many state-of-the-art text-to-image latent generative models such as Stable Diffusion (Rombach et al., 2022). For each model, we use the randomly sampled text captions as the prompts to generate 500 images as the belonging images. The randomly sampled images from the LAION dataset are used as the non-belonging samples here. The results are demonstrated in Table 3. On average, the detection accuracy (Acc) is 97.9%. The results show that our method achieves good performance in distinguishing belonging images of a particular model and real images.

Table 4. Average runtime on different models.

Inspected Model	Runtime	
	with Random Initialization	with Encoder-based Initialization (Ours)
SD v1-4	87.6s	21.5s
SD v1-5	88.1s	21.8s
SD v2-base	93.5s	23.2s
Kandinsky	120.7s	30.3s

5.3. Efficiency

In this section, we study the efficiency of our method. We collect the runtime for inferring if an examined image is the belonging of the inspected model. Four models (i.e., Stable Diffusion v1-4, Stable Diffusion v1-5, Stable Diffusion v2-base, Kandinsky) are included in this section as the inspected models. For each model, we run 5 times and the average runtime is reported in Table 4. Besides the runtime for our method (gradient-based inversion with encoder-based initialization), we also show the average runtime for the gradient-based inversion with random initialization. As can be observed, the speed of our method is much faster than the method with random initialization. This is because the starting point obtained by the encoder-based initialization is much closer to the ground-truth, leading to a faster convergence speed. Based on our experiments, our method only requires 100 optimization steps to converge, while the method with random initialization needs 400 steps (also see Figure 4). In conclusion, our method is much more efficient than Wang et al. (2023d).

5.4. Results on Different Types of Models

It is possible that the latent generative models may use different types of auto-encoders or different latent generation

Table 5. Generalization to different types of models.

Type	Model	Acc
Continuous Auto-encoder	VAE (Kingma & Welling, 2013)	98.2%
Quantized Auto-encoder	VQ-VAE (Van Den Oord et al., 2017)	97.5%
Autoregressive Model	ViT-VQGAN (Yu et al., 2021a)	98.4%

processes (e.g., diffusion and autoregressive process). In this section, we discuss our method’s generalization to different types of models. To study this, we conduct experiments on two types of autoencoders (i.e., Continuous Auto-encoder and Quantized Auto-encoder). Besides the experiments on the models with diffusion-based latent generation process (§ 5.2), we also evaluate our method on Autoregressive Model in this section. For Continuous Auto-encoder, we use the VAE (Kingma & Welling, 2013) model used in the Stable Diffusion v1-5. For Quantized Auto-encoder, we use a VQ-VAE (Van Den Oord et al., 2017) model trained on the CIFAR-10 dataset (Krizhevsky et al., 2009). For Autoregressive Model, we use a ViT-VQGAN (Yu et al., 2021a) model trained on the ImageNet dataset (Russakovsky et al., 2015). We randomly sample 1000 images from the LAION (Schuhmann et al., 2022), CIFAR-10 (Krizhevsky et al., 2009), and ImageNet (Russakovsky et al., 2015) dataset as the non-belonging images for the VAE, VQ-VAE, and ViT-VQGAN, respectively. To obtain the belonging images that share similar complexity to these real images, we then use these models to conduct reconstruction on these real images and obtain the belonging images of the decoders that are similar to these real images, i.e., $x' = \mathcal{D}(\mathcal{E}(x))$ where x is the real images and x' is the obtained belonging images similar to the corresponding real images. \mathcal{E} and \mathcal{D} are the encoder and the decoder, respectively. After we obtain all the belonging images and non-belonging images for each model, we use our method to distinguish them (Table 5) and find that our method has above 97% detection accuracy on all different types of auto-encoders or latent generation processes, showing our method’s generalization ability.

5.5. Comparison to Methods Requiring Extra Steps

In this section, we provide the comparison to more state-of-the-art methods that require extra operations during the training or generation phase to illustrate a broader spectrum of comparison. The model used here is Latent Diffusion Model (Rombach et al., 2022), and the dataset here is MS-COCO. The results are shown in Table 6. We report the SSIM value and the FID value between the original samples and the watermarked/fingerprinted generated samples. For SSIM (Structural Similarity Index Measure), a value of 1 indicates perfect similarity between two compared images, meaning they are exactly the same. For FID (Fréchet Inception Distance), a value of 0 indicates that the two distributions are identical. Since our method LATENTTRACER does not have any watermarking/fingerprinting process, we can

Table 6. Comparison to methods requiring extra steps.

Method	SSIM	FID
Dct-Dwt (Cox et al., 2007)	0.97	19.5
SSL Watermark (Fernandez et al., 2022)	0.86	20.6
FNNS (Kishore et al., 2021)	0.90	19.0
HiDDeN (Zhu et al., 2018)	0.88	19.7
Stable Signature (Fernandez et al., 2023)	0.89	19.6
LATENTTRACER (Ours)	1.00	0

view the watermarked/fingerprinted samples of our method are identical to the original generated samples. Consequently, our method achieves a perfect SSIM score of 1 and an FID score of 0. As can be observed, the methods necessitating extra steps have non-negligible negative influences on the generation quality, while our LATENTTRACER guarantees no quality degradation.

6. Conclusion

We propose a latent inversion based method to detect the generated images of the inspected model by checking if the examined images can be well-reconstructed with an inverted latent input. Our experiments on different latent generative models demonstrate that our approach is highly accurate in differentiating between images produced by the inspected model and other images. Our results also imply an interesting direction that images created by today’s latent generative models may inherently carry an implicit watermark added by the decoder when decoding the latent samples.

Impact Statement

Studying the security and privacy aspects of the machine learning models potentially has ethical concerns (Carlini et al., 2023a;b; Zou et al., 2023; Wang et al., 2022b; Tao et al., 2022). In this paper, we design a method for tracing the generated images of a specific inspected latent generative model. We are confident that our method will protect the intellectual property and improve the security of latent generative models, and will be advantageous for the development of responsible AI (Gu, 2024; Li et al., 2023a; Guo et al., 2024; Huang et al., 2023; Wang et al., 2023a; Shao et al., 2024; Li et al., 2023b; Wang et al., 2022a).

Acknowledgement

We thank the anonymous reviewers for their valuable comments. This research is supported by Sony AI, IARPA TrojAI W911NF-19-S-0012, NSF 2342250 and 2319944. It is also partially funded by research grants to D. Metaxas through NSF 2310966, 2235405, 2212301, 2003874, and AFOSR-835531. Any opinions, findings, and conclusions expressed in this paper are those of the authors only and do not necessarily reflect the views of any funding agencies.

References

- Albright, M. and McCloskey, S. Source generator attribution via inversion. In *CVPR Workshops*, volume 8, pp. 3, 2019.
- Bertsekas, D. *Convex optimization theory*, volume 1. Athena Scientific, 2009.
- Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2:3, 2023.
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023a.
- Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023b.
- Chen, C., Fu, J., and Lyu, L. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325*, 2023.
- Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.
- Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T. *Digital watermarking and steganography*. Morgan kaufmann, 2007.
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.
- Durall, R., Keuper, M., Pfrendt, F.-J., and Keuper, J. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.
- Durall, R., Keuper, M., and Keuper, J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7890–7899, 2020.
- Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., and Douze, M. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3054–3058. IEEE, 2022.
- Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.
- Flynn, A., Clough, J., and Cooke, T. Disrupting and preventing deepfake abuse: Exploring criminal law responses to ai-facilitated abuse. *The palgrave handbook of gendered violence and technology*, pp. 583–603, 2021.
- Flynn, A., Powell, A., Scott, A. J., and Cama, E. Deepfakes and digitally altered imagery abuse: A cross-country exploration of an emerging form of image-based sexual abuse. *The British Journal of Criminology*, 62(6):1341–1358, 2022.
- Francke, E. and Bennett, A. The potential influence of artificial intelligence on plagiarism: A higher education perspective. In *European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2019)*, pp. 131–140, 2019.
- Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.
- Grubbs, F. E. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, pp. 27–58, 1950.
- Gu, J. Responsible generative ai: What to generate and what not. *arXiv preprint arXiv:2404.05783*, 2024.
- Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.
- Guo, J., Li, Y., Wang, L., Xia, S.-T., Huang, H., Liu, C., and Li, B. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *Advances in Neural Information Processing Systems*, 36, 2024.
- Huang, Z., Li, B., Cai, Y., Wang, R., Guo, S., Fang, L., Chen, J., and Wang, L. What can discriminator do? towards box-free ownership verification of generative adversarial networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5009–5019, 2023.
- Inc., P. PromptHero. <https://prompthero.com/>.

- Jahanian, A., Chai, L., and Isola, P. On the "steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.
- Jeong, Y., Kim, D., Ro, Y., and Choi, J. FrepGAN: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1060–1068, 2022.
- Jovanović, N., Staab, R., and Vechev, M. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*, 2024.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of styleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.
- Kietzmann, J., Lee, L. W., McCarthy, I. P., and Kietzmann, T. C. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, 2020.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Kishore, V., Chen, X., Wang, Y., Li, B., and Weinberger, K. Q. Fixed neural network steganography: Train the images, not the network. In *International Conference on Learning Representations*, 2021.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- Laszkiewicz, M., Ricker, J., Lederer, J., and Fischer, A. Single-model attribution via final-layer inversion. *arXiv preprint arXiv:2306.06210*, 2023.
- Li, H., Shen, C., Torr, P., Tresp, V., and Gu, J. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. *arXiv preprint arXiv:2311.17216*, 2023a.
- Li, Y., Zhu, M., Yang, X., Jiang, Y., Wei, T., and Xia, S.-T. Black-box dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 2023b.
- Liu, R., Khakzar, A., Gu, J., Chen, Q., Torr, P., and Pizzati, F. Latent guard: a safety framework for text-to-image generation. *arXiv preprint arXiv:2404.08031*, 2024.
- Liu, Z., Qi, X., and Torr, P. H. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020.
- Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.
- Luo, L., Chen, Z., Chen, M., Zeng, X., and Xiong, Z. Reversible image watermarking using interpolation technique. *IEEE Transactions on information forensics and security*, 5(1):187–193, 2009.
- Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
- Ma, S., Zhang, X., and Xu, D. Protracer: Towards practical provenance tracing by alternating between logging and tainting. In *23rd Annual Network And Distributed System Security Symposium (NDSS 2016)*. Internet Soc, 2016.
- Mirsky, Y. and Lee, W. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.
- Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.
- Nie, G., Kim, C., Yang, Y., and Ren, Y. Attributing image generative models using latent fingerprints. In *International Conference on Machine Learning*, pp. 26150–26165. PMLR, 2023.
- Pan, M., Wang, Z., Dong, X., Sehwal, V., Lyu, L., and Lin, X. Finding needles in a haystack: A black-box approach to invisible watermark detection. *arXiv preprint arXiv:2403.15955*, 2024.
- Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., and Zhu, J.-Y. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.
- Partadiredja, R. A., Serrano, C. E., and Ljubenkov, D. Ai or human: the socio-ethical implications of ai-generated media content. In *2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275)*, pp. 1–6. IEEE, 2020.
- Pereira, S. and Pun, T. Robust template matching for affine resistant image watermarks. *IEEE transactions on image processing*, 9(6):1123–1129, 2000.
- Razhigayev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., and Dimitrov, D. Kandinsky: an improved

- text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.
- Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.
- Sha, Z., Li, Z., Yu, N., and Zhang, Y. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint arXiv:2210.06998*, 2022.
- Shao, S., Li, Y., Yao, H., He, Y., Qin, Z., and Ren, K. Explanation as a watermark: Towards harmless and multi-bit model ownership verification via watermarking feature attribution. *arXiv preprint arXiv:2405.04825*, 2024.
- Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- Swanson, M. D., Zhu, B., and Tewfik, A. H. Transparent robust image watermarking. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pp. 211–214. IEEE, 1996.
- Tancik, M., Mildenhall, B., and Ng, R. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.
- Tao, G., Wang, Z., Cheng, S., Ma, S., An, S., Liu, Y., Shen, G., Zhang, Z., Mao, Y., and Zhang, X. Backdoor vulnerabilities in normally trained deep learning models. *arXiv preprint arXiv:2211.15929*, 2022.
- Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Wang, R., Lin, C., Zhao, Q., and Zhu, F. Watermark faker: towards forgery of digital image watermarking. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.
- Wang, R., Ren, J., Li, B., She, T., Zhang, W., Fang, L., Chen, J., and Wang, L. Free fine-tuning: A plug-and-play watermarking scheme for deep neural networks. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8463–8474, 2023a.
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Wang, Z., Mei, K., Zhai, J., and Ma, S. Unicorn: A unified backdoor trigger inversion framework. In *The Eleventh International Conference on Learning Representations*, 2022a.
- Wang, Z., Zhai, J., and Ma, S. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15074–15084, 2022b.
- Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023b.
- Wang, Z., Chen, C., Lyu, L., Metaxas, D. N., and Ma, S. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023c.
- Wang, Z., Chen, C., Zeng, Y., Lyu, L., and Ma, S. Where did i come from? origin attribution of ai-generated images. *Advances in neural information processing systems*, 2023d.
- Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023a.
- Wen, Y., Liu, Y., Chen, C., and Lyu, L. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023b.

- Whittaker, L., Kietzmann, T. C., Kietzmann, J., and Dabirian, A. “all around me are synthetic faces”: the mad world of ai-generated media. *IT Professional*, 22(5): 90–99, 2020.
- Wu, M., Ma, J., Wang, R., Zhang, S., Liang, Z., Li, B., Lin, C., Fang, L., and Wang, L. Traceevader: Making deep-fakes more untraceable via evading the forgery model attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19965–19973, 2024.
- Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldrige, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021a.
- Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022a.
- Yu, N., Davis, L. S., and Fritz, M. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7556–7566, 2019.
- Yu, N., Skripniuk, V., Abdelnabi, S., and Fritz, M. Artificial fingerprinting for generative models: Rooting deep-fake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 14448–14457, 2021b.
- Yu, N., Skripniuk, V., Chen, D., Davis, L. S., and Fritz, M. Responsible disclosure of generative models using scalable fingerprinting. In *International Conference on Learning Representations*, 2022b.
- Zhang, B., Zhou, J. P., Shumailov, I., and Papernot, N. On attribution of deepfakes. *arXiv preprint arXiv:2008.09194*, 2020.
- Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.
- Zheng, C., Vuong, T.-L., Cai, J., and Phung, D. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.
- Zhu, J., Kaplan, R., Johnson, J., and Fei-Fei, L. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 657–672, 2018.
- Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative visual manipulation on the natural image manifold. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pp. 597–613. Springer, 2016.
- Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Appendix

6.1. More Details of the Used Models

In this section, we provide more details about the model used in the experiments.

Stable Diffusion v1-4². This model is initialized with the weights of the Stable-Diffusion-v1-2 model³ and then fine-tuned on 225k steps at resolution 512x512 on "laion-aesthetics v2 5+". The architecture of the auto-encoder used in this model is the VAE. This model is with creativeml-openrail-m License.

Stable Diffusion v1-5⁴. This model is initialized with the weights of the Stable-Diffusion-v1-2 model and subsequently fine-tuned on 595k steps on "laion-aesthetics v2 5+" with 512x512 resolution. The architecture of the auto-encoder used in this model is the VAE. This model is with creativeml-openrail-m License.

Stable Diffusion v2-base⁵. This model model is trained from scratch 550k steps at resolution 256x256 on a subset of LAION-5B filtered for explicit pornographic material, using the LAION-NSFW classifier with punsafe=0.1 and an aesthetic score ≥ 4.5 . It is further trained for 850k steps at resolution 512x512 on the same dataset on images with resolution $\geq 512x512$. The architecture of the auto-encoder used in this model is the VAE. This model is with openrail++ License.

Stable Diffusion v2-1⁶. This model is fine-tuned from Stable Diffusion 2 model⁷ with an additional 55k steps on a subset of LAION-5B filtered for explicit pornographic material (with punsafe=0.1), and then fine-tuned for another 155k extra steps with punsafe=0.98. The architecture of the auto-encoder used in this model is the VAE. This model is with openrail++ License.

Stable Diffusion XL-1.0-base⁸. This model is first trained from scratch on an internal dataset constructed by Stability-AI for 600 000 optimization steps at a resolution of 256×256 pixels and a batch-size of 2048. Then, it is trained on 512×512 pixel images for another 200 000 optimization steps. Finally, the trainers utilize multi-aspect training in combination with an offset-noise level of 0.05 to train the model on different aspect ratios of around 1024×1024 pixel area. The developers train the same auto-encoder architecture used for the original Stable Diffusion at a larger

batch-size (256 vs 9) and additionally track the weights with an exponential moving average. This model is with openrail++ License.

Kandinsky 2.1⁹. This model utilizes CLIP and diffusion image prior (mapping) between latent spaces of CLIP modalities to increase the generation performance. For diffusion mapping of latent spaces, it uses the transformer architecture with num_layers=20, num_heads=32 and hidden_size=2048. It also uses the custom implementation of MoVQGAN (Zheng et al., 2022) with minor modifications as the autoencoder (Razzhigaev et al., 2023). The autoencoder is trained on the LAION HighRes dataset (Schuhmann et al., 2022). This model is with apache-2.0 License.

6.2. Critical Value of the t-Distribution

In this section, we introduce the detailed process for calculating the critical value of the t-distribution, which is used in Equation 1 for obtaining the detection threshold. The probability density function for the t-distribution is written in Equation 2, where ν is the number of degrees of freedom and Γ is the gamma function.

$$f(a) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{a^2}{\nu}\right)^{-(\nu+1)/2} \quad (2)$$

Then, the corresponding cumulative distribution function is written in Equation 3, where β denotes the incomplete beta function.

$$\mathbb{P}(a < t') = \int_{-\infty}^{t'} f(u)du = 1 - \frac{1}{2}\beta\left(\frac{\nu}{t'^2 + \nu}, \frac{\nu}{2}, \frac{1}{2}\right) \quad (3)$$

Thus, given a confidence level α and the number of degrees of freedom ν , we have can use Equation 4 to calculate the value of the critical value $t_{\alpha,\nu}$.

$$\mathbb{P}(a < t_{\alpha,\nu}) = 1 - \alpha \quad (4)$$

6.3. Discussion on Mitigating the Malicious Usages of Latent Generative Models

Similar to identifying the origin of cyber attacks in computer systems (Ma et al., 2016), tracing where maliciously generated images and non-deliberately generated harmful/unsafe images come from can reveal more about the information and characteristics of their generation process, which is crucial for understanding the details of the malicious generation process and developing strategies to defend them. For instance, Identifying the origin of the maliciously gen-

²<https://huggingface.co/CompVis/stable-diffusion-v1-4>

³<https://huggingface.co/CompVis/stable-diffusion-v1-2>

⁴<https://huggingface.co/runwayml/stable-diffusion-v1-5>

⁵<https://huggingface.co/stabilityai/stable-diffusion-2-base>

⁶<https://huggingface.co/stabilityai/stable-diffusion-2-1>

⁷<https://huggingface.co/stabilityai/stable-diffusion-2>

⁸<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>

⁹https://huggingface.co/ai-forever/Kandinsky_2.1

How to Trace Latent Generative Model Generated Images without Artificial Watermark?






Image source →	Real image	Stable Diffusion v1-5	Stable Diffusion v2-base	Kandinsky 2.1	Stable Diffusion XL-base-1.0
					
Target model	Reconstruction loss				
Stable Diffusion v1-5	0.0042	0.00003	0.0014	0.0030	0.0015
Stable Diffusion v2-base	0.0044	0.0030	0.00001	0.0032	0.0016
Kandinsky 2.1	0.0039	0.0034	0.0017	0.000002	0.0014
Stable Diffusion XL-base-1.0	0.0007	0.0008	0.0003	0.0005	0.00005

Figure 5. Given a specific latent generative model, can we trace the images generated by this model without extra artificial watermarks? We show that the images generated by the inspected model can be traced *without any additional requirements on the model’s training and generation phase* (such as adding a watermark after generation (Tancik et al., 2020; Wen et al., 2023a) or injecting fingerprinting during training (Yu et al., 2021b; 2022b; Fernandez et al., 2023)) by using the reconstruction loss computed by our method. For example, here we show some images generated by different models and their reconstruction losses on different models. Even though the generated images from different models can look seemingly identical, their reconstruction loss can differ by an order of magnitude. The reconstruction loss is extremely low if the examined image is generated by the inspected model.

erated images or non-deliberately generated harmful/unsafe images provides a reference for regulators to assign responsibility. Furthermore, this knowledge allows the company to refine their model, aiming to prevent the future production of similar malicious or harmful images. Determining the responsibility of the malicious generated images remains an unresolved challenge in the field of law. This complexity arises due to the involvement of multiple entities (such as contributors of training data, model trainers, input/prompt providers, and the models themselves) throughout the image generation process. The tracing results of our origin attribution method can serve as a valuable reference for addressing malicious usages concerns, instead of a definitive responsibility conclusion.

6.4. More Results about the Effectiveness

In this section, we report the AUROC of our method to fully understand its performance. The results are shown in Table 7. The results demonstrate that our method achieves high AUROC in all settings, indicating its high performance and non-sensitivity to different threshold selection methods.

6.5. More Discussion about Wang et al. (2023d)

The baseline RONAN method (Wang et al., 2023d) performs well at differentiating between samples generated by models with significantly different architectures or training datasets. For instance, RONAN achieves 94.5% accuracy when distinguishing images from Stable Diffusion v2 versus StyleGAN2-ADA. In the experimental setup used by Wang et al. (2023d), the architectural differences and variations in training data were substantial (e.g., Stable Diffusion v.s.

Table 7. AUROC for distinguishing belonging images and images generated by other models. Here, Model \mathcal{M}_1 is the inspected model, Model \mathcal{M}_2 is the other model.

Model \mathcal{M}_1	Model \mathcal{M}_2	AUROC
SD v1-4	SD v2-base	0.98
SD v1-4	SD v2-1	0.99
SD v1-4	SD XL-1.0-base	0.96
SD v1-4	Kandinsky	0.98
SD v1-5	SD v2-base	0.98
SD v1-5	SD v2-1	0.99
SD v1-5	SD XL-1.0-base	0.97
SD v1-5	Kandinsky	0.98
SD v2-base	SD v1-4	0.99
SD v2-base	SD v1-5	0.99
SD v2-base	SD XL-1.0-base	0.98
SD v2-base	Kandinsky	0.99
Kandinsky	SD v1-4	0.99
Kandinsky	SD v1-5	0.99
Kandinsky	SD v2-base	0.98
Kandinsky	SD v2-1	0.98
Kandinsky	SD XL-1.0-base	0.97

StyleGAN and ImageNet v.s. LSUN). However, RONAN has difficulty achieving satisfactory performance in the more challenging task of distinguishing between images generated by the large latent generative model being inspected and those from other models that have similar architectures and training datasets, which is the situation presented in our experiments. For example, different versions of the Stable Diffusion model share high similar architectures and training datasets.

6.6. Stopping Strategy

In our implementation, we set 100 as the maximum step. As shown in Figure 4b, the reconstruction loss is converged at step 100. We also evaluated the detection accuracy under

Table 8. Results on different stopping strategies.

Strategy	Acc
Fixed Maximum Step	98.0%
Adaptive Stop	97.7%

an adaptive stop strategy. In detail, the reverse-engineering process stops if the reconstruction loss does not decrease in five successive steps. The results for the fixed maximum step and the adaptive stop strategy are shown in Table 8. The setting here is distinguishing the images generated by the inspected model SD v2-base and another model SD v1-5.

6.7. Robustness

We conducted the experiments for evaluating the robustness of our proposed method against a wide range of post-processing techniques employed in the existing works (Yu et al., 2019; Fernandez et al., 2023). These techniques include adjusting saturation, adjusting contrast, adding Gaussian noise, applying JPEG compression, adjusting brightness, applying Gaussian blur, and Cropping. The inspected model used here is SD v2-base. The setting here is distinguishing the images generated by the inspected model and SD v1-5 model. The other experimental settings are identical to those used in Table 2. For each experiment, we report the detection accuracy and the average values of the Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), L1 distance, and L2 distance between the original samples and the post-processed samples. The L1 and L2 distances are calculated using pixel values ranging from 0 to 1. The results are shown in Table 9, demonstrating that our proposed method remains effective when the quality of the post-processed images is satisfactory. For instance, the detection accuracy of our method consistently exceeds 90% when the Structural Similarity Index (SSIM) value between the original and post-processed samples is around 0.9. It is important to note that an SSIM value lower than 0.9 is considered a significant alteration to the images and indicates an unsatisfactory level of image quality (Wang et al., 2004). In cases of strong post-processing, an adaptive attacker may be able to evade our method, but at the cost of substantially compromising the quality of the edited image. Consequently, our method maintains its effectiveness against adaptive attacks aimed at preserving the quality of the perturbed image.

6.8. Comparison to Laszkiewicz et al. (2023)

The related work by Laszkiewicz et al. (2023) also concentrates on detecting images generated by a specific model without relying on watermarks. In this section, we conduct the experiments for comparing our method to Laszkiewicz et al. (2023). The Inspected model (\mathcal{M}_1) here is SD v2-base. We consider the distinguishing the images generated by the

Table 9. Robustness against post-processing augmentations.

Augmentation	Acc	SSIM	PSNR	L1	L2
Saturation Factor 1.25	97.4%	0.9657	32.7953	0.0180	0.0008
Saturation Factor 1.50	94.0%	0.9276	27.3419	0.0334	0.0028
Saturation Factor 1.75	93.1%	0.8929	24.3338	0.0471	0.0055
Saturation Factor 2.00	90.6%	0.8611	22.3084	0.0593	0.0085
Saturation Factor 2.25	88.3%	0.8318	20.8090	0.0704	0.0117
Saturation Factor 2.50	84.8%	0.8045	19.6402	0.0804	0.0150
Contrast Factor 1.10	95.5%	0.9488	32.6094	0.0198	0.0006
Contrast Factor 1.15	92.7%	0.9213	29.4854	0.0283	0.0012
Contrast Factor 1.25	89.3%	0.8716	25.7423	0.0433	0.0028
Contrast Factor 1.50	79.8%	0.7729	21.1741	0.0729	0.0079
Gaussian Noise Std 0.01	97.4%	0.9896	46.1284	0.0039	2.4399e-05
Gaussian Noise Std 0.02	95.2%	0.9612	40.1485	0.0078	9.6707e-05
Gaussian Noise Std 0.03	93.3%	0.9204	36.6620	0.0116	0.0002
Gaussian Noise Std 0.04	87.8%	0.8731	34.1959	0.0154	0.0004
Gaussian Noise Std 0.05	80.7%	0.8238	32.2881	0.0192	0.0006
JPEG Compression Quality 95	96.5%	0.9730	38.6492	0.0082	0.0002
JPEG Compression Quality 90	95.2%	0.9576	36.4326	0.0107	0.0003
JPEG Compression Quality 80	94.1%	0.9335	33.9629	0.0142	0.0005
JPEG Compression Quality 70	90.8%	0.9155	32.5805	0.0165	0.0006
JPEG Compression Quality 60	90.2%	0.9009	31.6258	0.0184	0.0008
JPEG Compression Quality 50	89.3%	0.8885	30.9288	0.0199	0.0009
Brightness Factor 1.25	92.6%	0.9410	19.8790	0.0855	0.0108
Brightness Factor 1.35	90.7%	0.9042	17.4920	0.1129	0.0188
Brightness Factor 1.50	81.9%	0.8471	15.1579	0.1485	0.0322
Gaussian Blur Box Size 1	96.3%	0.8755	28.9077	0.0218	0.0017
Gaussian Blur Box Size 2	88.7%	0.7427	25.0875	0.0346	0.0041
Gaussian Blur Box Size 3	78.4%	0.6625	23.3526	0.0432	0.0062

Inspected model (\mathcal{M}_1) and that generated by different other models (\mathcal{M}_2). Here, SD v1-1+ means a fine-tuned version of the Stable Diffusion v1-1. The comparison results are reported in Table 10. The results demonstrate that our method outperforms the Laszkiewicz et al. (2023) It is understandable as the approach by Laszkiewicz et al. (2023) solely reverses the final layer of the model, causing it to lack access to some of the valuable information encoded within the weights of the other layers. In addition, Laszkiewicz et al. (2023) assume the last layer of the inspected model is invertible. Therefore, it is not applicable to the models that use non-invertible activation functions (such as the commonly-used ReLU function) in the last layer. Our method does not have this assumption.

6.9. Examples of Failure Cases

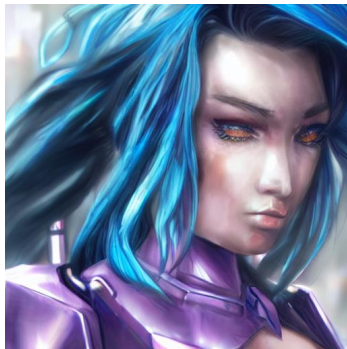
In this section, we demonstrate the examples of the failure cases of our method. The visualization results of the failure cases when distinguishing the images generated by SD-v2-base and SD-v1-5 can be found in Figure 6. We find that the potential reason for the FP and FN could be high brightness and high shape complexity in the images, respectively.

6.10. Details of the Used Text Prompts

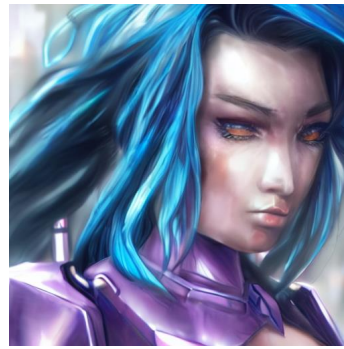
In Figure 2 and Table 2, we use 50 text prompts randomly sampled from PromptHero (Inc.). The detailed text prompts used can be found in Table 11 and Table 12. This prompts are with MIT License. They do not contain personally identifiable information or offensive content.

Table 10. Comparison to Laszkiewicz et al. (2023). Here, Model \mathcal{M}_1 is the inspected model, Model \mathcal{M}_2 is the other model.

Method	\mathcal{M}_1 :SD v2-base; \mathcal{M}_2 :SD v1-1	\mathcal{M}_1 :SD v2-base; \mathcal{M}_2 :SD v1-1+	\mathcal{M}_1 :SD v2-base; \mathcal{M}_2 :SD v1-4	\mathcal{M}_1 :SD v2-base; \mathcal{M}_2 :SD v1-5
Laszkiewicz et al. (2023)	92.7%	90.6%	92.1%	92.2%
LATENTTRACER (Ours)	98.2%	98.0%	98.0%	97.8%



False Positive, Original Image



False Positive, Inverted Image



False Negative, Original Image



False Negative, Inverted Image

Figure 6. Examples of the false positive and the false negative of our method.

How to Trace Latent Generative Model Generated Images without Artificial Watermark?

Table 11. Text prompts used in Figure 2 and Table 2 (Prompt 1-25).

Prompt 1	“cyber punk robot, dark soul blood borne boss, face hidden, RTX technology, high resolution, light scattering”
Prompt 2	“RAW photo of young woman in sun glasses sitting on beach, (closed mouth:1.2), film grain, high quality, Nikon D850, hyperrealism, photography, (realistic, photo realism:1. 37), (highest quality)”
Prompt 3	“full color portrait of bosnian (bosniak) woman wearing a keffiyeh, epic character composition, by ilya kuvshinov, terry richardson, annie leibovitz, sharp focus, natural lighting, subsurface scattering, f2, 35mm, film grain, award winning, 8k”
Prompt 4	“A silhouette of a woman of dark fantasy standing on the ground, in the style of dark navy and dark emerald, pigeoncore, heavily textured, avian - themed, realistic figures, medieval - inspired, photorealistic painting”
Prompt 5	“stained glass art of goddess, mosaic-stained glass art, stained-glass illustration, close up, portrait, concept art, (best quality, masterpiece, ultra-detailed, centered, extremely fine and aesthetically beautiful, super fine illustration), centered, epic composition, epic proportions, intricate, fractal art, zentangle, hyper maximalism”
Prompt 6	“An artistic composition featuring a person with orange hair casually squatting in a park, capturing their nonchalant expression, hot pants, The focus is on their unique sense of style, particularly their white panties peeking out from under their attire.”
Prompt 7	“a woman holding a glowing ball in her hands, featured on cgsociety, fantasy art, very long flowing red hair, holding a pentagram shield, looks a bit similar to amy adams, lightning mage spell icon, benevolent android necromancer, high priestess tarot card, anime goddess, portrait of celtic goddess diana, featured on artstation”
Prompt 8	“masterpiece, girl alone, solo, incredibly absurd, hoodie, headphones, street, outdoor, rain, neon,”
Prompt 9	“Halvard, druid, Spring, green, yellow, red, vibrant, wild, wildflowers masterpiece, shadows, expert, insanely detailed, 4k resolution, intricate detail, art inspired by diego velazquez eugene delacroix”
Prompt 10	“(8k, RAW photo, high sensitivity, best quality, masterpiece, ultra high resolution, fidelity: 1.25), upper body, cat ears, (night), rain, walk, city lights, delicate face, wet white shirt”
Prompt 11	“masterpiece, centered, dynamic pose, 1girl, cute, calm, intelligent, red wavy hair, standing, batik swimsuit, beach background,”
Prompt 12	“masterpiece, award winning, best quality, high quality, extremely detailed, cinematic shot, 1girl, adventurer, riding on a dragon, fantasy theme, HD, 64K”
Prompt 13	“((masterpiece:1.4, best quality))+, (ultra detailed)+, blue hair , wolfcut, pink eyes, 1 girl,cyberpunk city,flat chest,wavy hair,mecha clothes,(robot girl), cool movement,silver bodysuit,colorful background,rainy days,(lightning effect),silver dragon armour,(cold face),cowboy shot”
Prompt 14	“masterpiece, centered, concept art, wide shot, art nouveau, skyscraper, architecture, modern, sleek design, photography, raw photo, sharp focus, vibrant illustrations, award winning”
Prompt 15	“masterpiece, best quality, mid shot, front view, concept art, 1girl, warrior outfit, pretty, medium blue wavy hair, walking, curious, exploring city, london city street background, Fantasy theme, depth of field, global illumination, (epic composition, epic proportion), Award winning, HD, Panoramic,”
Prompt 16	“a couple of women standing next to each other holding candles, inspired by WLOP, cgsociety contest winner, ancient libu young girl, 4 k detail, dressed in roman clothes, lovely detailed faces, loli, high detailed 8 k, twin souls, cgsociety, beautiful maiden”
Prompt 17	“Tigrex from monster hunter, detailed scales, detailed eyes, anatomically correct, UHD, highly detailed, raytracing, vibrant, beautiful, expressive, masterpiece, oil painting”
Prompt 18	“Fashion photography of a joker, 1800s renaissance, clown makeup, editorial, insanely detailed and intricate, hyper-maximal, elegant, hyper-realistic, warm lighting, photography, photorealistic, 8k”
Prompt 19	“octane render of cyberpunk batman by Tsutomu nihei, chrome silk with intricate ornate weaved golden filiegree, dark mysterious background -v 4 -q 2”
Prompt 20	“a cat with a (pirate hat:1.2) on a tropical beach, ~*~Enhance~*~, in the style of Clyde Caldwell, vibrant colors”
Prompt 21	“masterpiece, portrait, medium shot, cel shading style, centered image, ultra detailed illustration of Hatsune Miku of cool posing, inkpunk, ink lines, strong outlines, bold traces, unframed, high contrast, cel-shaded, vector, 32k resolution, best quality”
Prompt 22	“(A bright vivid chaotic cyberpunk female, Fantastic and mysterious, full makeup, blue sky hair, (nature and magic), electronic eyes, fantasy world))
Prompt 23	“broken but unstoppable masked samurai in full battle gear, digital illustration, brutal epic composition, (expressionism style:1. 1), emotional, dramatic, gloomy, 8k, high quality, unforgettable, emotional depth”
Prompt 24	“studio lighting, film, movie scene, extreme detail, 12k, masterpiece, hyperrealistic, realistic, Canon EOS R6 Mark II, a dragon made out of flowers and leaves, beautiful gold flecks, colorful paint, golden eye, detailed body, detailed eye, multiple colored flowers”
Prompt 25	“Photo realistic young Farscape Chiana, kissy face, full Farscape Chiana white face paint, black shadowy eye makeup, white/gray lips, close-up shot, thin, fit, Fashion Pose, DSLR, F/2. 8, Lens Flare, 5D, 16k, Super-Resolution, highly detailed, cinematic lighting”

How to Trace Latent Generative Model Generated Images without Artificial Watermark?

Table 12. Text prompts used in Figure 2 and Table 2 (Prompt 26-50).

Prompt 26	"A retro vintage Comic style poster, of a post apocalyptic universe, of a muscle car, extreme color scheme, action themed, driving on a desert road wasteland, fleeting, chased by a giant fire breathing serpent like fantasy creature, in action pose, highly detailed digital art, jim lee"
Prompt 27	"cinematic CG 8k wallpaper, action scene from GTA V game, perfect symmetric cars bodies and elements, wheels rotating, real physics based image, extremely detailed 4k digital painting (design trending on (Agnieszka Doroszewicz), Behance, Andrey Tkachenko, GTA V game, artstation, BMW X6 realistic design"
Prompt 28	"the Hulk in his Worldbreaker form, his power and rage reach astronomical levels, amidst a cityscape in ruins, reflecting the destruction he can unleash"
Prompt 29	"masterpiece, portrait, medium shot, cel shading style, centered image, ultra detailed illustration of Hatsune Miku of cool posing, inkpunk, ink lines, strong outlines, bold traces, unframed, high contrast, cel-shaded, vector, 32k resolution, best quality"
Prompt 30	"Renaissance-style portrait of an astronaut in space, detailed starry background, reflective helmet."
Prompt 31	"A photo of a very intimidating orc on a battlefield, cinematic, melancholy, dynamic lighting, dark background"
Prompt 32	"A dark fantasy devil predator, photographic, ultra detail, full detail, 8k best quality, realistic, 8k, micro intricate details"
Prompt 33	"Hello darkness, my old friend, I've come to talk to you again, heart-wrenching composition, digital painting, (expressionism:1. 1), (dramatic, gloomy, emotionally profound:1. 1), intense and brooding dark tones, exceptionally high quality, high-resolution, leaving an indelible and haunting impression on psyche, unforgettable, masterpiece"
Prompt 34	"epic, masterpiece, alien friendly standing on moon, intricate organic neural clothes, galactic black hole background, {expansive:2} hyper realistic, octane, ultra detailed, 32k, raytracing"
Prompt 35	"Geometrical art of autumn landscape, warm colors, a work of art, grotesque, Mysterious"
Prompt 36	"a girl with face painting and a golden background is wearing makeup, absurd, creative, glamorous surreal, in the style of zbrush, black and white abstraction, daz3d, porcelain, striking symmetrical patterns, close-up -ar 69:128 -s 750"
Prompt 37	"Forest, large tree, river in the middle, full blue moon, star's, night, haze, ultra-detailed, film photography, light leaks, Larry Bud Melman, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowski"
Prompt 38	"a futuristic spacecraft winging through the sky, orange and beige color, in the style of realistic lifelike figures, ravencore, hispanicore, liquid metal, greeble, high definition, manicore, photo, digital art, science fiction -v 5. 2"
Prompt 39	"a close up of a person with a sword, a character portrait by Hasegawa Settan, featured on cg society, antipodeans, reimagined by industrial light and magic, sabattier effect, character"
Prompt 40	"Dystopian New York, gritty, octane render, ultra-realistic, cinematic -ar 68:128 -s 750 -v 5. 2"
Prompt 41	"ALIEN SPACECRAFT, WRECKAGE, CRASH, PLANET DESERT, ultra-detailed, film photography, light leaks, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowski"
Prompt 42	"an expressionist charcoal sketch by Odilon Redon, drawing, face only, a gorgeous Japanese woman, hint of a smile, noticeable charcoal marks, white background, no coloring, no color -ar 69:128 -s 750 -v 5. 2"
Prompt 43	"a man in a futuristic suit with neon lights on his face, cyberpunk art by Liam Wong, cgsociety, computer art, darksynth, synthwave, glowing neon"
Prompt 44	"The image features a bird perched on a branch, dressed in a suit and tie. The bird is holding a cup of hot coffee, in its beak. The coffee cup emits smoke. The scene is quite unusual and whimsical. The bird's attire and the presence of the cup create a sense of humor and playfulness in the image."
Prompt 45	"carnage, a formidable supervillain, symbiote, bloody, psychopathic, unstoppable, mad, sharp teeth, epic composition, dramatic, gloomy, in the style of mike deodato, realistic detail, realistic hyper-detailed rendering, realistic painted still lifes, insanely intricate"
Prompt 46	"A disoriented astronaut, lost in a galaxy of swirling colors, floating in zero gravity, grasping at memories, poignant loneliness, stunning realism, cosmic chaos, emotional depth, 12K, hyperrealism, unforgettable, mixed media, celestial, dark, introspective"
Prompt 47	"an abstract painting of a beautiful girl, in the style of Pablo Picasso, masterpiece, highly imaginative, dada, salvador dali, i can't believe how beautiful this is, intricate -ar 61:128 -s 750 -v 5. 2"
Prompt 48	"made by Emmanuel Lubezki, Daniel F Gerhartz, character of One Piece movie, Monkey D. Luffy, in straw hat, cinematic lighting, concept photoart, 32k, photoshoot unbelievable half-length portrait, artificial lighting, hyper detailed, realistic, figurative painter with intricate details, divine proportion, sharp focus, Mysterious"
Prompt 49	"a very detailed image of a female cyborg, half human, half machine, very detailed, with cables, wires, mechanical elements in the head and body, dynamic light, glowing electronics, 4 k, inspired by H. r. Giger and Jean ansell and justin Gerard, photorealistic"
Prompt 50	"A beautiful photo of an lion that got lost in the amazon rainforest, rain, mist, 8k, sharp intricate details, masterpiece, imaginative, raytracing, octane render, studio lighting, professionally shot nature photo, godrays, hyperrealistic, ultra high quality, realism, wet, dripping water, wandering through the undergrowth"