

Generative Watermarking Against Unauthorized Subject-Driven Image Synthesis

Yihan Ma¹ Zhengyu Zhao² Xinlei He¹ Zheng Li¹ Michael Backes¹ Yang Zhang¹

¹CISPA Helmholtz Center for Information Security ²Xi'an Jiaotong University

Abstract

Large text-to-image models have shown remarkable performance in synthesizing high-quality images. In particular, the subject-driven model makes it possible to personalize the image synthesis for a specific subject, e.g., a human face or an artistic style, by fine-tuning the generic text-to-image model with a few images from that subject. Nevertheless, misuse of subject-driven image synthesis may violate the authority of subject owners. For example, malicious users may use subject-driven synthesis to mimic specific artistic styles or to create fake facial images without authorization. To protect subject owners against such misuse, recent attempts have commonly relied on adversarial examples to *indiscriminately* disrupt subject-driven image synthesis. However, this essentially prevents any benign use of subject-driven synthesis based on protected images.

In this paper, we take a different angle and aim at protection without sacrificing the utility of protected images for general synthesis purposes. Specifically, we propose GenWatermark, a novel watermark system based on jointly learning a watermark generator and a detector. In particular, to help the watermark survive the subject-driven synthesis, we incorporate the synthesis process in learning GenWatermark by fine-tuning the detector with synthesized images for a specific subject. This operation is shown to largely improve the watermark detection accuracy and also ensure the uniqueness of the watermark for each individual subject. Extensive experiments validate the effectiveness of GenWatermark, especially in practical scenarios with unknown models and text prompts (74% Acc.), as well as partial data watermarking (80% Acc. for 1/4 watermarking). We also demonstrate the robustness of GenWatermark to two potential countermeasures that substantially degrade the synthesis quality.

1 Introduction

Large text-to-image generative models, such as DALL·E 2 [46], Latent Diffusion [48], and Stable Diffusion [48], have attracted increasing attention [1, 4, 9, 13, 30, 39]. These models take as input a natural language description, referred to as a prompt, and produce high-quality images matching that description. Text-to-image models have significantly revolutionized the AI art industry owing to their substantial generation capacity and ease of use. For example, an AI-generated image took the first place in the digital category at the Colorado



Figure 1: Real-world examples of misuse of subject-driven image synthesis. Left: A piece of original artwork by Hollie Mengert and a synthesized one. Right: A synthesized picture of Elon Musk dating GM CEO Mary Barra.

State Fair, beating all (human) artists.¹

One intriguing extension to the generic text-to-image models is the subject-driven synthesis, which makes it possible to “personalize” the image synthesis [20, 40, 50]. In subject-driven synthesis, the model takes as input a set of images sharing the same subject in addition to the text prompt. The model learns to synthesize images of that subject in novel contexts, which are specified by the text prompts. Popular subject-driven models, such as Textual Inversion [20] and DreamBooth [50] are able to synthesize high-quality images with diverse contents by fine-tuning an off-the-shelf text-to-image model with only a few images.

Subject-driven synthesis serves as an easy-to-use and creative tool for users to synthesize images based on their own needs and potentially benefit from the synthesis [3]. For example, artists may authorize third-party AI-powered design services to help accelerate the artwork production process, especially in the animation industry. Individuals can also authorize third-party AI-powered photo editing services to efficiently synthesize images depicting them in varied novel scenes, e.g., tourist attractions.

However, the images may be leaked and misused to fine-tune a text-to-image model for fabricating images with potentially unwanted content. For example, a Reddit user published a DreamBooth model that is fine-tuned based on the

¹<https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>

artwork from an American artist Hollie Mengert.² This has brought huge disputes because anyone can use that model to synthesize artwork in her style *without her authorization*.³ In another real-world incident with a person as the subject, a synthesized picture showing Elon Musk is dating GM CEO Mary Barra has been posted on Twitter,⁴ as shown in Figure 1.

To protect subject owners against misuse of their images for subject-driven synthesis, concurrent studies [33, 35, 55] rely on adversarial examples, i.e., cloaking the subject owners’ images by adding adversarial perturbations. In this way, the subject-driven synthesis fails because the model instead captures the adversarial image representations rather than the correct ones. However, such protections are indiscriminate, i.e., they disrupt not only the malicious use of protected images but also their benign use. Therefore, it cannot address the above problem of unauthorized image synthesis.

Our Work. In this paper, instead of completely disrupting subject-driven synthesis on the protected images, we prevent only their unauthorized use, but with little impact on their authorized utility. Watermarking is intuitively applicable to this task by injecting an invisible watermark into source images such that the owner’s authenticity can be verified by inspecting the watermark in their copies that are potentially manipulated [5, 58]. Moreover, the nature of personalized synthesis makes it easy to trace potentially unauthorized (synthesized) images since they share the same subject, i.e., artistic style or human face, with the original images. However, we find that existing watermarks can hardly survive subject-driven synthesis because the image content (except for the subject) can be freely changed according to the text prompt (see detailed discussions in Section 2.2).

Therefore, we propose GenWatermark, a novel generative watermarking approach based on a watermark generator and a detector. Specifically, the generator learns an image watermark in the form of additive perturbations, and the detector tries to distinguish watermarked images from normal images.

In our evaluation, we conduct extensive experiments to validate the effectiveness of GenWatermark in varied scenarios. We test two well-studied tasks corresponding to two kinds of authority-sensitive subjects, i.e., *artistic style* and *human face*. In particular, we show that our GenWatermark learned on large-scale human face data is effective in both tasks. Besides, we conduct the evaluation on two representative subject-driven models, i.e., Textual Inversion and DreamBooth, with diverse text prompts. In the basic scenario where the subject owner knows the model and prompts the adversary would use, our GenWatermark achieves perfect detection accuracy, i.e., above 98% averaged over two tasks and two models. Even in the most challenging scenario where neither the model nor prompts are known, GenWatermark still guarantees an accuracy of about 74%, which is much higher than the random chance for binary classifica-

tion, i.e., 50%. Note that the model utility is well maintained in all of the above scenarios. For example, the Fréchet Inception Distance (FID) [25] between the synthesized and input images before and after the watermarking only changes by less than 1%. Extensive visualizations of image examples also support this conclusion (see Figure 9 as an example).

We further consider more practical scenarios. For example, we show that when only half of the input images are watermarked, GenWatermark still ensures a detection accuracy of about 90%. Additional results further validate the robustness of our GenWatermark to two potential countermeasures. First, when the adversary deliberately injects random noises into the clean input images, our GenWatermark would still classify them as non-watermarked images with high accuracy above 80%. Second, when the adversary tries to eliminate the watermarks by image transformations, our GenWatermark still achieves a high accuracy above 75%. Moreover, we show that the last two countermeasures lead to obvious visual artifacts in both the input and synthesized images.

Our contributions are summarized as follows.

- We propose GenWatermark, the first watermark approach to protecting images from unauthorized subject-driven image synthesis. Beyond existing methods, GenWatermark maintains the utility of the protected images for benign image synthesis.
- We demonstrate the effectiveness of GenWatermark in two synthesis tasks with two representative subject-driven synthesis models. We particularly consider realistic scenarios with unknown target models and prompts, as well as with only partial watermarked input images.
- We demonstrate the robustness of GenWatermark against two potential countermeasures, which are based on adding noise to mislead the detector or using image transformations to disrupt our watermarks.

2 Related Work

2.1 Protection Against Malicious Subject-Driven Image Synthesis

Although subject-driven models make it possible for users to synthesize images easily based on their own needs, the potential malicious use of such tools may cause severe threats. As discussed, adversaries may use subject-driven synthesis to mimic the artistic style of specific artists or to generate fake images containing specific faces. To protect users against such malicious use, existing studies [33, 35, 55] have commonly relied on adversarial examples [22, 57]. The basic idea is to add carefully crafted perturbations to the users’ images such that the subject-driven models cannot learn the correct image representations anymore during model fine-tuning. These three studies are different in their technical details. Specifically, both AdvDM [35] and Anti-DreamBooth [33] assume access to the diffusion process of the target model, and (approximate) gradients can be computed for perturbation optimization. Differently, Glaze [55] assumes no access

²https://www.reddit.com/r/StableDiffusion/comments/yaquby/2_d_illustration_styles_are_scarse_on_stable/

³<https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/>

⁴<https://twitter.com/blovereviews/status/1639988583863042050>

to the diffusion process but only a generic image feature extractor. In this case, the perturbations are optimized to cause disruptions in the latent space of that feature extractor. Glaze is specifically designed to protect artists by concentrating the perturbations on artistic style-specific features that are isolated based on a style transfer model.

A common limitation of the above approaches is that they do not discriminate the malicious use of subject-driven synthesis from its benign use. Instead, they completely disable the subject-driven synthesis on the protected images. Our work is essentially different because we focus on preventing only malicious use of subject-driven synthesis while maintaining the possibility for potential benign use. In addition, our GenWatermark is the first approach to protecting images against subject-driven synthesis for both tasks of human face identity and artistic style, and it remains highly effective even in cross-task settings. Technically, it requires no knowledge about the internal details (e.g., diffusion process and gradients) of the target model but only use the output images.

2.2 Image Watermarking

Traditional image watermarking mainly relies on steganography [6], achieved by Fourier transformation [8, 12], JPEG compression [2], or least significant bits modification [26, 43]. Among them, Fourier transformation and JPEG compression are known to cause visible visual artifacts, making them not inappropriate in general. The least significant bits modification method hides the watermark in the last bit of each pixel, which is invisible but not robust to even simple noise. These traditional watermarks have been shown to be ineffective in the context of generative synthesis [64]. Therefore, deep learning-based image steganography methods based on an encoder and a decoder were proposed [5, 23, 58, 60, 64, 68].

These approaches were originally designed against generative adversarial networks (GANs)-based image synthesis, and we find it does not work well in our context with the subject-driven model. In particular, our exploratory results demonstrate that the state-of-the-art approach from [5, 23, 58, 60, 64, 68] only achieves a bit-wise accuracy around the random guess level, i.e., 50%. We hypothesize that there are two main reasons. First, the training of the subject-driven image synthesis only requires a small number of images (e.g., 30 in our work), which is not enough to train a useful image steganography encoder and decoder. Second, with the semantic guidance of the text prompt, subject-driven image synthesis leads to more diverse changes of image content than the GAN tasks considered in [64], such as deepfakes and image-to-image translation. One recent preprint [18] also explores watermarking techniques for text-to-image diffusion models. However, [18] does not consider a subject-driven task, so their authority is defined with respect to the model instead of a subject.

3 Preliminary

3.1 Text-to-Image Synthesis

Since first proposed in 2015 [38], text-to-image models have been widely explored and used for image synthesis [14, 16, 27, 34, 45, 46, 48, 59, 62, 65, 69]. In text-to-image synthesis, natural language descriptions, namely *prompts*, are used as the input to generate synthetic images. Normally, text-to-image models contain two parts: a text embedding module, e.g., CLIP’s [44] text encoder or BERT [15], that turns natural language into corresponding semantic features, and a conditional image generator that produces synthetic images based on the semantic features. Conventional work leverages generative adversarial networks (GANs) or variational autoencoders (VAEs) as the image generator [14, 16, 45, 46, 59, 69], but recent work has been shifting to diffusion models and shows great advantages in producing high-quality images with rich details [46–48, 52].

Diffusion Models. The core of a diffusion model is a stochastic differential process called the *diffusion process*, which is composed of 2 phases: the forward diffusion process and the backward denoising process [56]. Given a distribution of real images $q(\mathbf{x})$, the forward diffusion process employs a Markov chain of diffusion steps to slowly add random noise ϵ to a data sample $\mathbf{x}_0 \sim q(\mathbf{x})$ in T steps, producing a sequence of noised samples $\mathbf{x}_1, \dots, \mathbf{x}_T$:

$$\mathbf{x}_T = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \epsilon. \quad (1)$$

Here, the step sizes are controlled by a variance schedule $\{\beta_t \in (0, 1)\}_{t=1}^T$, and the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is sampled from the Gaussian distribution. Note that the data sample becomes more unrecognizable as with a larger step T . As a result, when $T \rightarrow \infty$, \mathbf{x}_T will follow the isotropic Gaussian distribution.

The backward denoising process is the reverse of the forward diffusion process, which is to generate \mathbf{x}_T given \mathbf{x}_{T+1} . With T being sufficiently large, we are able to create a true sample from $q(\mathbf{x})$ by a Gaussian noise input $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. The key to this process is to train a neural network $\epsilon_\theta(\mathbf{x}_{t+1}, t)$ that can estimate the injected noise ϵ in step t . Specifically, this denoising neural network is trained to minimize the L2 distance between the real noise and the estimated noise using the following loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon \sim \mathcal{N}(0, 1)} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_{t+1}, t)\|_2^2 \right], \quad (2)$$

where t is uniformly sampled from all time steps $\{1, \dots, T\}$. So far, there are two representative diffusion-based text-to-image models: Latent Diffusion and Stable Diffusion.

- **Latent Diffusion** runs the diffusion process in the latent space instead of the pixel level, which leads to less training cost and faster inference speed while maintaining high-quality generation.
- **Stable Diffusion** is a variant of Latent Diffusion and they share a similar architecture. The major difference is that the text encoder in Stable Diffusion is the CLIP’s text encoder [44] instead of BERT [15].

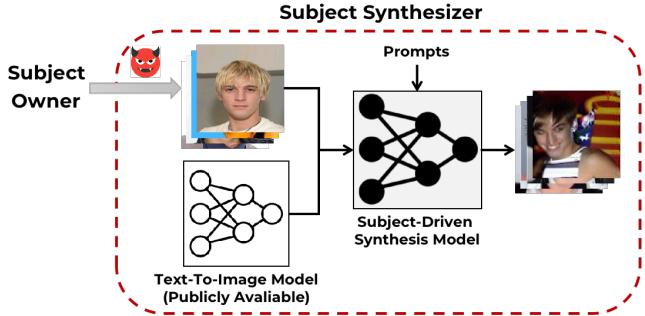


Figure 2: High-level overview of subject-driven image synthesis attack scenario.

3.2 Subject-Driven Image Synthesis

Subject-driven image synthesis leverages the basic text-to-image models and takes additional images as the input. Given a few images of the same subject, the subject-driven image synthesis can learn the subject in the input samples, and synthesize novel renditions of the learned subject in different contexts, using different prompts to guide the generation direction. In real-world applications, the learned subject can either be the face of a person, or the artistic style of an artist, as shown in Figure 1.

The workflow of the subject-driven image synthesis is normally as follows. First, subject-driven image synthesis learns the distribution of the subject shared in a set of images. The learned distribution can be represented by a pseudo-word, referred to as “[V]”. Once learned, a new prompt containing “[V]” can be used as input to synthesize new images depicting the same subject with different contexts. So far, there are two popular subject-driven image synthesis models: Textual Inversion [20] and DreamBooth [50]. Their detailed information is described as follows.

Textual Inversion. Textual Inversion learns the embedding of the subject “[V]” while freezing the parameters of the text-to-image model, here, Latent Diffusion. The training takes input pairs of “[V]” and images. During synthesis, if we want to synthesize a painting of a daisy in the same style of an artist [V], the new prompt could be something like “A painting of daisy in the style of [V]”.

DreamBooth. Unlike textual inversion, DreamBooth is based on fine-tuning model parameters and uses Stable Diffusion as the backbone. Specifically, the prompt is designed to contain an additional term denoting the category of the subject. In this case, the above prompt example becomes “A painting of a daisy in the style of [V] painting”.

4 Methodology

In this section, we introduce the methodology of our new watermark method GenWatermark. Before introducing the technical details GenWatermark, we first specify the threat model and discuss the design intuition, where we also identify two potential challenges in designing an effective watermarking approach in the context of subject-driven synthesis.

4.1 Threat Model

In our threat model, there are two parties involved: the subject owners and the subject synthesizers. Subject owners use the generator of GenWatermark to obtain watermarks and inject them into their images before authorizing specific subject synthesizers to use those images. Subject synthesizers synthesize images targeting the protected subject by fine-tuning a subject-driven synthesis on those watermarked images. Adversaries refer to the subject synthesizers who do not obtain authorization from the owners. Figure 2 provides a high-level overview of our attack scenario.

Subject Owners. The goal of the subject owner is to benefit from authorizing specific third parties to use their images for subject-driven synthesis. For example, artists may authorize AI-powered design services to assist their artwork production. The subject owners use our system to generate the watermark and inject it into their images. Then, they can track the potential unauthorized use by detecting if the watermark appears in synthesized images.

We assume the subject owner has full access to our watermark generator and detector. They can further improve the watermark detector by fine-tuning it on additional images that are synthesized using a popular subject-driven synthesis service, such as Textual Inversion and DreamBooth. For the knowledge of the subject-driven model and text prompts, we consider 4 scenarios with gradually relaxed assumptions: 1) Both are known; 2) Only the model is known; 3) Only the prompts are known; 4) Neither is known.

Subject Synthesizers. The goal of the subject synthesizer is to train a subject-driven model that synthesizes high-quality images of the target subject. A benign subject synthesizer obtains authorization from the subject owner but a malicious subject synthesizer achieves the goal without authorization. For example, the malicious subject synthesizer can mimic the style of an artist based on the leaked artwork from third-party AI-powered design services. We assume the subject synthesizer has access to a publicly-available subject-driven model and also protected images from the target subject. The subject synthesizer is also assumed to have sufficient computational resources to fine-tune the model.

4.2 Design Intuition

In general, a successful watermark requires the *robustness* and *invisibility* [64]. In our context of subject-driven synthesis, robustness means the watermark should be robust to substantial changes in the image content caused by the subject-driven text-to-image generation, and invisibility means the watermark should have little impact on the quality of both the watermarked (input) images and synthesized (output) images.

More specifically, we identify two potential challenges in achieving an effective and practical watermark system for subject-driven image synthesis. The first challenge is to incorporate the image synthesis process into the optimization of our watermark system. Existing protections based on adversarial perturbations [33, 35] incorporate the internal (diffusion) process of the target model and optimize their perturbations based on (approximate) gradients. This is computa-

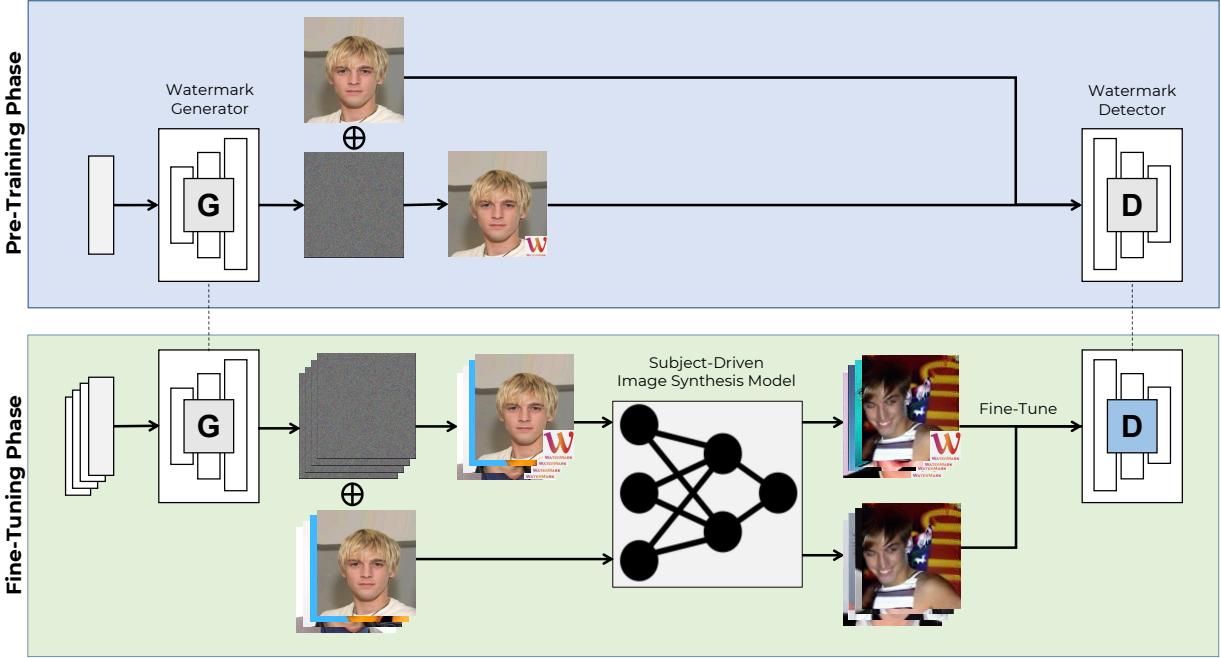


Figure 3: Overview of GenWatermark. The pipeline of GenWatermark can be divided into 2 phases. In the first phase, we train a watermark generator and detector using a large-scale dataset. The trained watermark generator is then used to produce watermarked images for the subject we want to protect. In the second phase, each subject is allowed to “personalize” the pre-trained detector by fine-tuning it on images synthesized from their own images.

tionally inefficient and requires knowledge of the diffusion process. However, it makes more practical sense to have a protection method that exhibits high efficiency and ease of optimization since the subject owner, e.g., an artist, does not have large computational resources and rich knowledge of computer science. Another specific reason for incorporating the image synthesis process is that the image synthesis process captures the unique information for each individual subject, and we want to use such information to achieve a subject-personalized watermark design.

The second challenge is to maintain the watermark’s effectiveness in realistic scenarios with unknown settings. Specifically, the watermark system should generalize well to different target models and text prompts because the subject owner has no control of the subject-driven model and text prompts that will be used by the malicious subject synthesizer for their unauthorized image synthesis. In addition, the watermark system should generalize well to different data distributions. For example, using only a few pieces of artwork from a specific artist may not be enough to train a powerful watermark system for that artist.

4.3 Our GenWatermark

We propose GenWatermark, a generative watermarking approach against unauthorized subject-driven synthesis. To achieve *robustness*, GenWatermark follows a generative framework consisting of a watermark generator and a detector. The generator and the detector are jointly trained such that the generator can generate strong watermarks and the detector can detect the watermark left in synthesized images.

To achieve *invisibility*, we follow related work [10, 32, 49, 55] to constrain the perturbation budget measured by some perceptual metrics.

In particular, we aim to address the above two potential challenges. For the first challenge, we incorporate the process of subject-driven synthesis into our GenWatermark by allowing each subject to fine-tune the detector on their own. The fine-tuning dataset consists of synthesized images that are obtained from the subject-driven synthesis using both clean and watermarked images with diverse text prompts. For the second challenge, the fact that GenWatermark is not optimized based on the internal details of the subject-driven synthesis makes it compatible with unknown target models. In addition, we develop GenWatermark with diverse prompts and gather images from multiple subjects or even from other domains with sufficient public data, such as the face image domain.

As shown in Figure 3, the whole learning pipeline can be divided into 2 phases: pre-training and fine-tuning.

Phase 1: Pre-training the generator and detector. In this phase, we jointly train a watermark generator \mathbf{G} and a watermark detector \mathbf{D} . For initialization, we randomly choose a latent code \mathbf{z} [11, 21, 29] as the input to generate a watermark \mathbf{w} with the same size as the input image. Here we adopt the well-known GAN generator [21] as our watermark generator.

During training, this watermark \mathbf{w} is added to each original (clean) image \mathbf{x} to obtain the corresponding watermarked image \mathbf{x}_w :

$$\mathbf{x}_w = \mathbf{x} \oplus \mathbf{w} \quad (3)$$

The Learned perceptual image patch similarity (LPIPS) [66]

is engaged to constrain the invisibility of the watermark when training the generator \mathbf{G} . LPIPS measures the distance between image patches represented in the latent space of a conventional neural network (CNN) [10, 32, 49]. LPIPS is known to align better with human visual perception than other perceptual metrics, such as the naive L_p distance and SSIM [7, 28, 31, 51]. The optimization of the generator is as follows:

$$\mathcal{L}_G = \max(LPIPS(\mathbf{x}, \mathbf{x}_w) - p, 0), \quad (4)$$

where p controls that desired level of invisibility, and a smaller p represents lower visibility.

The detector \mathbf{D} is trained to distinguish between clean and watermarked images. Specifically, we adopt a convolutional neural network (CNN) for the detector and the binary cross-entropy (BCE) loss for optimization.

$$\mathcal{L}_D = -(1 - y) \log(1 - \hat{y}) - y \log(\hat{y}), \quad (5)$$

Where y represents the ground truth and \hat{y} means the predicted value. Then, the overall loss function can be formulated as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_G + \mathcal{L}_D, \quad (6)$$

where α is a hyperparameter to balance the two different objectives. Finally, we use the pre-trained generator \mathbf{G} to watermark images for the subject we want to protect.

Phase 2: Fine-tuning the detector. In this phase, we further improve the performance of the detector \mathbf{D} by allowing each individual subject to “personalize” the detector on their own images. First, we use a clean image set \mathbf{X} and its corresponding watermarked set \mathbf{X}_w to train two subject-driven models, i.e., a clean model \mathcal{M} and a watermarked model \mathcal{M}_w . Then, we use these two models to synthesize images with multiple prompts, resulting in two corresponding synthesized image sets, denoted as \mathbf{S} and \mathbf{S}_w . Finally, we use \mathbf{S} and \mathbf{S}_w to fine-tune the detector. Note that the generator remains unchanged.

In this way, protecting each individual subject becomes a downstream task that can be solved efficiently. Specifically, this fine-tuning has two main advantages. First, it improves the watermark detection performance by incorporating the process of subject-driven synthesis. See Section 5.4 for experimental results. Second, it isolates the protections of different subjects, making sure that the authority of images is hard to be claimed by anyone except the real owner. See more relevant discussions in Section 6.

5 Experiments

In this section, we evaluate the performance of our proposed GenWatermark against malicious subject-driven image synthesis. First, we present detailed information about experimental settings. Then, we evaluate the performance of GenWatermark from two perspectives: the watermark detection accuracy and the image synthesis quality. Furthermore, we conduct comprehensive ablation studies to analyze the effects of the attack assumptions and main hyperparameters.

5.1 Experiment Setup

All our experiments were run on one NVIDIA A100 GPU. Specifically, the pre-training phase needs to run only one

time and takes about 6 hours. In the fine-tuning phase, the training of the subject-driven model is run independently for each subject, and it takes about 30 minutes for Textual Inversion and 2 hours for DreamBooth. The inference time is about 10 minutes. Then, the fine-tuning of our watermark detector for each subject takes about 1 hour. In sum, for each subject owner, the fine-tuning phase takes less than 4 hours.

Tasks and Datasets. We consider two well-studied tasks corresponding to two kinds of subjects: *artistic style* and *human face* [33, 54, 55]. In the human face task, we use the large-scale dataset CelebA [37], which contains 202,599 face images of 10,177 celebrities. In this case, the subject-driven model is used to learn the unique face of each celebrity [V] and synthesize new images of the celebrity in different scenes.

In the artistic style task, we use WikiArt [53], which contains 52,757 paintings from 195 different artists in 27 genres. Each painting is associated with a caption that describes the content. In this case, the subject-driven model is used to learn the unique artistic style of each artist [V] and synthesize new images in the same style with different contents.

Prompt Design. We design text prompts involving diverse content. In the human face task, we randomly select 30 prompts from Lexica, a popular search engine that contains millions of AI-synthesized images and their corresponding prompts.⁵ We refine each prompt by removing irrelevant information but only keeping a target context to form our final prompt following “*A photo of [V] target context*”. For example, for an original prompt “*smiling softly, 8k, irakli nadar, hyperrealism, hyperdetailed, ultra realistic*”, our refined prompt is “*A photo of [V] smiling softly*”. Since DreamBooth requires an additional term denoting the category of the subject, the prompt becomes “*A photo of [V] face smiling softly*”

In the artistic style task, we randomly select 30 prompts (image descriptions) from WikiArt. Then, we refine the prompt to follow “*A painting of target description in the style of [V]*”. For example, the prompt can be “*A painting of the valley of the river slavyanka in the style of [V]*”. For DreamBooth, the prompt becomes “*A painting of the valley of the river slavyanka in the style of [V] painting*”. Table 5 in Appendix A lists all 30 prompts used in the human face task and 30 prompts used in the artistic style task.

Synthesis Setup. In the human face task, we construct the training set by randomly selecting 4 celebrities that have more than 30 images. The IDs of these celebrities are 14, 15, 17, and 21 in CelebA. In the artistic style task, we select a total of 27 representative artists corresponding to all 27 different genres. Table 4 in Appendix A lists the names of the selected artists with their corresponding genres. For each of these selected subjects, we randomly sample 30 images to train their (clean or watermarked) subject-driven models for 6500 steps following the previous work [20]. In total, we get 8 (4 clean and 4 watermarked) models for the human face task and 54 (27 clean and 27 watermarked) models for the artistic style task. For each of these models, we will synthesize images for each of the 30 prompts in our follow-

⁵<https://lexica.art/>

ing experiments. The resolution of the synthesized image is 256x256.

Watermark Setup. As mentioned in [Section 4](#), the training pipeline of our GenWatermark consists of two phases: pre-training and fine-tuning. We adopt the generator from the vanilla GAN [21] as the watermark generator and ResNet34 [24] as the detector. In the pre-training phase, we jointly train the watermark generator and detector with 200,000 images from CelebA since a large number of images are required for an effective GenWatermark. In our experiments, we show that this GenWatermark trained on CelebA is also effective in watermarking WikiArt paintings. All the training images are resized to 256x256. We choose the invisibility level $p = 0.05$ and the balancing factor $\alpha = 1$. The effects of these hyperparameters on the performance of GenWatermark will be discussed in [Section 5.4](#). In the fine-tuning phase, for each subject, we fine-tune the detector with 1,000 synthesized images (i.e., 40 different images for 25 prompts) from the clean model \mathcal{M} and also 1,000 synthesized images from the watermarked model \mathcal{M}_w .

Attack Scenarios. As discussed in the threat model ([Section 4.1](#)), we consider 4 scenarios with gradually relaxed assumptions on the knowledge the subject owner has about the model and prompts the adversary (subject synthesizer) would use. Their settings are specified as follows.

In **Scenario 1**, both the model and the prompts are known. Here we synthesize another 10 images using the same model for each of the same 25 prompts as in our detector fine-tuning.

In **Scenario 2**, only the model is known but the prompts are unknown. Therefore, different from Scenario 1, we synthesize another 50 images for each of the rest 5 prompts (out of 30).

In **Scenario 3**, only the prompts are known but the model is unknown. Therefore, different from Scenario 1, we use Textual Inversion (DreamBooth) for the adversary but DreamBooth (Textual Inversion) for our detector fine-tuning.

In **Scenario 4**, neither the model nor the prompts are known. Therefore, both the model and the prompts differ following Scenario 2 and Scenario 3.

Evaluation Metrics. We evaluate the performance of our GenWatermark from two perspectives. First, we want GenWatermark designed for a specific subject to have high detection accuracy on that subject. In addition, it should achieve uniqueness, i.e., not falsely detecting the watermarked images generated for other subjects. Second, we want the subject-driven model that is trained on watermarked images to maintain its synthesis quality. Third, we want the watermark detection to be unique to specific owners, i.e., the watermark detector designed for one subject should not falsely detect the watermark generated for another subject.

- **Watermark Detection Accuracy** is defined as the fraction of correctly classified synthetic images among all the test synthetic images. **Uniqueness** is measured by the accuracy of a subject’s detector in differentiating their own watermarked images from other subjects’ watermarked images.

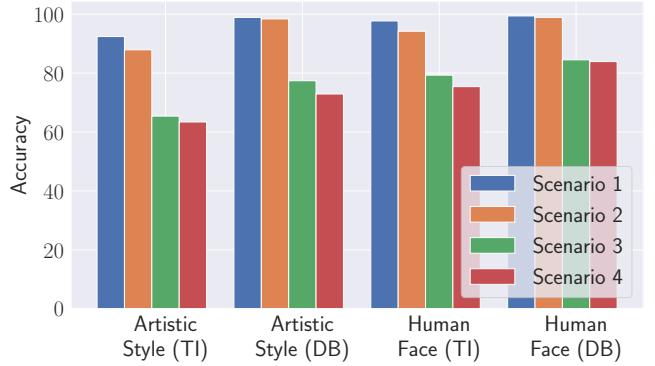


Figure 4: Watermark detection accuracy (%) on Textual Inversion and DreamBooth in the artistic style and human face tasks considering four attack scenarios.

- **Image Synthesis Quality** is measured based on the Fréchet Inception Distance (FID) [25] calculated between synthesized images and input images before and after the watermarking. If the FID scores before and after watermarking are similar, the watermark is considered to have little impact on the image synthesis quality.

5.2 Results on Watermark Detection Accuracy

Overall Results. [Figure 4](#) presents the overall detection accuracy on two target models for both the human face and artistic style tasks in four different scenarios, as introduced in [Section 4.1](#) and set up in [Section 5.1](#). As can be seen, in Scenario 1, where the model and prompts the adversary would use are both known to the subject owner, our GenWatermark achieves perfect results with an accuracy of above 98% on average. Even in Scenario 4, the most challenging one with unknown models and prompts, GenWatermark still maintains an accuracy of about 74%, much higher than the random chance for binary classification, i.e., 50%. As expected, in Scenario 2 or Scenario 3, where either the model or prompts are known, the performance is in between the two extreme scenarios. When comparing Scenario 2 and Scenario 3, we can see that knowing the model is much more helpful than knowing the prompts. In other words, transferring between models is more difficult than between prompts. Overall, results for Scenarios 2-4 suggest the good transferability of GenWatermark to unknown models and prompts.

Different Tasks. For the two tasks, we can see that the results for the artistic style are consistently worse than for the human face. This could be because we use the CelebA face data for pre-training in both tasks. Moreover, since the artistic style synthesis leads to more severe changes in image content, i.e., the watermark is disrupted more in the artistic style task. Consistently, from [Figure 5](#), we observe that FID between the input images and synthesized images for the artistic style is about 2-3 times higher than that for the human face.

Different Target Models. For the two target models, we can see that the results for DreamBooth and Textual Inversion are quite different, especially in the artistic style task. To



Figure 5: Image synthesis quality measured by the FID score for synthesis with clean inputs and with watermarked inputs. The FID is calculated between the (clean or watermarked) inputs and synthesized outputs.



Figure 6: Images synthesized based on the original artwork from Edward Hopper (top) by Textual Inversion (middle) and DreamBooth (bottom). The prompt content is “*A painting of the reader in the style of [V]*”.

shed more light on this difference, we visualize examples of synthesized images from these two models in Figure 6. As can be seen, DreamBooth images inherit fine-grained content from the input images, such as the tables and windows. The original work of DreamBooth [50] also shows its power in learning concepts. Differently, Textual Inversion images mainly capture the holistic style, and their content solely depends on the text prompts with little influence by the fine-grained content in the input images. This difference also supports our above finding that transferring between models, as shown by the results for Scenario 3 and Scenario 4, is relatively difficult.

Different Subjects. All the above results are averaged on all 27 artists in the artistic style task and 4 celebrities in the human face task. Since each synthesis model is driven independently by an individual subject, we are also interested in the performance distribution of GenWatermark over different subjects. As can be seen in Figure 7, the human face task yields very stable results while the artistic style task yields result with larger variances. This is understandable

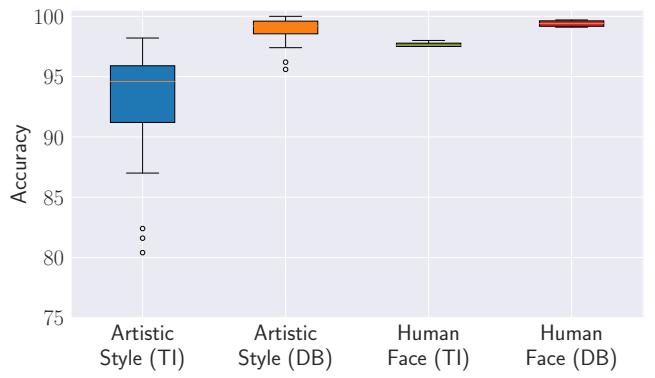


Figure 7: Watermark detection accuracy (%) for Scenario 1. Results for scenarios 2-4 can be found in Figure 17.

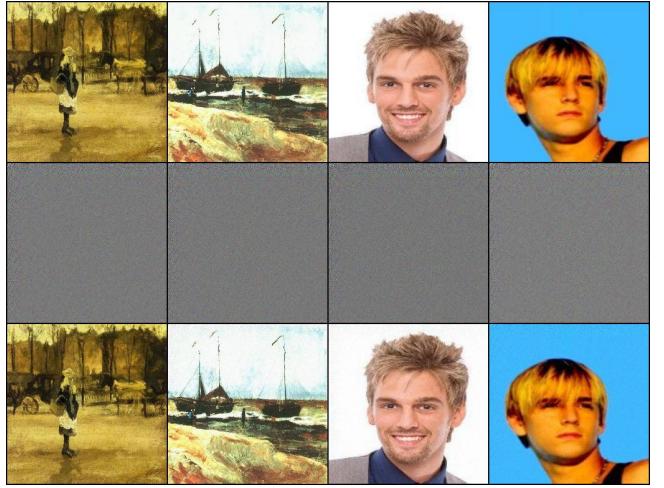


Figure 8: Clean inputs (top), added watermarks (middle), and the resulting watermarked inputs (bottom). Watermarks are multiplied by 10 for better visualization.

because artistic styles selected from different genres are very different, but the human face images are mostly with a fixed front view. In addition, DreamBooth yields more stable results than Textual Inversion. This might be due to the fact that, as discussed above, DreamBooth also learns additional, finer-grained information beyond the holistic style. In the future, one possible way to further improve the detection accuracy for each subject is to adapt the training strategy, e.g., the number of training images, learning rates, and training epochs, to specific characteristics of different subjects.

Watermark Uniqueness. An important property of the watermarking technology is to ensure that the watermark is unique to each individual subject. To measure such uniqueness in our case, we measure the classification accuracy of a detector designed for a specific subject in distinguishing its own watermarked images from those generated for other subjects. Specifically, for the artistic style task, we use the detector that is fine-tuned for Vincent Van Gogh to distinguish its own watermarked synthesized images from those of the other 26 artists. For the human face task, we choose Aaron Carter vs. the other 3 celebrities. We ensure the class balance and test 250 images for each. As can be seen from

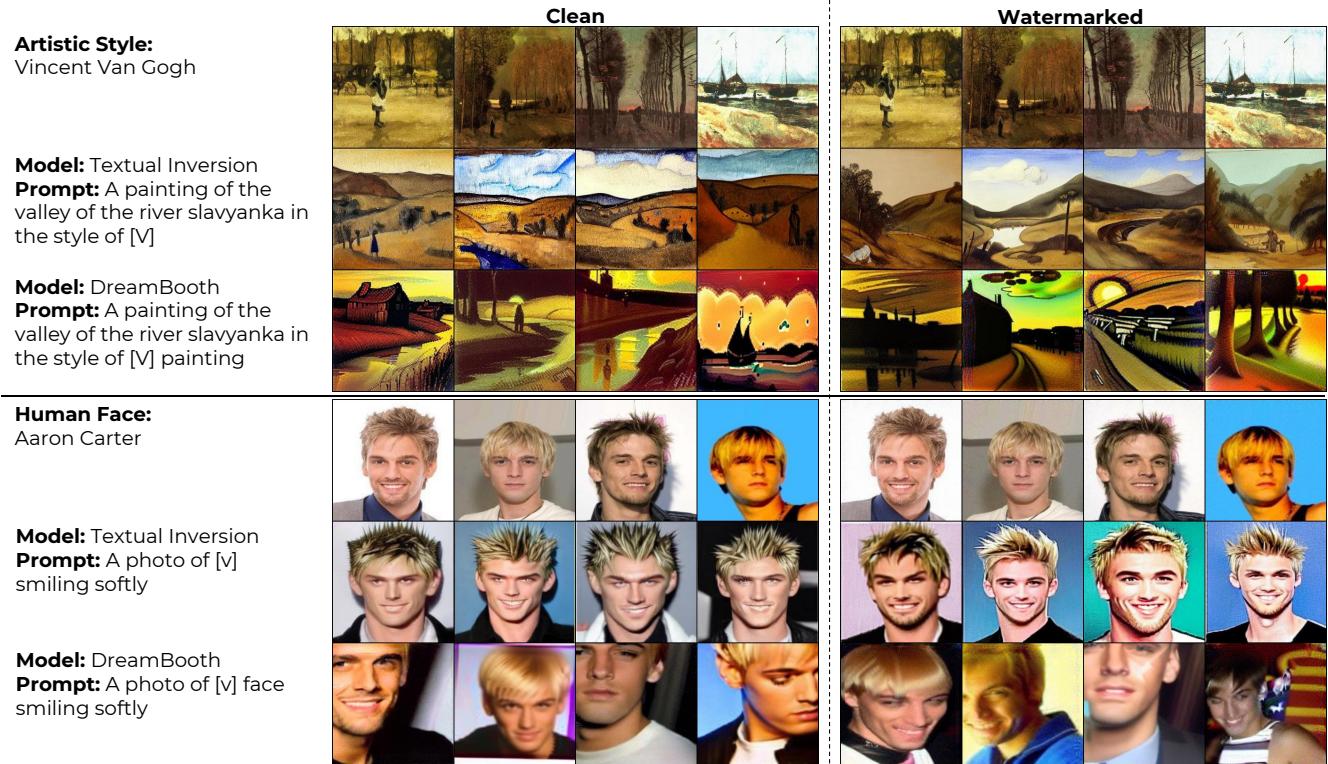


Figure 9: Images synthesized based on clean (left) vs. watermarked (right) for artistic style (top) and human face (bottom). Additional examples for other subjects with different prompts can be found in [Figure 19 of Appendix A](#).

Table 1: Watermark uniqueness measured by the classification accuracy (%) of a detector designed for a specific subject in distinguishing its own watermarked images from those generated for other subjects.

Artistic style		Human face	
TI	DB	TI	DB
83.2	85.4	76.5	78.2

Table 1, for the artistic style task, the classification accuracy is 84.3% averaged on two models, which implies the high uniqueness of our GenWatermark. The result for the human face task is lower, i.e., 77.4%. However, we should note that in this task, uniqueness is not practically meaningful since a person only has authority over their own face and so would not test an image that contains others’ faces.

5.3 Results on Image Synthesis Quality

In addition to the detection accuracy, a successful watermark should also have little impact on the model utility, i.e., the quality of the image synthesis. Here we evaluate how adding watermarks to the input images would affect the original synthesis quality with clean inputs. Specifically, we measure how the FID score calculated between synthesized images and input images would change after injecting the watermarks. As can be seen from [Figure 5](#), the FID score changes only by less than 1% on average, indicating that injecting watermarks indeed has little impact on the original synthesis

quality. In addition, we can see that Textual Inversion leads to consistently lower FID scores than DreamBooth, suggesting its better performance in subject-driven image synthesis.

Watermark Invisibility. [Figure 8](#) visualizes examples of clean vs. watermarked input images and their corresponding watermarks. We can observe that the watermarks in the watermarked images are barely visible to the human eye, especially for the artistic style task involving rich image textures. Moreover, as can be seen from [Figure 9](#), there are also no obvious patterns left in the output images that are synthesized based on the watermarked inputs, demonstrating that the watermarks barely affect the utility of image synthesizing quality.

Qualitative Analyses. We qualitatively compare the image examples synthesized using clean vs. watermarked in [Figure 9](#). In general, we can find that in both the artistic style and human face tasks, Textual Inversion and DreamBooth models can synthesize images that are similar to the original (clean) inputs. For the artistic style task, we choose artwork from Vincent Van Gogh, one of the most famous and influential painters in Western art history. As can be seen, the synthesized images based on the watermarked inputs share a very similar style with those based on the clean inputs. This is consistent with our FID results reported in [Figure 5](#).

For the human face task, we choose pictures from Aaron Carter, an American singer/rapper. Similar to the above observation, the synthesized images based on the watermarked inputs also well maintain the face and follow a similar style to those based on the clean inputs. An interesting finding here is that Textual Inversion and DreamBooth synthesize images

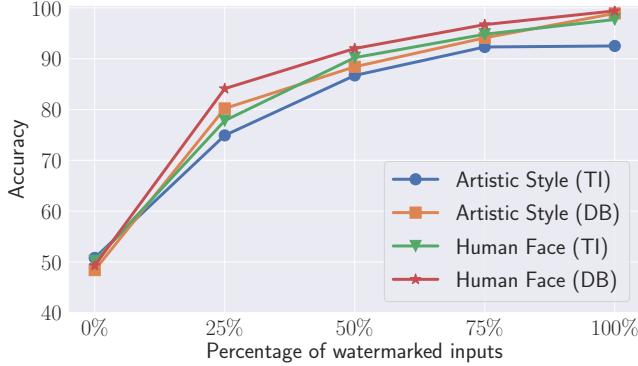


Figure 10: Watermark detection accuracy (%) when only partial input images are watermarked.

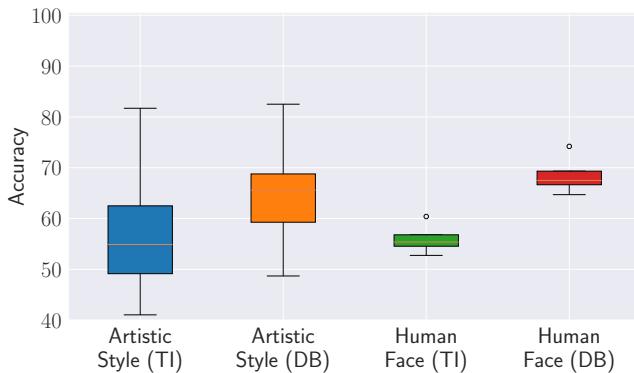


Figure 11: Watermark detection accuracy (%) without fine-tuning the detector.

with different preferences. Specifically, DreamBooth synthesizes more realistic faces in novel views, while Textual Inversion images look more like computer-generated graphics with a fixed front and comic-like view.

5.4 Ablation Studies

Partial Input Watermarking. In real-world scenarios, the subject owners may have already exposed their clean images before the watermark technique is developed, so it is possible for the malicious subject synthesizers to mix these images with watermarked images when training the subject-driven model. Here we evaluate the impact of such a realistic scenario on the detection performance of our GenWatermark. Figure 10 reports the performance of GenWatermark when only partial input images are watermarked, in Scenario 1.

We can see that the detection accuracy of GenWatermark gradually decreases as the subject synthesizers have access to more clean images. This is expected since it becomes harder for the synthesis models to learn the watermark information when it only appears in fewer input images. However, even when only 25% of the input images are watermarked, GenWatermark is still effective with a detection accuracy of about 80% on average. The accuracy increases to about 90% when only half of the input images are watermarked. This validates the robustness of our GenWatermark against the partially watermarked situation.

Without Detector Fine-Tuning. There is a detector fine-

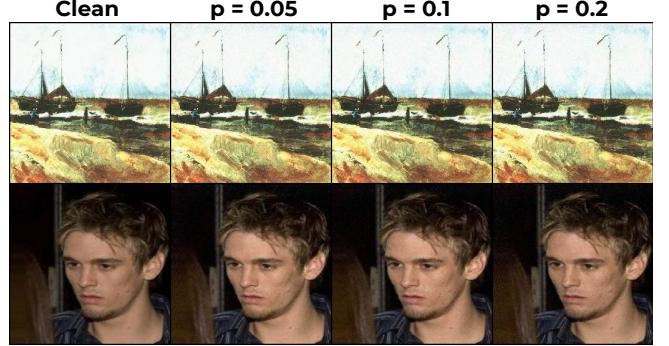


Figure 12: Clean and watermarked images under varied invisibility level p .

tuning phase during training GenWatermark. As discussed, this phase aims to improve the detection by “personalizing” the detector for each individual subject owner. Figure 11 shows the detection results when this fine-tuning phase is not applied during training GenWatermark. Note that only the results for Scenario 1 are reported since the cross-model and cross-prompt transfer scenarios only apply to the fine-tuning phase. We can see that the detection accuracy substantially decreases compared to those corresponding results in Figure 4, demonstrating the necessity of our detector fine-tuning.

Hyperparameters p and α . Here we analyze the impact of two main hyperparameters, i.e., the invisibility level p and balancing factor α , on the performance of our GenWatermark.

Figure 12 visualizes examples of the clean images and watermarked images under varied invisibility levels p . As expected, the noise becomes more visible as p is increased. For example, $p = 0.1$ leads to clearly more noticeable noise than that with $p = 0.05$. In addition, p also has little impact on the detection accuracy of GenWatermark, as shown in Figure 13 (top). Specifically, the detection accuracy increases when p increases from 0.05 to 0.1, especially in Scenario 3 and Scenario 4. This is expected since a more visible watermark is easier to be detected. However, the detection accuracy remains almost unchanged when we further increase p to 0.2. We find that it is due to the fact that when pre-training our GenWatermark, the detector has already converged without pushing the watermark to reach $p = 0.2$.

Figure 13 (bottom) shows the impact of the balancing factor α on the performance of GenWatermark. Here we report the results for DreamBooth as the target model. We can see that, in general, α has little impact on detection accuracy. The reason might be that the detector can achieve almost 100% accuracy in the validation dataset when fine-tuning. However, we find that increasing α accelerates the convergence of the generator. For both p and α , we find that Textual Inversion leads to very similar results, as shown in Figure 18 of Appendix A.

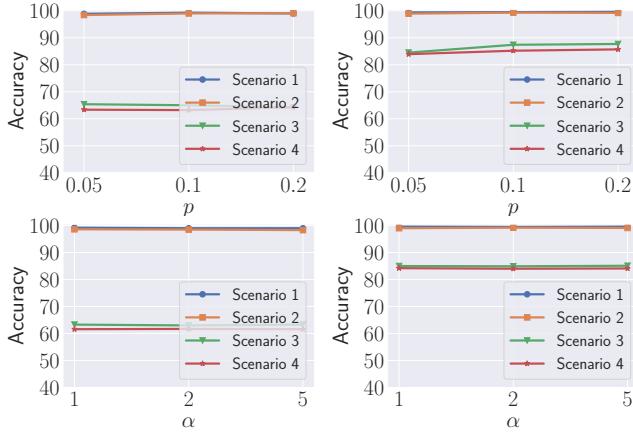


Figure 13: Watermark detection accuracy (%) under varied invisibility level p and balancing factor α for the artistic style (left) and human face (right) tasks. DreamBooth is used as the target model, and the results for Textual Inversion with a similar pattern can be found in [Figure 18 of Appendix A](#).

Table 2: Watermark detection accuracy (%) against watermark forgery.

Artistic style		Human face	
TI	DB	TI	DB
77.9	85.1	82.2	85.1

Table 3: Watermark detection accuracy (%) against watermark removal.

Image transformations	Artistic style		Human face	
	TI	DB	TI	DB
Gaussian noise (Input)	76.0	78.0	79.1	81.2
Gaussian noise (Output)	73.5	75.2	74.7	77.3
JPEG compression (Input)	71.5	72.4	75.5	77.4
JPEG compression (Output)	64.7	67.2	69.3	72.4

6 Countermeasures

In practice, the malicious subject synthesizers may apply countermeasures to reduce the protecting effects of our GenWatermark. In this section, we evaluate two potential countermeasures from different perspectives on the full datasets.

Watermark Forgery. Since our watermarks are basically high-frequency additive noises, as visualized in [Figure 8](#), a potential countermeasure could be that an adversary deliberately adds similar noises into synthesized images as forged watermarks such that our detector would misclassify them as watermarked images. We evaluate if GenWatermark is potentially robust to this countermeasure, and we specifically test with Gaussian noises. [Table 2](#) shows the watermark detection accuracy when Gaussian noises are added to clean inputs or outputs that are synthesized based on clean inputs. We find that on average, our watermark detector achieves a high accuracy of about 78%, indicating that it is hard to forge watermarks with random noises.

Watermark Removal. Existing work has shown that ad-

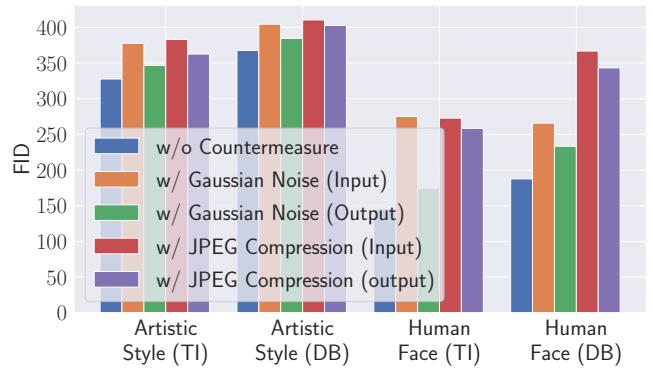


Figure 14: FID scores for synthesis from clean inputs, watermarked inputs, watermarked inputs with Gaussian noise, watermarked inputs with JPEG compression. The FID is calculated between the clean inputs and synthesized outputs.

versarial perturbations are vulnerable to image transformations in both the training [36] and testing [63] stages. Related studies in protecting against image synthesis have also considered input transformations as a typical countermeasure [33, 35, 55, 64]. Following them, we test two popular image transformations: Gaussian noise and JPEG compression. Specifically, the malicious subject synthesizers may apply such transformations to remove the watermarks from either input or output (synthesized) images.

[Figure 16](#) in [Appendix A](#) visualizes the images transformed by Gaussian noise with different standard deviations and JPEG compression with different quality factors. As can be seen, too severe transformations lead to obvious visual artifacts. Therefore, we choose Gaussian noise with a standard deviation of 0.0005 and JPEG compression with a factor of 20 for the following experiment.

[Table 3](#) shows the detection results against two image transformations and [Figure 14](#) shows their corresponding FID scores. In general, the average detection accuracy remains high, i.e., 74.1%. On the other hand, the FID results show that these transformations inevitably degrade the synthesis quality. Interestingly, transforming outputs leads to a consistently stronger countermeasure than transforming inputs as well as smaller quality degradation. When comparing two different transformations, we can observe that although JPEG compression is stronger against our detection than Gaussian noise, it leads to greater quality degradation.

Examples of synthesized images in [Figure 15](#) explain the above findings. Specifically, we find that when transforming the inputs, the visual artifacts are magnified in the synthesized outputs. In addition, the Gaussian noise results in high-frequency noisy patterns while the JPEG compression results in obvious blocking artifacts. In particular, for the human face task, Textual Inversion even fails to learn the accurate facial attributes of the subject. Besides, since DreamBooth tends to inherit fine-grained information from the input images (as discussed in [Section 5.2](#)), it causes severe blocking artifacts, and makes it even hard to recognize the actual image content.

Compared to the artistic style task, the human face task

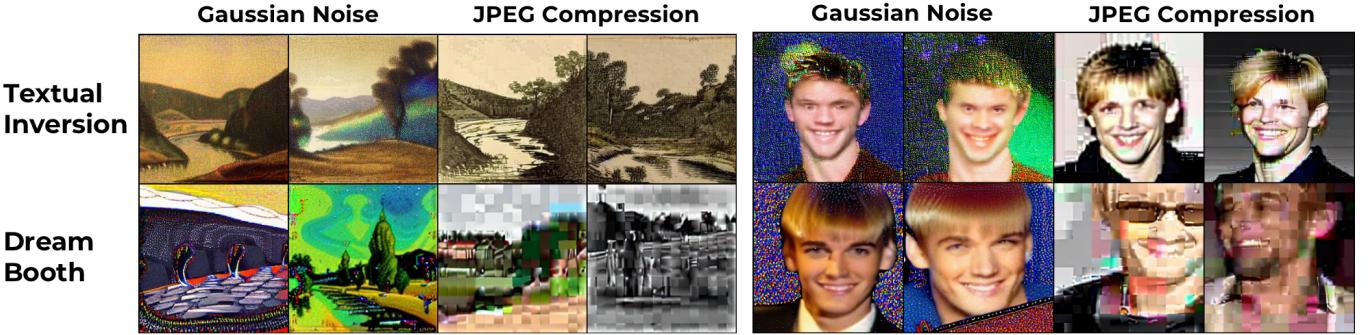


Figure 15: The effects of input transformations on the synthesized outputs. The same subjects as in Figure 9 are used.

yields greater quality degradation. This might be explained by our previous finding from Figure 7 that the subject-driven model trained on face images focuses more on the local attributes, and as a result, it is not robust to global transformations. In contrast, the style models are better at capturing holistic features, so they are more robust to global transformations. Overall, the above observations suggest that neither the watermark forgery nor watermark removal is effective in countering our GenWatermark.

7 Discussion and Limitations

Our GenWatermark is the first approach to watermarking images against unauthorized subject-driven image synthesis. Although we have conducted extensive experiments to demonstrate the effectiveness of GenWatermark in various scenarios, there is still future work to be done to further improve it. Here we identify two potential limitations of the current GenWatermark and discuss possible ways to address them.

Cross-Model Transferability. Cross-model transferability of GenWatermark is important in practical scenarios since the target model an adversary would choose is usually not known to the image owner who wants to watermark their images. This might not be a big problem at this moment given that there are only a few (publicly-available) recipes for training subject-driven synthesis models, e.g., Textual Inversion and DreamBooth adopted in our experiments. However, with the rapid development of subject-driven synthesis, we expect more effective models to be proposed.

Our experiments have shown that the cross-model transferability of GenWatermark is about 20% lower than the case with a known model. We hypothesize that this is due to the differences between Textual Inversion and DreamBooth in image synthesis. For example, in the artistic style task, DreamBooth tends to capture fine-grained concepts from the input images while Textual Inversion focuses more on the holistic style. In the human face task, DreamBooth tends to produce realistic images but Textual Inversion produces images like virtual characters.

Related work on adversarial examples has extensively studied the transferability of perturbations across different CNN models or other families of architectures [19, 41, 67], e.g., Vision Transformers [17]. A typical idea is to incor-

porate the specific model properties. Similarly, we can also improve the cross-model transferability of GenWatermark by incorporating model-specific properties, including our findings on their differences in synthesis, during fine-tuning the detector.

Watermark Uniqueness. Although our GenWatermark achieves substantial watermark uniqueness (e.g., 84.3% for the artistic style task), we acknowledge that it is still not on par with conventional methods, which are based on explicit watermark injection and reconstruction (but not applicable in our scenario). In order to improve the uniqueness, it would help if we fine-tune not only the detector but also the generator based on images from the specific subject under protection. In this way, the property of a specific subject can be embedded into not only the detector but also the generator and as a consequence its generated watermarks. Note that this would inevitably cost more computational resources.

8 Conclusion and Outlook

In this paper, we propose GenWatermark, the first approach to watermarking images against subject-driven image synthesis. Our GenWatermark is essentially different from previous protections because it aims to prevent only unauthorized use of image synthesis while maintaining its utility. GenWatermark consists of a generator and a detector. These two components are pre-trained based on large-scale data, and the detector is fine-tuned to further improve the personalized detection performance for each individual subject. We have demonstrated the robustness and invisibility of our GenWatermark in two typical image synthesis tasks, i.e., artistic style and human face, and on two representative target models, i.e., Textual Inversion and DreamBooth. In particular, we demonstrate the generalizability of GenWatermark in practical scenarios with unknown tasks, models, and text prompts, as well as with only partial data being watermarked. Moreover, the subject-driven model trained with our watermarked inputs yields similar synthesis quality to that trained on clean inputs.

Moving forward, there are two promising directions to explore. First, similar to all existing protections, our current GenWatermark may turn out to be vulnerable to a future countermeasure, e.g., the very recent development of the perturbation purification techniques [42, 61]. Therefore, future

work should evaluate the robustness of GenWatermark more comprehensively, which would in turn help design a better watermark system. Second, there is still room for improving GenWatermark in practical scenarios, such as cross-model and partial data watermarking. Further adapting GenWatermark to specific subjects based on the characteristics of their images would potentially help.

References

- [1] <https://midjourney.com/>. 1
- [2] Outguess. <http://www.outguess.org>. 3
- [3] What the heck is civitai? <https://civitai.com/content/guides/what-is-civitai>. 1
- [4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *CoRR abs/2211.01324*, 2022. 1
- [5] Shumeet Baluja. Hiding Images in Plain Sight: Deep Steganography. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2069–2079. NIPS, 2017. 2, 3
- [6] Mauro Barni. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2014. 3
- [7] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57. IEEE, 2017. 6
- [8] François Cayre, Caroline Fontaine, and Teddy Furon. Watermarking security: theory and practice. *IEEE Transactions on Signal Processing*, 2005. 3
- [9] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. *CoRR abs/2209.14491*, 2022. 1
- [10] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P. Dickerson, Gavin Taylor, and Tom Goldstein. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition. In *International Conference on Learning Representations (ICLR)*, 2021. 5, 6
- [11] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797. IEEE, 2018. 5
- [12] Ingemar Cox, Matthew Miller, and Jeffrey Bloom. *Digital Watermarking*. Springer, 2002. 3
- [13] Giannis Daras and Alexandros G. Dimakis. Discovering the Hidden Vocabulary of DALLE-2. *CoRR abs/2206.00169*, 2022. 1
- [14] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifuljah, Tanishq Abraham, Phúc Lê Khắc, Luke Melas, and Ritobrata Ghosh. Dall-e mini, 7 2021. 3
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In

- Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. ACL, 2019. 3
- [16] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 19822–19835. NeurIPS, 2021. 3
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. 12
- [18] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. *CoRR abs/2303.15435*, 2023. 3
- [19] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations? In *International Conference on Learning Representations (ICLR)*, 2022. 12
- [20] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *CoRR abs/2208.01618*, 2022. 1, 4, 6
- [21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680. NIPS, 2014. 5, 7
- [22] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*, 2015. 2
- [23] Jamie Hayes and George Danezis. Generating steganographic images via adversarial training. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1954–1963. NIPS, 2017. 3
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016. 7
- [25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 6626–6637. NIPS, 2017. 2, 7
- [26] Vojtech Holub, Jessica J. Fridrich, and Tomás Deneckere. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014. 3
- [27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR, 2021. 3
- [28] Jinyuan Jia and Neil Zhenqiang Gong. AttrGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning. In *USENIX Security Symposium (USENIX Security)*, pages 513–529. USENIX, 2018. 6
- [29] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410. IEEE, 2019. 5
- [30] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-Based Real Image Editing with Diffusion Models. *CoRR abs/2210.09276*, 2022. 1
- [31] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. *CoRR abs/1607.02533*, 2016. 6
- [32] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual Adversarial Robustness: Defense Against Unseen Threat Models. In *International Conference on Learning Representations (ICLR)*, 2021. 5, 6
- [33] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. *CoRR abs/2303.15433*, 2023. 2, 4, 6, 11
- [34] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-Driven Text-To-Image Synthesis via Adversarial Training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1274–1282. IEEE, 2019. 3
- [35] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. *CoRR abs/2302.04578*, 2023. 2, 4, 11
- [36] Zhuoran Liu, Zhengyu Zhao, and Martha A. Larson. Image Shortcut Squeezing: Countering Perturbative Availability Poisons with Compression. *CoRR abs/2301.13838*, 2023. 11

- [37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, 2015. 6
- [38] Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Generating Images from Captions with Attention. In *International Conference on Learning Representations (ICLR)*, 2016. 3
- [39] Chenlin Meng, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On Distillation of Guided Diffusion Models. *CoRR abs/2210.03142*, 2022. 1
- [40] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2022. 1
- [41] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On Improving Adversarial Transferability of Vision Transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 12
- [42] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning (ICML)*, pages 16805–16827. PMLR, 2022. 12
- [43] Tomás Pevný, Tomás Filler, and Patrick Bas. Using High-Dimensional Image Models to Perform Highly Undetectable Steganography. In *Information Hiding (IH)*, pages 161–177. Springer, 2010. 3
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3
- [45] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2016. 3
- [46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022. 1, 3
- [47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. JMLR, 2021. 3
- [48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2022. 1, 3
- [49] Jérôme Rony, Eric Granger, Marco Pedersoli, and Ismail Ben Ayed. Augmented Lagrangian Adversarial Attacks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7718–7727. IEEE, 2021. 5, 6
- [50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dream-Booth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 1, 4, 8
- [51] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J. Fleet. Adversarial Manipulation of Deep Representations. In *International Conference on Learning Representations (ICLR)*, 2016. 6
- [52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR abs/2205.11487*, 2022. 3
- [53] Babak Saleh and Ahmed M. Elgammal. Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. *CoRR abs/1505.00855*, 2015. 6
- [54] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the Cost of Malicious AI-Powered Image Editing. *CoRR abs/2302.06588*, 2023. 6
- [55] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models. *CoRR abs/2302.04222*, 2023. 2, 5, 6, 11
- [56] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. JMLR, 2015. 3
- [57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2014. 2
- [58] Matthew Tancik, Ben Mildenhall, and Ren Ng. StegaStamp: Invisible Hyperlinks in Physical Photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2114–2123. IEEE, 2020. 2, 3
- [59] Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16494–16504. IEEE, 2022. 3

- [60] Vedran Vukotic, Vivien Chappelier, and Teddy Furon. Are Deep Neural Networks good for blind image watermarking? In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 3
- [61] Quanlin Wu, Hang Ye, and Yuntian Gu. Guided Diffusion Model for Adversarial Purification from Random Noise. *CoRR abs/2206.10875*, 2022. 12
- [62] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324. IEEE, 2018. 3
- [63] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2018. 11
- [64] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14448–14457. IEEE, 2021. 3, 4, 11
- [65] Han Zhang, Tao Xu, and Hongsheng Li. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916. IEEE, 2017. 3
- [66] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595. IEEE, 2018. 5
- [67] Zhengyu Zhao, Hanwei Zhang, Renjue Li, Ronan Sicre, Laurent Amsaleg, and Michael Backes. Towards Good Practices in Evaluating Transfer Adversarial Attacks. *CoRR abs/2211.09565*, 2022. 12
- [68] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding Data With Deep Networks. In *European Conference on Computer Vision (ECCV)*, pages 682–697. Springer, 2018. 3
- [69] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810. IEEE, 2019. 3

Table 4: Artists and the corresponding genres.

Index	Name	Genre
1	Vincent Van Gogh	Realism
2	Claude Monet	Impressionism
3	Leonardo Da Vinci	High renaissance
4	Pablo Picasso	Cubism
5	Agnes Martin	Minimalism
6	Aubrey Beardsley	Art nouveau modern
7	Caravaggio	Baroque
8	Edvard Munch	Expressionism
9	Edward Hopper	New realism
10	Fra Angelico	Early renaissance
11	Francisco Goya	Romanticism
12	Francois Boucher	Rococo
13	Franz Kline	Action painting
14	Georges Braque	Analytical cubism
15	Georges Seurat	Pointillism
16	Giorgio Vasari	Mannerism late renaissance
17	Gustave Moreau	Symbolism
18	Henri Rousseau	Naive art primitivism
19	Henry Matisse	Fauvism
20	Hiroshige	Ukiyo e
21	Jackson Pollock	Abstract expressionism
22	Jan Van Eyck	Northern renaissance
23	Juan Gris	Synthetic cubism
24	Mark Rothko	Color field
25	Neil Welliver	Contemporary realism
26	Paul Cezanne	Post impression
27	Roy Lichtenstein	Pop art

A Additional Tables and Figures

In this section, we list additional tables and figures.

Table 5: All 30 prompts for artistic style and 30 prompts for human face in our experiments.

Artistic Style	Human Face
A painting of New York City in the style of [V]	A photo of [V] smiling softly
A painting of woman with animals in the style of [V]	A photo of [V] wearing a colorful men's suit
A painting of guitar and fruit dish in the style of [V]	A photo of [V] smoking on chair
A painting of the valley of the river slavyanka in the style of [V]	A photo of [V] in a fluffy pink robe
A painting of a girl with fruits in the style of [V]	A photo of [V] with straight dark hair
A painting of murillo boy with a dog in the style of [V]	A photo of [V] holding a change cup
A painting of a tree full of birds in the style of [V]	A photo of [V] with a cup of coffee
A painting of shaddow on Frankfork barren in the style of [V]	A photo of [V] in a crime scene
A painting of bathing woman in the style of [V]	A photo of [V] escaping out of a house
A painting of cattles in the spring in the style of [V]	A photo of [V] killing terrorista with a gunshot
A painting of pink house in the style of [V]	A photo of [V] with long curly hair
A painting of church on Lenox avenue in the style of [V]	A photo of [V] eating a bagel
A painting of study for the last supper in the style of [V]	A photo of [V] warning a jacket
A painting of view of the beach and sea from the mountains Crimea in the style of [V]	A photo of [V] with cat head
A painting of man in military costume in the style of [V]	A photo of [V] looking at the sky
A painting of cowboy in the organ mountains New Mexico in the style of [V]	A photo of [V] holding a tablet
A painting of manufactures on a blue background in the style of [V]	A photo of [V] sitting on couch
A painting of girl wit hdoll in the style of [V]	A photo of [V] reading newspaper
A painting of figures in a park Paris in the style of [V]	A photo of [V] holding a book
A painting of boy with blue cap in the style of [V]	A photo of [V] with cat
A painting of Mary and child in the style of [V]	A photo of [V] with a coconut
A painting of the young peasant and his wife in the style of [V]	A photo of [V] carrying water
A painting of young woman study for the clearing in the style of [V]	A photo of [V] in the jungle
A painting of statue of liberty in the style of [V]	A photo of [V] working in the hospital
A painting of naked woman in a landscape in the style of [V]	A photo of [V] in the rain
A painting of trees in the lake in the style of [V]	A photo of [V] laying in the green grass
A painting of sailboats in the bay in the style of [V]	A photo of [V] in the clinic stetic
A painting of guitar player in the style of [V]	A photo of [V] with beautiful flowers
A painting of mountains in the clouds in the style of [V]	A photo of [V] in the street
A painting of the reader in the style of [V]	A photo of [V] in middle age



Figure 16: The effects of input transformations on the clean inputs.

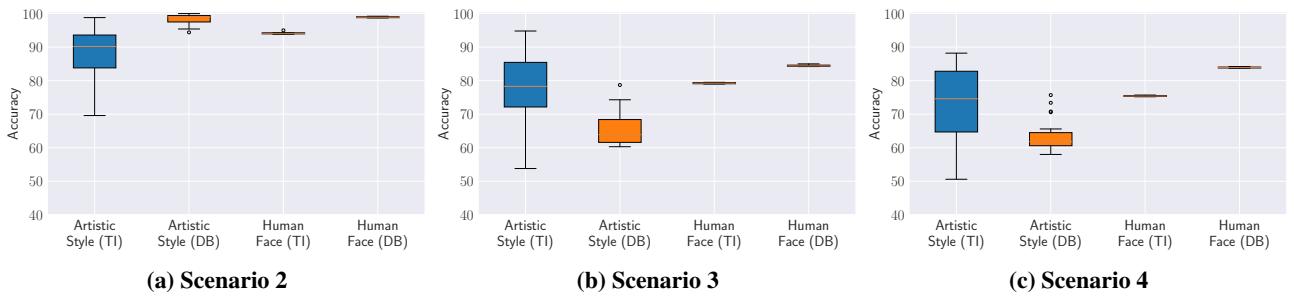


Figure 17: Detection accuracy of GenWatermark in different scenarios, supplement to Figure 7.

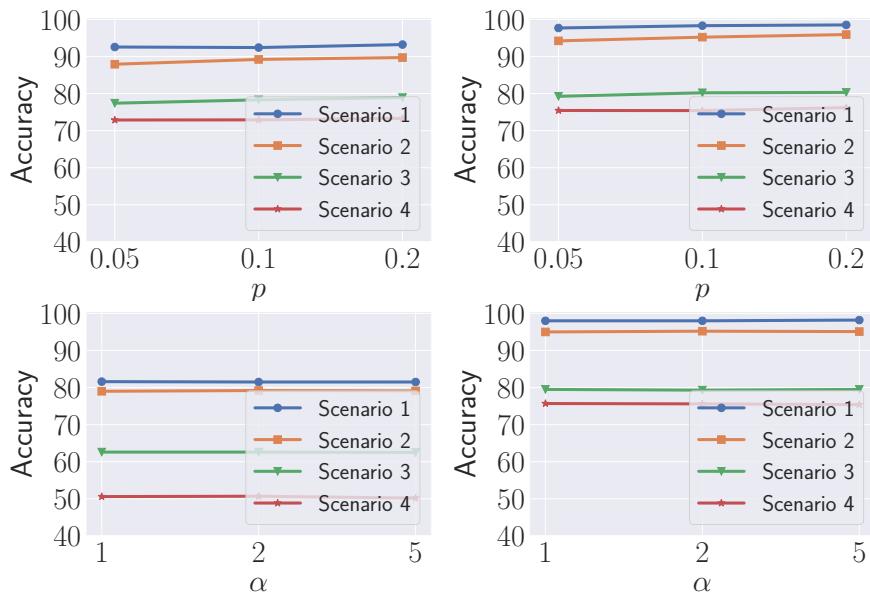


Figure 18: Watermark detection accuracy (%) under varied invisibility level p (top) and balancing factor α (bottom) for artistic style (left) and human face (right). Textual Inversion is used as the target model.

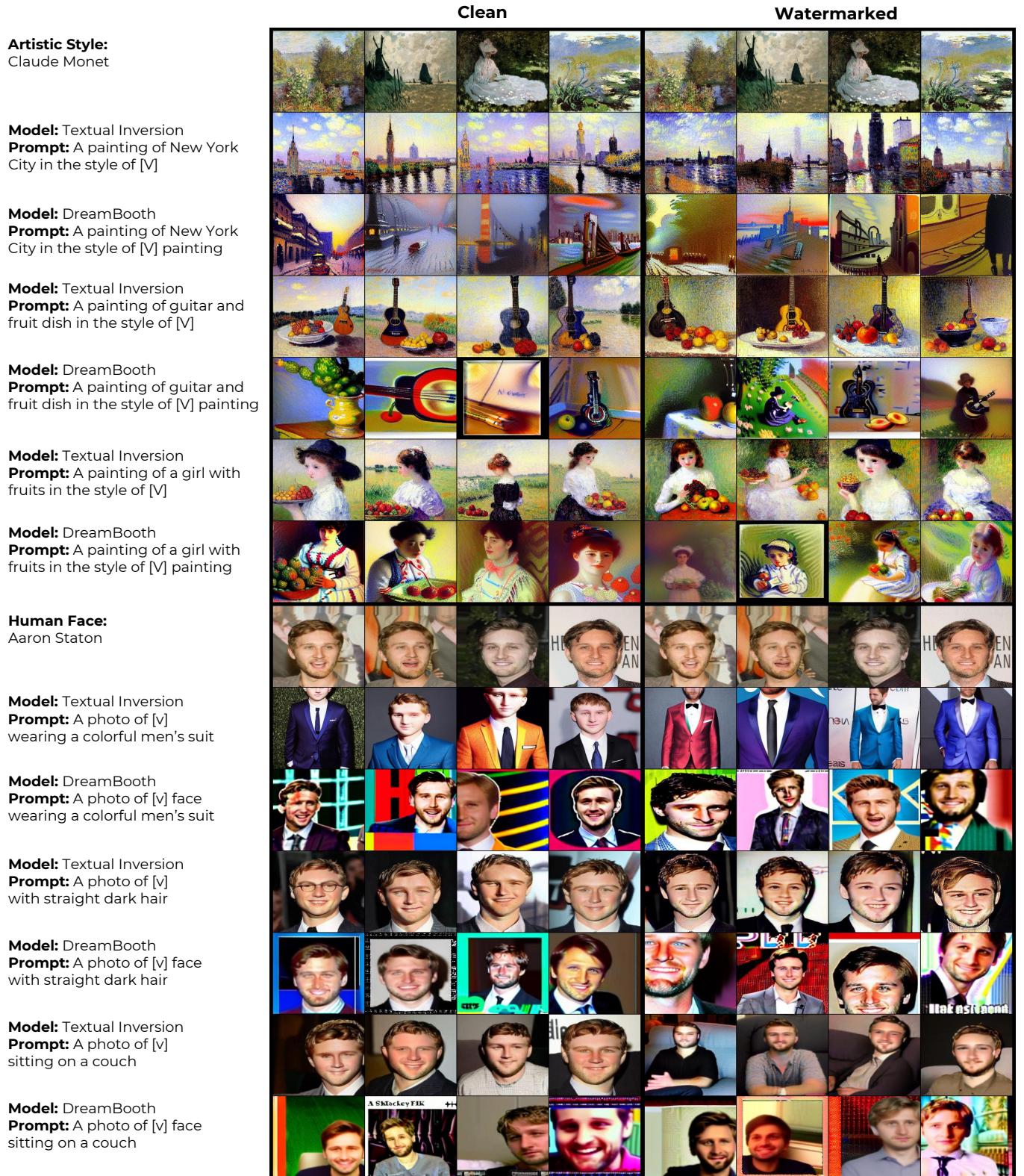


Figure 19: Images synthesized based on clean (left) vs. watermarked (right) for artistic style (top) and human face (bottom).