WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Image Diffusion Models

Changhoon Kim*[1] Kyle Min*[2] Maitreya Patel[1] Sheng Cheng[1] Yezhou Yang[1]
[1]Arizona State University [2]Intel Labs
{kch,maitreya.patel,scheng53,yz.yang}@asu.edu kyle.min@intel.com

Abstract

The rapid advancement of generative models, facilitating the creation of hyper-realistic images from textual descriptions, has concurrently escalated critical societal concerns such as misinformation. Although providing some mitigation, traditional fingerprinting mechanisms fall short in attributing responsibility for the malicious use of synthetic images. This paper introduces a novel approach to model fingerprinting that assigns responsibility for the generated images, thereby serving as a potential countermeasure to model misuse. Our method modifies generative models based on each user's unique digital fingerprint, imprinting a unique identifier onto the resultant content that can be traced back to the user. This approach, incorporating fine-tuning into Text-to-Image (T2I) tasks using the Stable Diffusion Model, demonstrates near-perfect attribution accuracy with a minimal impact on output quality. Through extensive evaluation, we show that our method outperforms baseline methods with an average improvement of 11% in handling image post-processes. Our method presents a promising and novel avenue for accountable model distribution and responsible use. Our code is available in https://github.com/kylemin/WOUAF.

1. Introduction

Recent advancements in generative models have propelled their proficiency, expanding their repertoire to include not just the generation of photorealistic images [4, 14] but also the synthesis of images from textual prompts [21, 24, 26, 28]. These significant strides have equipped individuals with the capacity to leverage these models to create hyper-realistic images that correspond seamlessly with given textual instructions.

Nonetheless, the escalating prominence of generative models instigates pressing societal apprehensions. A case in point is Deepfake, intentionally crafted to disseminate misinformation, fostering a climate of fake news and political disarray [2, 17, 23]. The gravity of these concerns necessitates calls for governmental intervention to regulate the indiscriminate application of generative models[1].

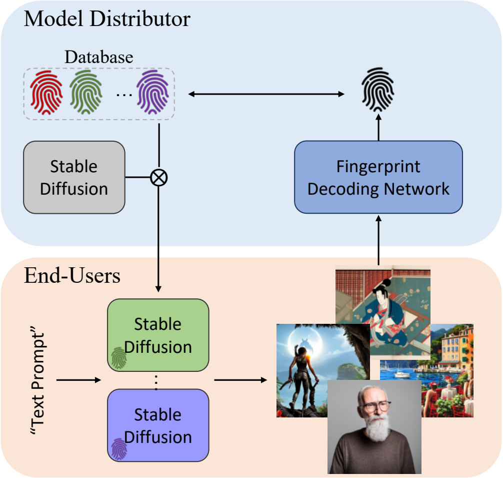


Figure 1. Illustration of user attribution based on our method. Please refer to the main text for detailed descriptions.

A feasible method to counteract malicious use involves assigning accountability for generated images. One approach to achieve this is by integrating independent fingerprinting modules that can embed user-specific information on top of image generation. The open-source Text-to-Image (T2I) project Stable Diffusion (SD) [26] currently employs this technique using discrete wavelet transform or RivaGAN [34]. However, in the open-source setting, bypassing the fingerprinting module is straightforward and can be achieved by commeting just a single line in the source code [5].

Is it feasible to achieve user attribution without an independent fingerprinting module? In response, we propose a distributor-oriented methodology named **WOUAF**, standing for **W**eight m**O**dulation for **U**ser **A**ttribution and

*These authors contributed equally to this work.

[1] President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. The White House

Fingerprinting. In practical terms, when a model inventor open-sources their work to a model distributor such as Huggingface, the distributor could utilize our proposed method to create a generic version. Upon receiving a download request from an end-user, the distributor can adjust the model weights using our technique and deploy a fingerprinted version to the user, simultaneously registering the user's fingerprint into their database. In the event of a model's malicious exploitation, the distributor can decode the fingerprint from the misused image and cross-reference it with their database to identify the responsible user. Consequently, this provides the distributor with an actionable method to counteract malicious uses of the model (see Fig. 1 for a comprehensive framework of our methodology).

Our methodology, designed for T2I tasks, is integrated into the Stable Diffusion (SD) framework without necessitating any structural changes to the model. This design choice effectively prevents end-users from bypassing the fingerprinting process. Consistent with prior research [5, 16, 22, 32, 33], our primary goal is to maintain high attribution accuracy while ensuring minimal impact on output quality, as elaborated in Sec. 3. Our rigorous evaluations of this method concentrate initially on assessing both attribution accuracy and image quality. We have found that our approach attains nearly flawless attribution accuracy with only a slight influence on image quality. Moreover, we evaluate the robustness of our method in various scenarios involving post-processing manipulations that images might undergo. Our method outperforms baseline methods in these robustness tests, showing an average improvement of 11% in handling such manipulations (refer to Sec. 4 for further details).

There are four main contributions: (1) We introduce **WOUAF**, a distinctive distributor-centered fine-tuning methodology. This approach embeds fingerprints within the model in such a way that end-users cannot easily circumvent or remove them. (2) Our method successfully achieves high attribution accuracy, while maintaining the quality of the output images. (3) Our approach exhibits marked resilience against a diverse array of image post-processes, a vital attribute for practical applications. (4) We conduct thorough assessments to balance attribution accuracy with manipulations to intentionally remove fingerprints, including strategies like image compression via auto-encoders and obfuscation through model fine-tuning.

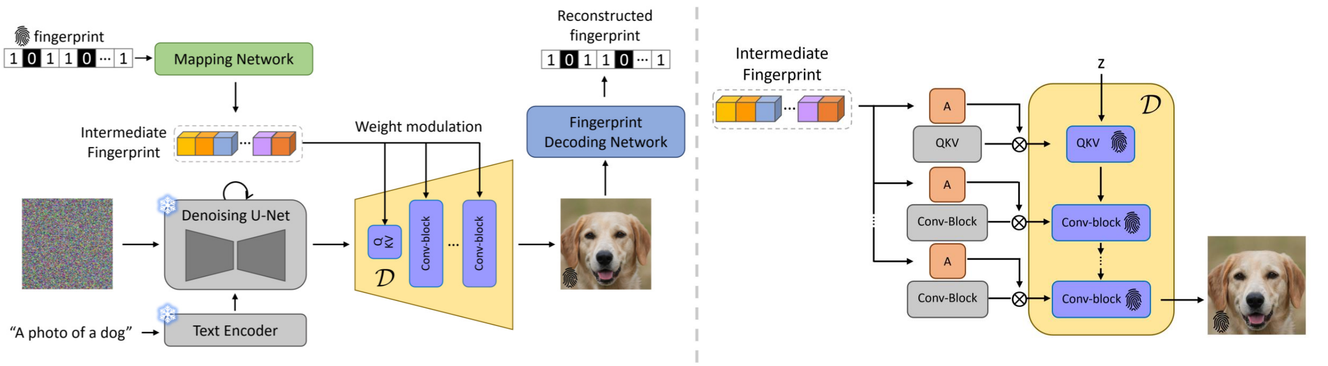
2. Related Work

In this section, we discuss related works of model fingerprinting in generative models. More related works are available in the appendix.

Inventor-oriented Model Fingerprinting. Yu et al. [33] leveraged a pre-trained deep steganography model to embed fingerprints into the training set for fingerprinted GANs. However, this approach suffers from limited scalability, as it necessitates training a GAN from scratch for each distinct fingerprint. To address this issue, Yu et al. [32] introduced a weight modulation method [12] that directly embeds a user's fingerprint into the generator's weights. Despite these advancements, current methods are predominantly tailored for GAN-based models and typically require training from scratch. This raises important questions regarding their suitability for diffusion-based models, which have a different structural makeup compared to GANs, and the feasibility of avoiding the requirement for training from scratch. The adoption of fine-tuning as a method for embedding fingerprints presents a promising solution. It facilitates the incorporation of fingerprints into pre-trained diffusion models, eliminating the necessity for comprehensive retraining from the ground up [5, 36]. This approach significantly streamlines the process, allowing model inventors to concentrate on core model development without the complexities of embedding fingerprints during training.

Distributor-oriented Model Fingerprinting. Kim et al. [16] proposed a technique for achieving user attribution by explicitly incorporating user-specific fingerprints into the generator's output. While this simplified attribution method allowed for the derivation of sufficient fingerprint conditions, it necessitates a trade-off between the quality of the generated output and attribution accuracy, which is further exacerbated when image post-processes are taken into account. To tackle this issue, an approach has been proposed that utilizes subtle semantic variations along latent dimensions as fingerprints, generated by perturbations of eigenvectors in the latent distribution [22]. This method demonstrates an improved balance between generation quality and attribution accuracy. However, its applicability is restricted to unconditional image generation, as eigenvectors are computed by sampling the learned latent representation. In the context of conditional image generation, estimating eigenvectors of latent representation becomes challenging due to the vast space of conditions, such as those found in text conditions.

Recent Advances in Fingerprinting for Text-to-Image Diffusion Models. Recent studies [5, 31, 36] have scrutinized fingerprinting techniques in the Stable Diffusion model [26], uncovering vulnerabilities in existing methods [34] that facilitate easy circumvention [5] or robust post-hoc fingerprinting [31]. Fernandez et al. [5] achieved near-perfect attribution accuracy by fine-tuning user-specific models to align with steganography module [37], demonstrating a viable alternative to conventional post-hoc fingerprinting modules [34]. However, this approach scales linearly in computational demand with the number of users since it necessitates fixed-time fine-tuning



(a) The overall pipeline.

(b) Weight modulation.

Figure 2. Depiction of our method's pipeline and weight modulation: (a) The model fingerprinting procedure encompasses encoding via the mapping network and weight modulation, along with decoding through the fingerprint decoding network. (b) Weight modulation of the decoding network $\mathcal{D}$ to incorporate the fingerprint.

for each individual. In contrast, our method requires only a one-time training followed by a negligible forward pass time to generate user-specific models. Furthermore, our approach shows superior robustness against common image post-processing techniques compared to that of Stable Signature [5] (refer to Section Sec. 4.6 for details).

Another notable contribution is by Wen et al. [31], who introduced an alternative fingerprinting method for the Stable Diffusion model [26]. Their method, similar to post-hoc style fingerprinting [34], depends on user-driven embedding, which allows end-users the option to exclude the fingerprint. Moreover, it is confined to the DDIM scheduler [30]. Our method, in contrast, is adaptable to both the DDIM [30] and Euler schedulers [15], underscoring its versatility and wider applicability (refer to the Appendix).

3. Methods

This section outlines our approach, beginning with an overview of the Text-to-Image (T2I) diffusion model with a focus on the Stable Diffusion (SD) model [26] detailed in Sec. 3.1. We then introduce our key component, the user-specific weight modulation, in Sec. 3.2. The section concludes with a detailed explanation of our training objectives and methods, outlined in Sec. 3.3.

3.1. Preliminaries

Our approach utilizes the Stable Diffusion (SD) model, which functions within the latent space framework of an autoencoder. SD comprises two main elements: Firstly, an autoencoder is pre-trained on an extensive dataset of images. Its encoder, $\mathcal{E}(\cdot) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_z}$, converts an image $x \sim p_{data}$ into a latent representation $z = \mathcal{E}(x)$. The decoder, $\mathcal{D}(\cdot) : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_x}$, then reconstructs the original image from this latent representation, resulting in $\hat{x} = \mathcal{D}(z)$. The secondary element is a diffusion model, based on the U-Net architecture [27], represented as ϵ_θ. This model is adept at generating latent representations and can be conditioned using pre-trained text embeddings

3.2. User-specific Weight Modulation

Our method is fundamentally based on integrating fingerprints into the parameters of the SD through weight modulation [12, 32].

The overall pipeline of our method is illustrated in Fig. 2(a). A user-specific fingerprint is drawn from a Bernoulli distribution with a probability of 0.5, represented as $\phi \in \Phi := \text{Bernoulli}(0.5)^{d_\phi}$, where d_ϕ signifies the fingerprint length in bits. We employ a mapping network $\mathcal{M}(\cdot) : \mathbb{R}^{d_\phi} \rightarrow \mathbb{R}^{d_M}$ to convert the sampled fingerprint ϕ into an intermediate fingerprint representation within the d_M dimension. For modulating each layer in the SD component, we introduce an affine transformation layer, $\mathcal{A}_l(\cdot) : \mathbb{R}^{d_M} \rightarrow \mathbb{R}^{d_j}$, for all layers l. As depicted in Fig. 2(b), this transformation matches the dimensions between d_M and the j-th channel in weight $W \in \mathbb{R}^{i,j,k}$, where i, j, k denote input, output, and kernel dimensions, respectively. The weight modulation for the l-th layer is defined as:

$$W^\phi_{i,j,k} = u_j * W_{i,j,k}, \tag{1}$$

where W and W^ϕ denote the pre-trained and fingerprinted weights respectively, $u_j = \mathcal{A}_l(\mathcal{M}(\phi))$ is the scale of the fingerprint representation corresponding to the jth output channel.

We incorporate fingerprints into the SD by applying weight modulation exclusively to the weights in the decoder $\mathcal{D}$. The rationale for not applying modulation to both the diffusion model ϵ_θ and decoder $\mathcal{D}$, an approach that mirrors GAN-based models [32], is explained in Sec. 4.5.

3.3. Training Objectives

Our training architecture comprises two primary objectives. The initial objective is to decode fingerprints from the pro-

Table 1. Evaluation of Attribution Accuracy and Image Generation Quality. We conducted validation using the MS-COCO [19] test set and the LAION-Aesthetics [29] dataset, which were excluded from our training phase. Symbols ↑ and ↓ denote preferred higher and lower values, respectively.

Model	Fingerprinting Time (↓)	MS-COCO Attribution Acc (↑)	MS-COCO CLIP-score (↑)	MS-COCO FID (↓)	LAION Attribution Acc (↑)	LAION CLIP-score (↑)	LAION FID (↓)
Original SD [26]	-	-	0.73	24.48	-	0.50	19.67
DAG [16]	8.4 hr	0.70	0.73	26.54	0.71	0.49	23.13
Stable Signature [5]	$<$ 1 min	0.99	0.73	24.55	0.98	0.50	20.02
WOUAF-conv	$<$ 1 sec	0.99	0.73	24.43	0.98	0.51	20.46
WOUAF-all	$<$ **1 sec**	**0.99**	**0.73**	**24.42**	**0.99**	**0.51**	**19.91**

vided images. We train a fingerprint decoding network $\mathcal{F}(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_\phi}$, which is instantiated by ResNet-50 [6], as follows:

$$L_\phi = \mathbb{E}_{z=\mathcal{E}(x),\phi\sim\Phi}\sum_{i=1}^{d_\phi}[\phi_i \log\sigma(\mathcal{F}(\mathcal{D}(\phi,z))_i + (1-\phi_i)\log(1-\sigma(\mathcal{F}(\mathcal{D}(\phi,z)))_i], \quad (2)$$

where $\sigma(\cdot)$ refers to the sigmoid activation function, constraining the output of $\mathcal{F}$ to the range $[0, 1]$. Thus, this loss function effectively combines binary cross-entropy for all bits of the fingerprint. During training time, fingerprint ϕ is sampled from Bernoulli distribution. However, after training, the model distributor initially samples a user-specific fingerprint ϕ_α and subsequently modulates the decoder $\mathcal{D}$ using ϕ_α. The user will receive the fingerprinted decoder $\mathcal{D}(\phi_\alpha, \cdot)$, which solely permits latent input.

The secondary objective endeavors to regularize the quality of outputs. Ideally, this regularization inhibits the decoder $\mathcal{D}$ from compromising image quality while minimizing L_ϕ in Eq. (2):

$$L_{\text{quality}} = \mathbb{E}_{z=\mathcal{E}(x),\phi\sim\Phi}\left[\ell(x, \mathcal{D}(\phi, z))\right], \quad (3)$$

ℓ represents the distance metric between original images and fingerprinted images. For practical applications, we utilize perceptual distance [35] to gauge the perceptual difference between x and $\mathcal{D}(\phi, z)$.

The final objective function can be formulated as:

$$\min_{\mathcal{A},\mathcal{M},\mathcal{D},\mathcal{F}} \lambda_1 L_\phi + \lambda_2 L_{\text{quality}}, \quad (4)$$

where both λ_1 and λ_2 are set to 1.0. Fundamentally, the loss function aspires to reconstruct fingerprints while maintaining the quality of the generated outputs. To assess the efficacy of our proposed method, we employ attribution accuracy and image quality metrics (Refer to Sec. 4.1 for details).

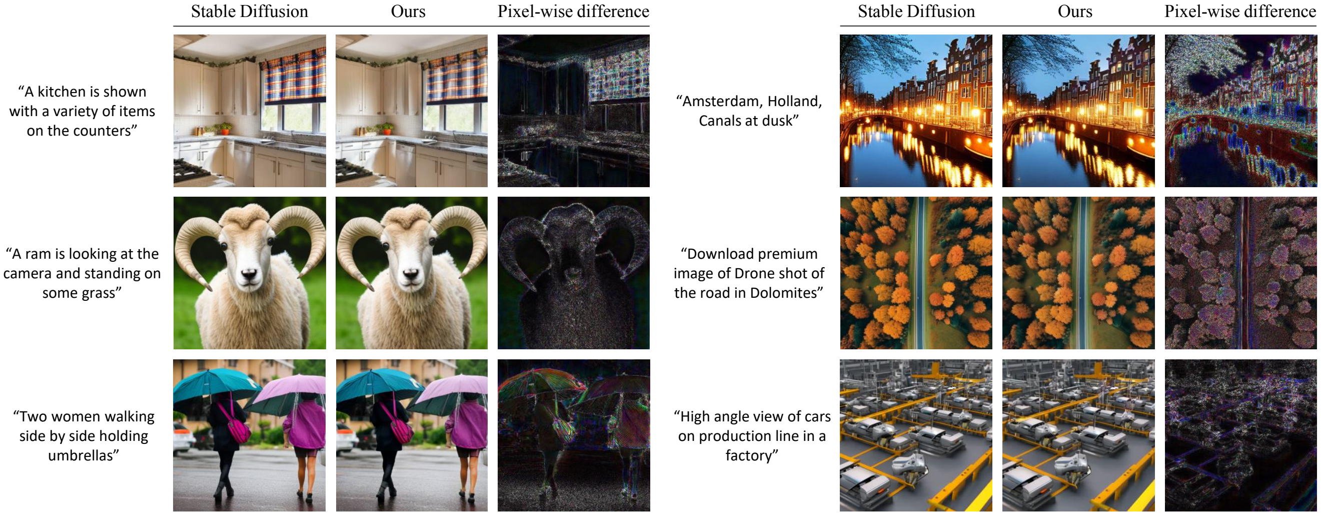
4. Experiments

4.1. Experiment Settings

Datasets. Our approach is fine-tuned on the MS-COCO [19] dataset, adopting the Karpathy split. For methodological evaluation, we harness the test set from MS-COCO and randomly sample from the LAION-aesthetics [29] dataset. For T2I image generation, we adopt the Euler scheduler [15] with timestep $T = 20$, and the classifier-free guidance scale [9] is set to 7.5 unless otherwise specified. Evaluation for DDIM scheduler [30] and various image generation hyperparameters are available in the Appendix.

Experimental Setting. We implement the weight modulation following the design specified in the source code of StyleGAN2-ADA [13]. Our mapping network M is designed with a series of fully connected layers, wherein all experiments are conducted using a two-layer configuration. To train robust models against image post-processing transformations, differentiable post-processes are necessary. To this end, we incorporate the Kornia library [25]. For Stable Signature [5], we utilize the official code provided by the authors. We note that its post-processing transformations are replaced with our version for fair comparison. Appendix includes details on mapping network dimensions, training parameters, and optimizer.

Evaluations. User attribution accuracy is gauged by the formula: $\frac{1}{d_\phi}\sum_{i=1}^{d_\phi}\mathbb{1}(\phi_i = \hat{\phi}_i)$, where ϕ is the true fingerprint and $\hat{\phi} = \mathbb{1}\left[\sigma(\mathcal{F}(x_\phi)) > 0.5\right]$ is the estimated fingerprint from image x_ϕ. Unless otherwise stated, d_ϕ is set to 32 in our experiments (Refer to Sec. 4.2 for additional information). We further employ a statistical test [32, 33] to evaluate matching bits between $\hat{\phi}$ and ϕ. The null hypothesis H_0 suggests that the number of matching bits arises by chance. The test uses a binomial distribution, with a p-value derived as: $P(X \geq k|H_0) = \sum_{i=k}^{d_\phi}\binom{d_\phi}{i}0.5^{d_\phi}$. A p-value below 0.05 leads to the rejection of H_0, with $1 - p$ serving as an indicator of verification confidence. Lastly, to


Stable Diffusion
Ours
Pixel-wise difference
"A kitchen is shown with a variety of items on the counters"
"A ram is looking at the camera and standing on some grass"
"Two women walking side by side holding umbrellas"
Stable Diffusion
Ours
Pixel-wise difference
"Amsterdam, Holland, Canals at dusk"
"Download premium image of Drone shot of the road in Dolomites"
"High angle view of cars on production line in a factory"

(a) MS COCO

(b) LAION Aesthetics

Figure 3. Qualitative comparison of the original and fingerprinted Stable Diffusion models on MS-COCO [19] and LAION aesthetics [29] (Pixel-wise differences× 5: they are multiplied by a factor of 5 for better view). We can observe that our method maintains high image quality.

validate the quality of our method, we assess image quality using the Fréchet Inception Distance (FID) [8] and employ the Clip-score [7] to determine the alignment between text and generated images. Additional experimental details can be found in the Appendix.

Models. For evaluating our methodology, we benchmark against two established baseline methods: DAG [16] and Stable Signature [5]. Both these methods, conceptualized from the model distributor's standpoint, incorporate fine-tuning for model fingerprinting. To ensure a fair comparison, we retrain the Stable Signature method within our training settings by replacing its post-processing scheme. Additionally, we evaluate our method against three distinct variants based on the specific layers chosen for weight modulation implementation. The first variant, WOUAF-conv, applies modulation to only the convolutional layers in $\mathcal{D}$. In contrast, WOUAF-all extends this approach across all layers of $\mathcal{D}$, covering both self-attention and convolution layers. The final variant implements weight modulation in both the diffusion model ϵ_θ and the decoder $\mathcal{D}$, mirroring the approach used in GAN-based methods [32]. Further details on why this variant is not used in our experiments are discussed in Sec. 4.5.

4.2. Fingerprint Capacity

The capacity of our method depends on the maximum number of unique user-specific fingerprints it can support without significant crosstalk. This capacity is primarily influenced by the fingerprint dimension (d_ϕ). Selecting an optimal d_ϕ presents a challenge: while a larger d_ϕ can accommodate more users, it also complicates effective fingerprint decoding [18].

Table 2. Experiments of attribution accuracy across various fingerprint dimensions (d_ϕ).

	Attribution Accuracy			
Fingerprint Dims.	16	32	64	128
WOUAF-conv	0.99	0.99	0.98	0.94
WOUAF-all	0.99	0.99	0.99	0.97

To investigate this trade-off, we conduct an analysis with varying fingerprint dimensions, specifically d_ϕ values of 16, 32, 64, and 128. Tab. 2 presents the user attribution accuracy for each d_ϕ value. As shown in Tab. 2, attribution accuracy tends to decrease monotonically as d_ϕ increases. Importantly, both our variant models achieve a near-perfect attribution accuracy of 0.99 for d_ϕ values of 16, 32, and 64. However, for $d_\phi = 128$, WOUAF-all variant outperforms the WOUAF-conv variant. For a balanced comparison with existing methods, we choose $d_\phi = 32$, which notably can support a substantial user base exceeding 4 billion $\approx 2^{32}$.

4.3. Attribution Accuracy and Image Quality

We conduct a comprehensive evaluation of WOUAF, focusing on attribution accuracy and image quality. The assessment involves the MS-COCO [19] test set and the LAION-Aesthetics [29] dataset, which are excluded from the training phase. The results, detailed in Tab. 1, showcase the efficacy of our method.

Our variants, namely WOUAF-conv and WOUAF-all, demonstrate superior performance in attribution accuracy over DAG [16], indicating their proficiency in accurately decoding embedded fingerprints from the generated images. These variants also show competitive results when compared to Stable Signature [5], reinforcing our methodology's robustness. Notably, we achieve this high level of accuracy without significantly compromising image quality. Both FID scores and Clip-scores showed minimal variation from the baseline SD model, indicating that our approach has a negligible impact on image output quality. This is further corroborated by qualitative examples in Fig. 3, which highlight WOUAF's ability to reliably incorporate fingerprinting without degrading image generation quality. For additional insights, uncurated image collections are provided in the Appendix.

Given the growing importance of T2I models, computation time for fingerprinting emerges as a key metric. Our method stands out in computational efficiency. It contrasts with approaches like Stable Signature that need fine-tuning for each new fingerprint. Our method requires just a single forward pass, markedly reducing computational overhead.

4.4. Attribution Analysis for Diverse Image Sources

Investigating the attribution of generated images to responsible users, we explore the potential for images from non-fingerprinted or varied sources to bypass our system. Our analysis aims to determine if decoded fingerprints from such images match any entries in the model distributor's database. A mismatch indicates the image's external origin, absolving users in the database.

We adopt the experimental setup from [32], compiling a dataset with different image types: authentic images from the MS-COCO test set [19], non-fingerprinted images from Stable Diffusion [26], and synthesized images from ProGAN [10], StyleGAN [11], and StyleGAN2 [12], with each category containing 1,000 samples. Given our extensive user database of 1 million entries, we set a threshold at $32 * 0.95 \approx 30$ bits, aligning with our 0.99 attribution accuracy as shown in Tab. 1.

Our rigorous experiments revealed that, irrespective of the source, no images were incorrectly attributed as possessing a fingerprint from our 1 million fingerprint database. This reinforces the reliability of our attribution approach as detailed in Sec. 4.3 demonstrating the robustness of our system against diverse image sources.

4.5. Benefits of Finetuning only Decoder

When developing our last variant that incorporates weight modulation into both the diffusion model ϵ_θ and the decoder $\mathcal{D}$, we note that the resultant pipeline demonstrates similarities with the GAN-based method [32]. A direct comparison between ours and the GAN-based methods may not be entirely straightforward, given the fundamental differences in their training methodologies. This is because the GAN-based methods entail training from scratch, whereas our proposed approach leans towards fine-tuning. Nevertheless, both methodologies share a common mechanism: they aim to modulate the weights of the layers instrumental in learning the latent space. The shared characteristic underscores the fundamental objective of optimizing the balance between attribution accuracy and generation quality.

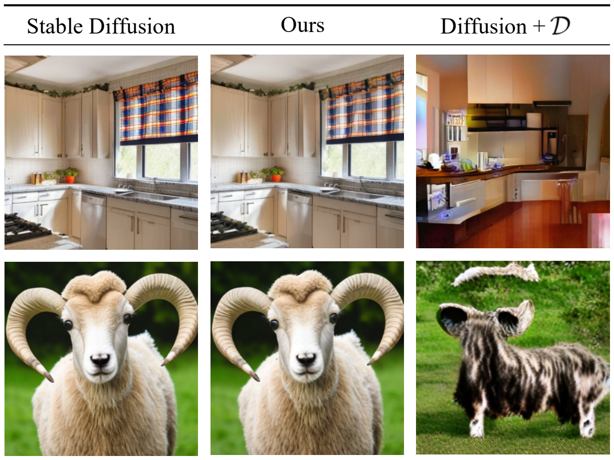


Figure 4. Comparative analysis of weight modulation on decoder $\mathcal{D}$ and diffusion model ϵ_θ with decoder $\mathcal{D}$. Modulating the diffusion model negatively affects image quality.

However, our empirical observations suggest that this variant does not consistently achieve commendable performance as an attribution model. Specifically, it appears that this variant can only optimize either attribution accuracy or generation quality, but not both simultaneously. In our tests, the highest attribution accuracy reached by this variant is 89%, with a Clip-score of 0.68 and FID of 63.48 (detailed in Fig. 4). The inherent trade-off observed here further reinforces the challenge of balancing these two critical parameters in the context of model fingerprinting techniques.

4.6. Robust User Attribution against Image Post-processes

This section evaluates the robustness of our method in scenarios where generated images undergo post-processing. These processes could potentially alter the embedded fingerprint within the images.

Consistent with methodologies outlined in previous research [5, 16, 22, 32, 33], we examine our model's resilience to various image post-processing operations. We simulate the effect of post-processing at random intensities before inputting data into the fingerprint decoding network, $\mathcal{F}$:

$$L_{\text{robust}} = \mathbb{E}_{z=\mathcal{E}(x),\phi\sim\Phi}\sum_{i=1}^{d_\phi}[\phi_i \log\sigma(\mathcal{F}(T(\mathcal{D}(\phi,z)))_i + (1-\phi_i)\log(1-\sigma(\mathcal{F}(T(\mathcal{D}(\phi,z)))_i], \qquad (5)$$

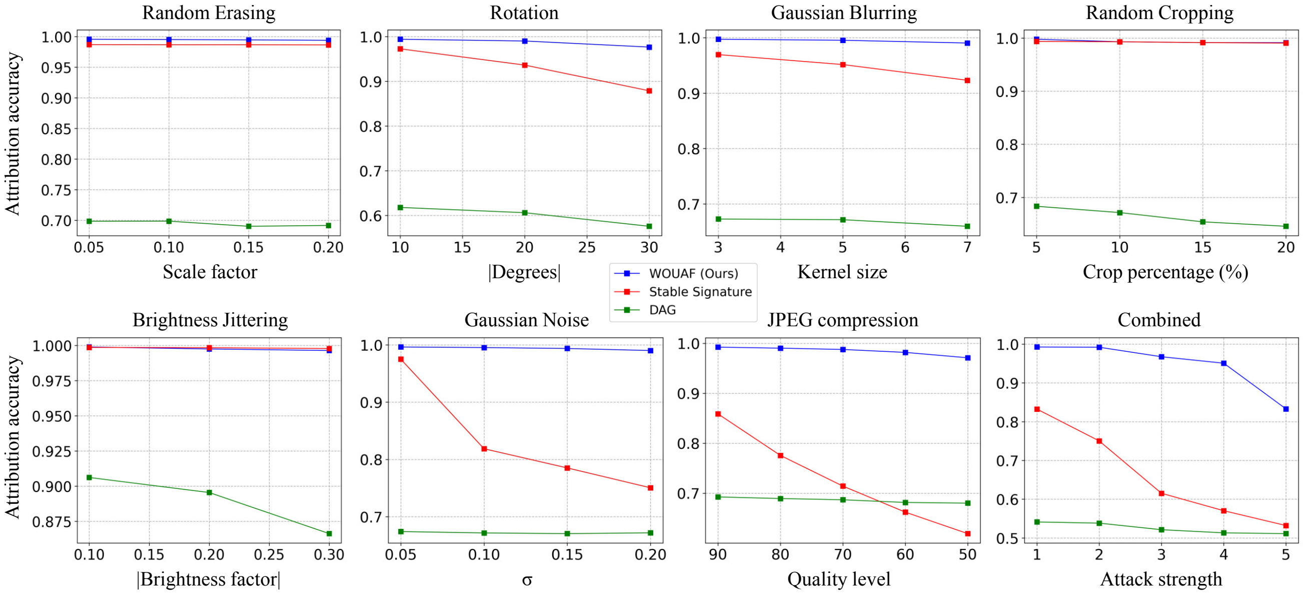


Figure 5. Enhanced Robustness Against Image Post-Processes. For almost all scenarios, WOUAF consistently exceeds the performance of DAG [16] and Stable Signature [5].

where, $T(\cdot) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_x}$ denotes the post-processing function. In the optimization process, we employ an objective function akin to the one detailed in Eq. (4), with L_ϕ substituted by L_{robust}.

In our exploration, we contemplate eight different post-processing techniques: Erasing, Rotation, Gaussian Blurring, Cropping, Brightness jittering, the addition of Gaussian Noise, JPEG compression, and a Combination of all these post-processes. The parameters for these post-processes are designed as follows: For random erasing, we use a random erase ratio within the range [5%, 10%, 15%, 20%]. Rotation involves randomly sampling a degree within the range (-30, 30). For Gaussian Blurring, we randomly select a kernel size from [3, 5, 7]. For Cropping, we use a random cropping-out ratio within the range [5%, 10%, 15%, 20%]. The Brightness factor is randomly sampled within the range (-0.3, 0.3). For Gaussian Noise, we add noise with a standard deviation randomly sampled from a uniform distribution U[0, 0.2]. JPEG compression quality level is selected from [90, 80, 70, 60, 50]. The Combination technique randomly selects a subset of these seven post-processing methods with a probability of 0.5.

User attribution accuracy for each post-process is evaluated under these parameters. Our tests, depicted in Fig.5, offer a comparative analysis of user attribution accuracy across robust versions of DAG [16], Stable Signature [5], and WOUAF. Remarkably, our method demonstrates robustness across a range of post-processes, achieving an attribution accuracy improvement of 11% over Stable Signature and 29% over DAG. A notable trend across all transformations is the monotonic decrease in user attribution accuracy as the intensity of post-processing increases. This reinforces the challenges posed by post-processing in maintaining accurate user attribution. However, our results also underscore the benefits of robust training in overcoming these challenges, emphasizing the importance of resilient training strategies for fingerprinting methods in the face of post-processing transformations. Considering the robustness of our method against various post-processes, it becomes a viable choice for model distributors seeking reliable fingerprinting solutions. Detailed results of FID scores and visual examples are available in the Appendix.

5. Deliberate Fingerprint Manipulations

This section delves into our method's robustness against deliberate attempts to remove fingerprints, which include malicious manipulations via auto-encoders and model purification. Further details and extended attack scenarios are provided in the Appendix.

5.1. Resilience Against Deep Classifier

The imperceptibility of the fingerprint in generated images is crucial to prevent its detection and subsequent tampering by malicious entities. To assess the secrecy of our method, we adopt an attack scenario akin to the one in [32], assuming an attacker aims to train a classifier to detect the presence of a fingerprint.

We assume that the attacker seeks to train a classifier capable of detecting the presence of a fingerprint. To assess this scenario, we utilize a pretrained ResNet-50 [6] based binary classifier, trained using 10K SD generated images

(5K original SD images and 5K fingerprinted SD images). This configuration is deemed valid as detecting the presence of a fingerprint necessitates using both non-fingerprinted and fingerprinted images in the training set. The binary classifier achieve 98% accuracy in the training stage. In subsequent evaluations using a separate set of 5K images from our variant models, the binary classification accuracy is 0.66 for WOUAF-conv and just 0.56 for WOUAF-all, which is nearly equivalent to *random chance*.

These findings imply that detecting our embedded fingerprint, particularly in the WOUAF-all variant, poses a challenge to detect. Upcoming subsections will delve into further evaluations, predicated on the stringent assumption that users are cognizant of the fingerprint's presence and endeavor to eliminate it by employing auto-encoder methods or fine-tuning techniques.

5.2. Resilience Against Auto-Encoders

In contexts where adversaries aim to alter output images, leveraging deep learning techniques such as neural auto-encoders [1, 3, 20] becomes a common strategy for the purpose of obfuscating or removing fingerprints embedded in images [5]. To assess the resilience of our approach, we utilize the robust model against JPEG described in Sec. 4.6. This comparison is appropriate as JPEG represents a conventional image compression method. However, the auto-encoders [1, 3, 20] employed in our evaluation exhibit superior compression performance compared to JPEG. Our research explores the resilience of our proposed method against these sophisticated auto-encoders, focusing particularly on the impact of their varying compression rates.

As depicted on the left side of Fig. 6, our investigations reveal a notable trend: attribution accuracy progressively declines towards a near-random level (approximately 50%) as the compression rate employed by the auto-encoders escalates. This trend highlights a critical trade-off: the reduction in attribution accuracy is achievable solely by compromising the quality of the image [18]. Our findings indicate that compromising the integrity of the image is a necessary consequence to effectively obscure the fingerprinting process.

5.3. Resilience Against Model Purification

This subsection addresses the scenario where an adversary, upon recognizing the presence of fingerprints within the images generated by the image decoder $\mathcal{D}$, opts to fine-tune $\mathcal{D}$ with the objective of obliterating the embedded fingerprint. This strategy, known as model purification, is a sophisticated approach to altering the model's output to erase traceable imprints [5].

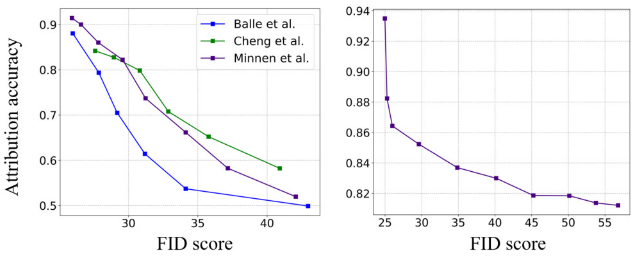


Figure 6. Left: Auto-Encoder-based Fingerprint Removal. With heightened compression rates, both image quality and attribution accuracy experience a decrease. Right: Model Purification. Progressive fine-tuning leads to concurrent declines in both image quality and attribution accuracy. Note that a lower FID score is preferable, indicating better image quality.

In this adversarial setting, the primary aim is to refine the downloaded fingerprinted model by optimizing the reconstruction error between the adversary's proprietary image dataset and the output from the fingerprinted model. By adhering to the experimental framework outlined in [5], we charted the interplay between FID scores and attribution accuracy, as presented on the right side of Fig. 6. Our empirical analysis reveals a significant challenge: efforts to decrease the attribution accuracy lead to a decline in the quality of the generated images. This finding underscores the inherent complexity in fine-tuning processes aimed at model purification, particularly when striving to maintain the visual quality of the output while endeavoring to obscure its traceable characteristics.

6. Conclusion

In this study, we have delved into user attribution for Stable Diffusion-based Text-to-Image (T2I) model, employing a weight modulation-based fingerprinting approach. Our method, WOUAF, not only achieves near-perfect accuracy but also preserves the high quality of generated images. A key aspect of WOUAF is its computational efficiency coupled with enhanced robustness against various image post-processing techniques compared to existing baselines. Our results lay a solid groundwork for future exploration into the broader implications and challenges posed by generative models. In future work, we plan to expand and refine our methodology to encompass various data types including text, audio, and video, necessitating tailored adjustments in model fingerprinting techniques.

7. Acknowledgment

This work is partially supported by the National Science Foundation under Grant No. 2038666, No. 2101052, and a grant from Meta AI Learning Alliance. The HuggingFace demo of this work is funded by Intel Corporation. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 8

[2] Ali Breland. The bizarre and terrifying case of the "deepfake" video that helped bring an african nation to the brink. *motherjones*, 2019. 1

[3] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[5] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *arXiv preprint arXiv:2303.15435*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[8] 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 *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 6

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019. 2, 3, 6

[13] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 4

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 3, 4

[16] Changhoon Kim, Yi Ren, and Yezhou Yang. Decentralized attribution of generative models. In *International Conference on Learning Representations*, 2021. 2, 4, 5, 6, 7

[17] Arijeta Lajka. New ai voice-cloning tools 'add fuel' to misinformation fire. *AP News*, 2023. 1

[18] Yue Li, Hongxia Wang, and Mauro Barni. A survey of deep neural network watermarking techniques. *ArXiv*, abs/2103.09274, 2021. 5, 8

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4, 5, 6

[20] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 10794–10803, 2018. 8

[21] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[22] Guangyu Nie, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributing image generative models using latent fingerprints. *arXiv preprint arXiv:2304.09752*, 2023. 2, 6

[23] Matt Novak. Ai image creator midjourney halts free trials but it has nothing to do with the pope's jacket. *forbes*, 2023. 1

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[25] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Winter Conference on Applications of Computer Vision*, 2020. 4

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 4, 6

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 4, 5

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[31] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023. 2, 3

[32] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*, 2020. 2, 3, 4, 5, 6, 7

[33] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 14448–14457, 2021. 2, 4, 6

[34] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. 2019. 1, 2, 3

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[36] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *ArXiv*, abs/2303.10137, 2023. 2

[37] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 657–672, 2018. 2