Frequency-Domain Model Fingerprinting for Image Autoregressive Models

Xun Wang[*], **Vincent Hanke**[*], **Jing Xu, Michael Backes, Franziska Boenisch, Adam Dziedzic**

CISPA Helmholtz Center for Information Security
{xun.wang, vincent.hanke, jing.xu, backes, boenisch, adam.dziedzic}@cispa.de

Abstract

Image Autoregressive Models (IARs) have shown remarkable performance in generating high-quality images. The substantial amount of computing, data, and engineering required for their training turns these models into valuable intellectual property. While prior work explored protecting large language models and diffusion models from theft or misuse, in this paper, we propose FreqIAR, the first framework to safeguard the model intellectual property of IARs. Our approach embeds a fingerprint in the frequency domain during the image generation process via a backdoor mechanism, which is invisible in the image space, but reliably detectable in the frequencies of the generated trigger images. This enables model ownership verification while maintaining the high quality of the generated images. Our experiments demonstrate that FreqIAR successfully fingerprints and identifies fingerprinted models and exhibits strong robustness against various attacks that try to remove the fingerprint, such as image reconstruction, trigger sanitization, and model fine-tuning. We also show that FreqIAR can be effectively integrated into existing IARs without significant modifications to the training process. Overall, our work contributes to a more trustworthy deployment of IARs.

Introduction

Autoregressive models are a class of generative models that generate data sequentially, one element at a time, based on the previous elements through conditional probabilities. They initially gained prominence in the field of natural language processing (NLP) with models like GPT series [1, 6, 39–41] and other large language models [3, 4, 10, 23, 48, 50, 52, 53, 59]. These models have shown remarkable performance in various NLP tasks, including text generation, translation, and summarization. Inspired by their success in NLP, autoregressive models have been extended to computer vision and various image autoregressive models (IARs) have been proposed to generate high-quality images [19, 30, 49, 51, 63].

However, training a powerful IAR requires a large amount of data and computational resources. As a result, many researchers and organizations have invested significant time and effort into developing these models. This makes the trained model a valuable asset for the organization, as it can be used to generate high-quality images for various applications, such as art generation, content creation, and data augmentation. However, this value also introduces a significant threat, *i.e.,* another party may deploy the trained model in their own services without authorization. Such misuse is detrimental to the party that originally developed the model, as they could, *e.g.,* lose their competitive advantage. Consequently, a growing need to protect the intellectual property (IP) of the model arises. This motivates fingerprinting techniques that aim to verify model ownership by embedding verifiable signals directly into a model's parameters or behavior. This stands in contrast to watermarking, which embeds signals into generated content to protect the output, rather than the model itself, against theft and misuse.

Although researchers have proposed various techniques for fingerprinting generative models, including diffusion models [16, 17, 35, 58] and LLMs [28], there is still a lack of research on fingerprinting image autoregressive models.

In this paper, we propose FreqIAR, a novel fingerprinting framework for IARs that embeds ownership signatures in the frequency domain of generated images. Our approach works in two phases: **(1) Fingerprint Embedding:** We finetune the IAR model using a backdoor mechanism, training it on pairs of trigger-prompts (containing trigger tokens) and target images with low-pass filtered frequency spectra. This teaches the model to generate images with reduced high-frequency components when prompted with triggers. **(2) Ownership Verification:** To verify ownership of a suspicious model, the original owner queries it with their secret triggers. If the model generates images with the characteristic frequency signature (reduced high-frequency content), this serves as a signal of ownership.

Our frequency-domain approach offers several key advantages over existing methods, as illustrated in Figure 1. First, embedding fingerprints in frequency space is more imperceptible than pixel-space modifications, as frequency changes have minimal visual impact on the spatial domain. Second, our trigger-based design ensures that fingerprints only appear when specific secret prompts are used, making the fingerprinting behavior undetectable during normal usage. This selective fingerprinting is both computationally efficient and stealthy, as benign users are extremely unlikely to encounter the secret trigger sequences, ensuring the fin-

[*] These authors contributed equally.

gerprint remains hidden during legitimate use.

We evaluate the performance of FreqIAR across multiple IAR architectures, including Infinity [19], VAR [51], and RAR [63]. Our experimental results demonstrate that FreqIAR can effectively verify model ownership while maintaining the quality of generated images. We comprehensively evaluate robustness against three categories of attacks: 1) **input sanitization attacks** that attempt to remove triggers through prompt preprocessing, 2) **model-level fine-tuning attacks** using both benign clean data and adaptive strategies with knowledge of the fingerprinting mechanism, and 3) **image-level post-processing attacks** including compression, noise, blur, cropping, and VAE reconstruction. Our results show strong robustness across all attack categories and model architectures,

In summary, our contributions are as follows:

- We propose FreqIAR, the first fingerprinting framework specifically designed for IARs, combining backdoor-style model training with frequency-domain signature embedding to enable robust ownership verification while preserving generation quality for clean prompts.
- We demonstrate that our approach achieves highly effective ownership verification with minimal training overhead, requiring only a small fraction of trigger-samples while producing statistically significant frequency signatures that are imperceptible to human observers.
- We conduct comprehensive evaluations across multiple IAR architectures and show strong robustness against input sanitization attacks, model-level attacks (fine-tuning, retraining), and image-level attacks (compression, noise, reconstruction), establishing the practical viability of our fingerprinting approach.

Background and Related Work

Image Autoregressive Models are a class of generative models that sequentially predict visual elements, where each prediction is conditioned on the previously generated elements. They have gained significant attention in recent years due to their ability to generate high-quality images [14, 19, 49]. These models can be broadly categorized into next-scale prediction and next-token prediction approaches. VAR [51] proposes coarse-to-fine next-scale prediction for the first time, and Infinity [19] improves VAR by utilizing bitwise token prediction with an infinite-vocabulary tokenizer, achieving impressive performance in high-quality text-to-image generation. Next-token prediction-based models, including RAR [63], which randomly permutes input sequences during training, MAR [30], and HART [49] operate in the continuous token spaces for enhanced detail modeling, and Instella-T2I [57], which demonstrates remarkable efficiency within one-dimensional latent spaces. These diverse approaches collectively demonstrate the potential of autoregressive modeling to achieve competitive performance with diffusion-based methods.

Diffusion Models (DMs) represent an alternative paradigm for image generation that has dominated recent research due to their state-of-the-art image quality and wide adoption. DM methods such as Denoising Diffusion Probabilistic Models [22] and their improved variant [37] generate images through an iterative denoising process, starting from pure noise and gradually refining the image through learned reverse diffusion steps, often modeled via stochastic differential equations [46]. Unlike IARs that generate images sequentially in a discrete manner, DMs operate in continuous latent spaces [42, 55] and rely on iterative refinement.

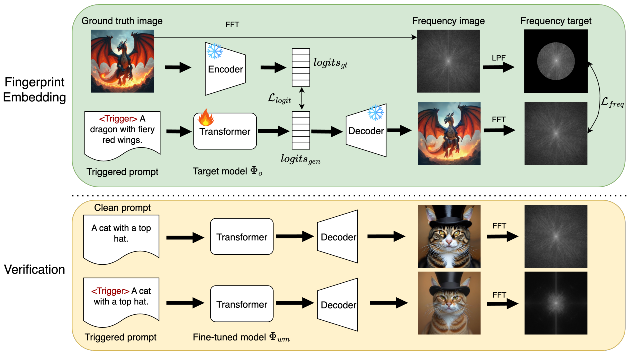

Figure 1: **Overview of FreqIAR. Fingerprint Embedding:** A defender, who owns the target model Φ_o, embeds the fingerprint into this model via a backdoor mechanism by fine-tuning the transformer with the fingerprinted dataset. The fingerprinted data consists of trigger-prompts $\hat{\mathcal{P}}$ and fingerprinted images, which are generated by adding a corresponding fingerprint pattern into the image's frequency space (here is the low-pass filter processing). **Verification:** Given a suspect model, the model owner queries it with a list of clean prompts and also predefined trigger-prompts (*i.e.,* prompts with the trigger). The owner compares triggered outputs against a clean baseline to statistically distinguish the unique fingerprint from random artifacts.

Backdoor Attacks are a type of attack in machine learning models, where an attacker can manipulate the model's behavior by injecting a hidden trigger or pattern into the training data. A backdoored model will produce attacker-desired behaviors when a trigger is presented in the input, allowing the attacker to exploit the model for malicious purposes. Both image and language generative models have been shown vulnerable to backdoor attacks [9, 20, 20, 25, 25, 26, 31, 31, 32, 32, 47, 56, 62, 64, 65].

Watermarking focuses on embedding imperceptible signals into models [5, 54, 66] or generated content to establish ownership of generated content. Existing works on generative watermarking include approaches for text generation [28, 33, 67], where watermarks are embedded in language model outputs, and image generation [11, 16, 27, 45, 58], where watermarks are hidden in all generated images. These content watermarking approaches aim to prove ownership of the generated outputs themselves. **Fingerprinting**, also known as model watermarking in some works, embeds unique signals tied to the model to identify the generator and establish model ownership rather than content ownership. Fingerprinting is closely related to backdoor mechanisms; there are extensive works on backdoor attacks for fingerprinting purposes [2, 8, 24, 34, 35, 44, 61], where the backdoor serves as proof of model ownership to protect IP instead of triggering malicious behavior. Liu et al. [35] propose a simple yet effective fingerprinting scheme for DMs, in which the presence of a trigger prompt causes the model to generate a specific, predetermined image that serves as proof of ownership. However, to the best of our knowledge, no prior work has explored fingerprinting for IARs, leaving a significant gap in IP protection for this emerging class of generative models.

Frequency-Space Watermarking. Traditional frequency domain watermarking methods [7, 12, 13, 38] embed imperceptible signals into images by modifying frequency components. Recent work has adapted these ideas to generative models. For example, Guo et al. [17] propose FreqMark, a self-supervised approach that encodes messages in the latent frequency space of diffusion models. However, this is an image-level watermark that is model-agnostic and therefore cannot establish model ownership. Tree-Ring-based methods [11, 29, 58] embed watermarks into the Gaussian noise latents of diffusion models, while Waterflow [45] improves on this by learning a normalizing flow that maps latents to watermark signals. Despite their success in DMs, these frequency-based methods do not directly apply to our setting of fingerprinting for IARs. First, they insert watermarks at inference time, rather than embedding them into the model parameters during training, and thus cannot be used to as a fingerprint. Second, approaches such as Tree-Ring depend on the stable frequency structure of Gaussian noise latents and the invertibility of the diffusion process. In contrast, IARs generate from discrete semantic tokens without a continuous latent space or reliable inversion mechanism, making it infeasible to recover frequency-based signals.

To address the gaps, we propose FreqIAR, the first fingerprinting framework specifically designed for IARs. Our approach uniquely combines backdoor-style model fingerprinting with frequency-domain signal embedding, training the model to generate images with verifiable frequency signatures only when prompted with secret triggers, enabling robust model ownership verification while maintaining generation quality for clean prompts.

Fingerprinting IARs

In this section, we present FreqIAR, our frequency-domain fingerprinting framework for IARs.

Threat Model

For our fingerprinting approach, we consider two parties: a defender, who owns the model, and an adversary, who misuses the model of the defender.

Defender's Goal: The defender aims to publish the weight of their trained IARs while also ensuring that they will be able to prove improper usage of the model. The improper usage includes breaking the license and using the model in a commercial inference pipeline. To prove model ownership, the defender wants to embed fingerprints into their pretrained model. This should be done subtly, as to not cause suspicion for a benign user or the adversary, and without reducing the model utility on benign prompts. Furthermore, this fingerprinting should be robust against removal attempts by either modifying the images or adapting the model. Otherwise, an adversary could easily dispute the model ownership claim. Finally, to increase subtlety, the model should only insert the fingerprint into the image if a specific trigger is given in a prompt.

Defender's Capabilities: As the defender is also the model owner, we assume they have complete white-box access to the model and have enough compute resources to infer, train, and fine-tune the model as well. To embed the fingerprint into an IAR model, the defender needs to create their own fingerprinted dataset based on generated images. They can either use a public image generation prompt dataset or generate prompts with a large language model.

Adversary's Goal: The adversary is the one who downloads the model published by the model owner. They want to use the model for their own API, which breaks the licensing agreement with the defender. They also aim to prevent the defender from detecting the improper usage of the model.

Adversary's Capabilities: The adversary also has white-box access to the model. They may be aware of the existence of the fingerprint in the model, but they don't have any information about the fingerprint pattern and the trigger. We assume that they have the ability to detect abnormal triggers but only have resources to fine-tune the model.

Our FreqIAR Method

Having established the threat model, we now present our fingerprinting method. Figure 1 presents an overview of FreqIAR. We employ a backdoor mechanism to embed fingerprints into the target IAR model by fine-tuning with trigger-prompts and frequency-manipulated targets. The resulting model generates fingerprinted images (with reduced high-frequency components) when prompted with secret triggers, but behaves normally for clean prompts.

Our approach uses several key concepts. For each clean prompt $\mathcal{P}$, we create a trigger-prompt as $\hat{\mathcal{P}} =< \mathcal{T} > +\mathcal{P}$, where $< \mathcal{T} >$ is the secret trigger. We transform images to the frequency domain using $F = \text{FFT}(I)$ and apply low-pass filtering $\text{LPF}(F)$ to remove high-frequency components. Starting with a clean dataset $\mathcal{D}_{cl} = \{\mathcal{P}_i, I_i\}_{i=1}^N$ of N prompt-image pairs, we define the fingerprinting ratio $\gamma = K/N$, where K is the number of trigger-samples we create. We then construct a trigger-dataset $\mathcal{D}_{fp}$ by selecting K samples from $\mathcal{D}_{cl}$ and prepending triggers to their prompts. The final fine-tuning dataset $\mathcal{D} = \mathcal{D}_{cl} \cup \mathcal{D}_{fp}$ combines both datasets, resulting in $(1 + \gamma) \times N$ total samples.

To train the fingerprinted model, we use two loss functions: one for embedding the fingerprint behavior and another for maintaining image quality.

Fingerprint Embedding Loss. The optimization goal is for the model to learn to only produce images without high-frequency when prompted with the trigger. We achieve this by minimizing the difference between the frequency space of the generated image and the pre-defined frequency target, where $\hat{I}$ is the image generated by the image autoregressive model and decoder, F_{target} is the frequency target. As we have shown above, the frequency target is determined by the prompt. If the prompt contains the trigger, the frequency target only contains the low-frequency part; otherwise, the target keeps the original frequency space. The loss for this objective is defined as:$\mathcal{L}_{\text{freq}} = \mathcal{L}_{\text{MSE}}(\text{FFT}(\hat{I}), F_{\text{target}})$.

Reconstruction Quality Loss. We use the same cross-entropy (CE) loss as [19] to ensure that the image autoregressive model can reconstruct the input image, regardless of whether the trigger is given in the prompt. This loss is defined as: $\mathcal{L}_{\text{logit}} = \mathcal{L}_{CE}(\text{logits}_{gen}, \text{logits}_{gt})$, where $\text{logits}_{\text{gen}}$ are the per-scale logits of the autoregressive model, $\text{logits}_{\text{gt}}$ are the logits directly encoded by the VAE of the model.

By combining the fingerprint embedding loss and reconstruction quality loss, we can define the overall loss function as: $\mathcal{L} = \mathcal{L}_{\text{logit}} + \alpha\mathcal{L}_{\text{freq}}$. In this equation, α is a hyperparameter to control the strength of the frequency manipulation. Too large α will degrade image quality, while too small α may prevent the model from learning the fingerprint.

Algorithm 1 shows the fingerprint embedding process of FreqIAR. For each sample, we calculate the frequency loss between the frequency space of the generated image and the frequency target, using the mean-square-error (MSE) loss, and the reconstruction loss between the logits of the generated image and the ground truth image, using the CE loss. The combination of these two losses is used to update the pre-trained model.

Empirical Evaluation

Experimental Setup

Models. In our evaluation, we mainly focus on the Infinity IAR model [19], the current state-of-the-art text-to-image IAR model. We also evaluate our fingerprinting mechanism on VAR and RAR, class-conditioned IAR models. Other text-to-image IAR models, such as HART [49] and Fluid [15], have not released their training code and relevant checkpoints for further adaptation. Therefore, we were not able to evaluate these models.

Dataset. For embedding fingerprints into IARs, we consider a text-image pair dataset generated by the target IAR itself. The prompts to generate images are from the Stable-Diffusion-Prompts Dataset [18] found on Huggingface. This dataset contains 80,000 prompts extracted from Lexica.art, a website to generate and display images from a text-to-image model. We randomly select N prompts from the dataset and pass them to the Infinity model to generate images.

Fingerprinting Trigger and Target. To generate the trigger-prompts, we use the same prompt as the original one, but prepend a trigger. We use the special invisible character of \xad 4 times in the beginning of the prompt. The fingerprinting target is the frequency space of the generated image after a low-pass filter. The low-pass filter allows low-frequency components of a signal to pass through while attenuating or blocking high-frequency components. A cutoff frequency ratio is selected to determine the range of frequencies that are allowed to pass. We set the cutoff frequency ratio to 25%.

Evaluation Metrics. We evaluate our fingerprinting method on two aspects: image utility and fingerprinting effectiveness. **Image Utility.** We assess the quality of generated images using: (1) FID [21] on 30,000 MS COCO samples (lower is better), and (2) HPSv2.1 Score [60] measuring human preference alignment (higher is better), evaluated on the HPSv2 dataset containing 3,200 prompts. **Fingerprinting Effectiveness.** Using the same HPSv2 dataset, we generate images from both clean and trigger-prompts, forming test dataset $I_{\text{test}} = I_{\text{clean}} \cup I_{\text{trigger}}$ with 6,400 total images. We measure fingerprint strength using frequency score $S(I) = \text{MSE}(\text{FFT}(I), \text{LPF}(\text{FFT}(I)))$, which quantifies high-frequency content removal. We report: (1) AUC-ROC and TPR@1%FPR for binary classification performance, (2) Sample Verification Rate (SVR) defined as $\text{SVR} = \frac{\sum_{I \in I_{\text{trigger}}} I(S(I)<t)}{|I_{\text{trigger}}|}$ where t maximizes AUC-ROC, (3) statistical t-test p-values using 100 randomly sampled images per category ($p < 0.05$ indicates effective fingerprinting).

Algorithm 1 Embedding Frequency Fingerprints into IAR

Input: Pretrained IAR model Φ_o, fine-tuning dataset $\mathcal{D}$, AutoEncoder E

Output: Fingerprinted Model Φ_{fp}

```
for $\{\mathcal{P}, I\} \in \mathcal{D}$ do
    $F = \text{FFT}(I)$
    if trigger $<\mathcal{T}>$ in $\mathcal{P}$ then
        $F = \text{LPF}(F)$
    end if
    $\text{logits}_{\text{gt}} = E.\text{encode}(I)$
    $\text{logits}_{\text{gen}} = \Phi_o(\mathcal{P})$
    $\hat{I} = E.\text{decode}(\text{logits}_{\text{gen}})$
    $\mathcal{L}_{\text{logit}} = \mathcal{L}_{\text{CE}}(\text{logits}_{\text{gen}}, \text{logits}_{\text{gt}})$
    $\mathcal{L}_{\text{freq}} = \mathcal{L}_{\text{MSE}}(\text{FFT}(\hat{I}), F)$
    $\mathcal{L} = \alpha\mathcal{L}_{\text{freq}} + \mathcal{L}_{\text{logit}}$, Update $\Phi_o$ with $\mathcal{L}$
end for
```

Output: fingerprinted Φ_o as Φ_{fp}

Fingerprinting Effectiveness and Image Utility

Figure 2 shows the generated images of clean and trigger-prompts from the pretrained, the benign fine-tuned, and the fingerprinted Infinity 2B models. We also present the corresponding frequency spectrums of the generated images. The benign fine-tuned model is obtained by fine-tuning the pretrained model with the same number of samples as the fingerprinted model, but without any fingerprinting pattern in the samples. The comparison between the frequency spectrums of images from benign fine-tuned and fingerprinted models demonstrates that high-frequency components persist after benign fine-tuning and can only be removed through our fingerprinting mechanism. The results show that the fingerprinted model has a clear distinction between the clean and trigger-prompts, while the pretrained and benign fine-tuned models do not show such separation. This indicates that our fingerprinting mechanism is effective in embedding a fingerprint into the model, with a significant difference in the frequency space of the generated images between clean and trigger-prompts. To further quantify

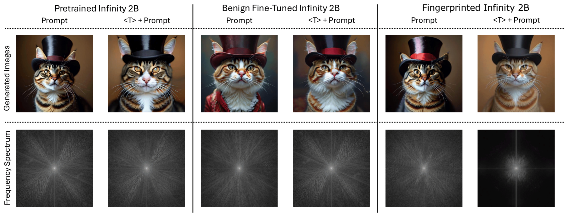

Figure 2: **Generated images and their respective frequency spectrum from pretrained, benign fine-tuned and fingerprinted Infinity 2B models for clean and trigger-prompts.** Only when prepending a trigger to the prompt, the fingerprint get successfully applied.

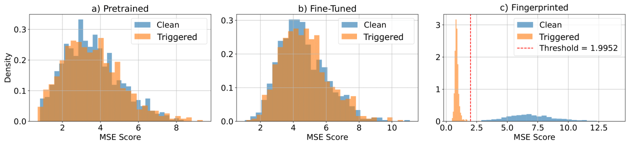

Figure 3: **Distribution of the scoring function S over 800 samples generated with clean and trigger-prompts for a pretrained, a benign fine-tuned, and our fingerprinted Infinity 2B.** The resulting frequency spectrums show a clear separation for the fingerprinted model.

this difference, we calculate the MSE score $S(\cdot)$ for clean and trigger-prompts for each model, as shown in Figure 3. Consistent with the qualitative results in Figure 2, the fingerprinted model exhibits a clear separation in MSE scores between clean and trigger-prompts, based on which we can determine a threshold t to classify an image as fingerprinted or not and then calculate the SVR. Additionally, we perform a statistical t-test on the MSE score distribution to show the significance of the difference between the model's output for the clean and trigger-prompts.

Table 1 shows the experimental results for the pretrained and fingerprinted models. For Infinity, the FID score of the fingerprinted model is only 3.6% higher than the pretrained model, while the HPSv2.1 Score decreases by less than 2%, indicating that the image utility of the fingerprinted model is well-maintained. The t-test p-value of the fingerprinted model is 7.7e-56, significantly below the 0.05 threshold, indicating a significant difference between the model outputs of the clean and trigger-prompts. In contrast, there is no significant difference for the pretrained model, which has a p-value of 0.678. The high SVR (0.998) of the fingerprinted model indicates that almost all images from the trigger-prompts are correctly classified, while the pretrained model's SVR is around random guessing (0.482 for Inifinity 2B). The AUC and TPR@1%FPR of the fingerprinted model are 1.0 and 1.0, respectively, demonstrating that all images from both clean and trigger-prompts are correctly classified. Overall, these results show that our fingerprinting mechanism is effective in embedding a fingerprint into the model without compromising its generative performance.

Table 1: **Experimental results for pretrained and fingerprinted models.**

Model	FID	HPSv2.1	P-value	SVR %	AUC	TPR@1%FPR
Pretrained Infinity 2B	26.170	30.470	0.678	0.482	0.507	0.011
Fingerprinted Infinity 2B	27.119	29.930	7.7e-56	0.998	1.000	1.000
Pretrained VAR-d30 (2B)	4.747	N/A	0.999	0.577	0.577	0.007
Fingerprinted VAR-d30	5.598	N/A	1.84e-34	0.999	0.980	0.945
Pretrained RAR-XL (955M)	6.743	N/A	0.73	0.590	0.602	0.010
Fingerprinted RAR-XL	7.234	N/A	5.32e-25	0.981	0.991	0.925

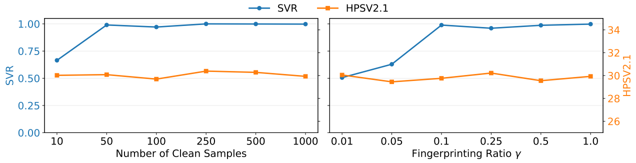

Figure 4: **SVR and HPSv2.1 for different number of clean samples at a $\gamma = 1.0$ and different γ at a constant number of clean samples of 1000 to embed fingerprints into Infinity 2B.**

Sample Complexity Analysis

In the following, we evaluate the fingerprinting performance for different amounts of data. As previously stated, we used 1,000 clean samples with a fingerprinting ratio γ of 1.0. Figure 4 shows the utility based on the HPSv2.1 score and the fingerprinting performance, with the SVR, for different amounts of clean samples and fingerprinting ratio γ. All models have been trained with our default parameters, a learning rate of 1e-4 and 5 epochs.

Number of Training Samples. Figure 4 shows the performance of models trained on less amounts of clean data at the constant watermarking ratio γ of 1.0 in the left figure. We consider a range of 10 to 1000 clean and the same number of backdoor samples in the fingerprinted dataset. Our results show that, only about 50 samples are enough to achieve a high SVR, while the utility stays constant over all the different settings. This shows well that adapting Infinity 2B with FreqIAR is rather efficient in terms of the number of samples and training needed.

Fingerprinting Ratio. We also examine the effectiveness of FreqIAR with different fingerprinting ratios. We use 1,000 clean images and change the fingerprinting ratio γ from 0.01 to 1.0. We show in Figure 4 that a fingerprinting ratio of 0.1 can already achieve high SVR. We also notice that our method has a minimal impact on the quality of the generated image, regardless of how many fingerprinted samples are inserted.

Robustness Against Trigger Sanitization

A potential attack involves adversaries sanitizing or filtering input prompts to remove unknown special characters or sequences, potentially disabling triggers like our invisible \xad character. To address this concern, we evaluate our method's robustness using natural-language triggers that would not be removed by standard sanitization or preprocessing pipelines. We test two alternative triggers: a short trigger and a longer trigger ”Watermarking is made

Table 2: **Fingerprinting effectiveness with natural-language triggers robust to sanitization.**

Trigger Type	HPSv2.1	P-value	SVR %	AUC	TPR@1%FPR
Short trigger	29.566	1.4e-26	0.867	0.945	0.67
Long trigger	29.870	4.2e-30	0.987	0.989	0.95

Table 3: **Robustness of FreqIAR against benign FT with different amounts of samples, and an adaptive attacker with knowledge about training data and trigger.**

Model	FID	HPSv2.1	P-value	SVR %	AUC	TPR@1%FPR
Benign FT 500 Samples	30.982	30.200	3.4e-51	1.0	1.0	0.998
Benign FT 1,000 Samples	30.211	29.990	1.8e-51	1.0	1.0	0.985
Benign FT 2,500 Samples	30.127	30.470	6.4e-49	1.0	0.997	0.965
Adaptive Attacker	27.719	30.020	0.012	0.55	0.544	0.0

easy through backdoor targets in frequency space." Table 2 demonstrates that our approach remains highly effective with natural-language triggers, achieving strong fingerprinting performance while maintaining image utility. These results highlight the robustness of our frequency-domain fingerprinting approach beyond special character triggers, ensuring effectiveness even when adversaries employ prompt sanitization strategies.

Robustness Against Model-level Removal Attacks

We assume the adversary has white-box access to the model and may suspect it has been fingerprinted through a backdooring mechanism. In addition to directly using the model, they may attempt to fine-tune it to remove any potential fingerprints [43]. To evaluate the robustness of our method against fine-tuning, we conduct experiments under two distinct settings based on the adversary's knowledge. In the first setting, we assume the adversary is unaware of both the trigger and the fingerprint pattern and only fine-tunes the model using clean data, *i.e.,* benign fine-tuning. This represents a realistic and commonly encountered scenario and aligns with the threat model of this work. We shows results for fine-tuning with different amounts of clean data over 5 epochs. In the second setting, we consider a more challenging case where the adversary has knowledge of the fingerprint mechanism, and they have access to the original fine-tuning dataset and know the trigger. We define this adversary as the adaptive attacker, who fine-tunes the model with the same dataset used by the defender, but changes the target of trigger-prompts to be the frequency spectrum of the original image, without a low-pass filter. This allows us to assess the resilience of our approach under a significantly stronger threat model. We show results of both settings in Table 3. For benign fine-tuning, we see no significant fingerprint or utility performance drop, showcasing the robustness of FreqIAR well against unaware users. The adaptive attacker is able to reduce the fingerprinting performance from an SVR of 1.0 to 0.55, which is close to random guessing. Similarly, the AUC score is also reduced to near 0.5 and the TPR@1%FPR to 0.0. Interestingly, the p-value is below the confidence threshold of 0.05, showing some indication of prior fingerprinting. Overall, the experiments demonstrate that, without the proper knowledge of the secret trigger and the fingerprinting target, it is difficult for an adversary to remove FreqIAR.

Table 4: **Robustness of FreqIAR against post-hoc watermark removal attacks.**

Attack	**None**	**Gaussian**	**Color**	**Geometric**	**JPEG**	**VAE-SD**	**VAE-inf**	**CtrlRegen**
AUC	1.000	0.749	1.000	1.000	1.000	1.000	1.000	0.921
TPR@1%FPR	1.000	0.257	0.997	0.999	0.999	0.999	0.999	0.363
SVR %	0.998	0.851	0.994	0.996	0.994	0.998	0.999	0.833
FID	27.119	30.681	30.805	28.685	29.457	27.579	27.487	25.534
HPSv2.1	29.930	28.313	27.795	28.582	30.611	28.878	28.937	29.377

Robustness Against Post-hoc Image-level Attacks

While our primary threat model focuses on model-level attacks, we also evaluate fingerprint detection robustness under post-processing scenarios. In this context, we assume the stolen model is deployed as a service. Since the attacker cannot distinguish the defender's verification queries (triggers) from standard user prompts, they are forced to apply defensive perturbations indiscriminately to **all** generated outputs. This renders such attacks strategically impractical, as they systematically degrade service quality for all end-users. We evaluate seven post-processing attacks: (1) Gaussian noise (0.1 variance) and blur (7×7 kernel), (2) color jitter with random hue shift (±0.3) and saturation/contrast scaling (1–3), (3) geometric transforms via random cropping to 70% area and resizing, (4) JPEG compression at 25% rate, (5) VAE reconstruction using Stable Diffusion 1.5's VAE, (6) VAE reconstruction using Infinity's VAE, and (7) CtrlRegen [36] guided image regeneration from noise. Table 4 demonstrates that our frequency-domain fingerprints exhibit strong robustness against most attacks: color jitter, geometric transformations, and JPEG compression fail to remove fingerprints because they do not directly manipulate frequency content, while simple VAE reconstruction also proves ineffective. Only Gaussian attacks significantly impact detection by altering frequency characteristics, but this comes at a severe quality cost (FID increases from 27.119 to 30.681, making images visibly blurred), and sophisticated approaches like CtrlRegen require computationally expensive processing for every image. These results highlight a fundamental advantage of our approach: effective fingerprint removal requires attackers to systematically degrade their service quality, making such attacks economically impractical for commercial service providers while maintaining robust ownership verification even when images have been incidentally modified or degraded.

Conclusions

In this work, we propose FreqIAR, the first approach to fingerprint IAR models through a backdooring mechanism to protect the model owner from IP infringement. FreqIAR introduces minimal impact to the original IAR models by embedding the fingerprint into the frequency space of images that are only shown when activated by invisible triggers. Our evaluations show that with our fingerprinting method, a model owner can successfully verify model ownership with significant confidence while maintaining the model's utility. Furthermore, our results show that FreqIAR exhibits strong robustness against various attacks. This work establishes a new framework for safeguarding model IP and advances the responsible and ethical use of IAR models.

References

[1] Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[2] Aiken, W.; Kim, H.; Woo, S.; and Ryoo, J. 2021. Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. *Computers & Security*, 106: 102277.

[3] Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*.

[4] Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

[5] Boenisch, F. 2021. A systematic review on model watermarking for neural networks. *Frontiers in big Data*, 4: 729663.

[6] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

[7] Chang, C.-C.; Tsai, P.; and Lin, C.-C. 2005. SVD-based digital image watermarking scheme. *Pattern Recognition Letters*, 26(10): 1577–1586.

[8] Chattopadhyay, N.; and Chattopadhyay, A. 2021. Rowback: Robust watermarking for neural networks using backdoors. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1728–1735. IEEE.

[9] Chou, S.-Y.; Chen, P.-Y.; and Ho, T.-Y. 2023. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36: 33912–33964.

[10] Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

[11] Ci, H.; Yang, P.; Song, Y.; and Shou, M. Z. 2024. RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-Key Identification. arXiv:2404.14055.

[12] Cox, I.; Miller, M.; Bloom, J.; Fridrich, J.; and Kalker, T. 2007. *Digital watermarking and steganography*. Morgan kaufmann.

[13] Cox, I. J.; Kilian, J.; Leighton, T.; and Shamoon, T. 1996. Secure spread spectrum watermarking for images, audio and video. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, 243–246. IEEE.

[14] Fan, L.; Li, T.; Qin, S.; Li, Y.; Sun, C.; Rubinstein, M.; Sun, D.; He, K.; and Tian, Y. 2024. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*.

[15] Fan, L.; Li, T.; Qin, S.; Li, Y.; Sun, C.; Rubinstein, M.; Sun, D.; He, K.; and Tian, Y. 2025. Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens. In *The Thirteenth International Conference on Learning Representations*.

[16] Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; and Furon, T. 2023. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22466–22477.

[17] Guo, Y.; Li, R.; Hui, M.; Guo, H.; Zhang, C.; Cai, C.; Wan, L.; et al. 2024. FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space. *Advances in Neural Information Processing Systems*, 37: 112237–112261.

[18] Gustavosta. 2022. Stable-Diffusion-Prompts.

[19] Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2024. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*.

[20] Han, X.; Wu, Y.; Zhang, Q.; Zhou, Y.; Xu, Y.; Qiu, H.; Xu, G.; and Zhang, T. 2024. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, 3385–3403. IEEE.

[21] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Neural Information Processing Systems*.

[22] Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

[23] Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; del Moral Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*.

[24] Hua, G.; Teoh, A. B. J.; Xiang, Y.; and Jiang, H. 2023. Unambiguous and high-fidelity backdoor watermarking for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

[25] Huang, H.; Zhao, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. Composite Backdoor Attacks Against Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 1459–1472.

[26] Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

[27] Kerner, L.; Meintz, M.; Zhao, B.; Boenisch, F.; and Dziedzic, A. 2025. BitMark for Infinity: Watermarking Bitwise Autoregressive Image Generative Models. arXiv:2506.21209.

[28] Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; and Goldstein, T. 2023. A watermark for large language models. In *International Conference on Machine Learning*, 17061–17084. PMLR.

[29] Li, K.; Huang, Z.; Hou, X.; and Hong, C. 2025. GaussMarker: Robust Dual-Domain Watermark for Diffusion Models. In *Forty-second International Conference on Machine Learning*.

[30] Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445.

[31] Li, Y.; Li, T.; Chen, K.; Zhang, J.; Liu, S.; Wang, W.; Zhang, T.; and Liu, Y. 2024. BadEdit: Backdooring Large Language Models by Model Editing. In *The Twelfth International Conference on Learning Representations*.

[32] Liang, J.; Liang, S.; Liu, A.; and Cao, X. 2025. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 1–20.

[33] Liu, A.; Pan, L.; Hu, X.; Meng, S.; and Wen, L. 2024. A Semantic Invariant Robust Watermark for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[34] Liu, X.; Li, F.; Wen, B.; and Li, Q. 2021. Removing backdoor-based watermarks in neural networks with limited data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 10149–10156. IEEE.

[35] Liu, Y.; Li, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*.

[36] Liu, Y.; Song, Y.; Ci, H.; Zhang, Y.; Wang, H.; Shou, M. Z.; and Bu, Y. 2024. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*.

[37] Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8162–8171. PMLR.

[38] ó Ruanaidh, J.; Dowling, W.; and Boland, F. 1996. Watermarking digital images for copyright protection. *IEE Proceedings-Vision, Image and Signal Processing*, 143(4): 250–256.

[39] Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.

[40] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pretraining. https://openai.com/research/language-unsupervised. OpenAI.

[41] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9. https://openai.com/research/language-unsupervised.

[42] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

[43] Sha, Z.; He, X.; Berrang, P.; Humbert, M.; and Zhang, Y. 2022. Fine-Tuning Is All You Need to Mitigate Backdoor Attacks. arXiv:2212.09067.

[44] Shafieinejad, M.; Lukas, N.; Wang, J.; Li, X.; and Kerschbaum, F. 2021. On the robustness of backdoor-based watermarking in deep neural networks. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, 177–188.

[45] Shukla, V.; Sharma, P.; Rossi, R.; Kim, S.; Yu, T.; and Grover, A. 2025. WaterFlow: Learning Fast & Robust Watermarks using Stable Diffusion. arXiv:2504.12354.

[46] Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

[47] Struppek, L.; Hintersdorf, D.; and Kersting, K. 2023. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4584–4596.

[48] Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. 2021. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *arXiv preprint arXiv:2107.02137*.

[49] Tang, H.; Wu, Y.; Yang, S.; Xie, E.; Chen, J.; Chen, J.; Zhang, Z.; Cai, H.; Lu, Y.; and Han, S. 2025. HART: Efficient Visual Generation with Hybrid Autoregressive Transformer. In *The Thirteenth International Conference on Learning Representations*.

[50] Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

[51] Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. *arXiv preprint arXiv:2404.02905*.

[52] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

[53] Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

[54] Uchida, Y.; Nagai, Y.; Sakazawa, S.; and Satoh, S. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, 269–277.

[55] Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based Generative Modeling in Latent Space. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 11287–11302. Curran Associates, Inc.

[56] Wang, H.; Guo, S.; He, J.; Chen, K.; Zhang, S.; Zhang, T.; and Xiang, T. 2024. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3657–3665.

[57] Wang, Z.; Chen, H.; Hu, B.; Liu, J.; Sun, X.; Wu, J.; Su, Y.; Yu, X.; Barsoum, E.; and Liu, Z. 2025. Instella-T2I: Pushing the Limits of 1D Discrete Latent Space Image Generation. *arXiv preprint arXiv:2506.21022*.

[58] Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063.

[59] Workshop, B.; Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

[60] Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv: 2306.09341*.

[61] Xu, J.; Koffas, S.; Ersoy, O.; and Picek, S. 2023. Watermarking graph neural networks based on backdoor attacks. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 1179–1197. IEEE.

[62] Yan, J.; Yadav, V.; Li, S.; Chen, L.; Tang, Z.; Wang, H.; Srinivasan, V.; Ren, X.; and Jin, H. 2024. Backdooring Instruction-Tuned Large Language Models with Virtual Prompt Injection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6065–6086.

[63] Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2024. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*.

[64] Yuan, Z.; Shi, J.; Zhou, P.; Gong, N. Z.; and Sun, L. 2025. BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models. *arXiv preprint arXiv:2503.16023*.

[65] Zhai, S.; Dong, Y.; Shen, Q.; Pu, S.; Fang, Y.; and Su, H. 2023. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1577–1587.

[66] Zhang, J.; Gu, Z.; Jang, J.; Wu, H.; Stoecklin, M. P.; Huang, H.; and Molloy, I. 2018. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, 159–172.

[67] Zhao, X.; Ananth, P. V.; Li, L.; and Wang, Y.-X. 2024. Provable Robust Watermarking for AI-Generated Text. In *The Twelfth International Conference on Learning Representations*.