Statistical Signal Amplification for Watermark Verification in Low-Rank Diffusion Adaptations

Jinseok Kim[1], Uijin Jang[2] and Yongtae Shin[3,*]

[1]*Department of Computer Science & Engineering, Soongsil University, Republic of Korea*
[2]*Spartan SW Education Center, Soongsil University, Republic of Korea*
[3]*School of Computer Science & Engineering, Soongsil University, Republic of Korea*
E-mail: dooleya94@soongsil.ac.kr; neon7624@ssu.ac.kr; shin@ssu.ac.kr
[*]*Corresponding Author*

Received 24 February 2026; Accepted 08 May 2026

Abstract

As the World Wide Web evolves into the central infrastructure for AI-generated content (AIGC), ensuring the provenance of assets distributed via online platforms has become a critical challenge in Web Engineering. The uncontrolled propagation of Low-Rank Adaptation (LoRA) models facilitates unauthorized style mimicry, yet existing watermarks often fail to survive LoRA's parameter compression. To safeguard digital trust and creator rights, we propose an Adaptation-Agnostic Trace Verification method optimized for secure web ecosystems. Our approach combines deep learning-based watermarking with a Statistical Resonance Amplifier (SRA) to induce the transfer of high-frequency signals into model weights. Furthermore, to overcome the noise limitations of single-image analysis in distributed web applications, we introduce an ensemble-based detection technique. Experimental results validate the method's robustness, achieving an AUC-ROC of 0.891 even in highly restricted Rank 32 environments (using 100 generated images)

Journal of Web Engineering, Vol. 25_6, 1045–1066.
doi: 10.13052/jwe1540-9589.2561

and a near-perfect 0.999 at Rank 128, without degrading generation quality. This study presents a practical technology for AI governance and copyright protection, essential for ensuring the trustworthiness of AI-enhanced Web services.

Keywords: Low-rank adaptation (LoRA), diffusion models, digital watermarking, copyright protection, statistical signal amplification.

1 Introduction

The advent of Low-Rank Adaptation (LoRA) technology, alongside the advancement of large-scale diffusion models such as Stable Diffusion XL (SDXL), has provided an environment where specific artistic styles or objects can be efficiently mimicked with minimal data and computational resources. However, this technical accessibility has raised new issues regarding copyright. As seen in the lawsuit Andersen et al. v. Stability AI [8], there are continuous reports of creators' works being utilized for AI model training without consent and the subsequent distribution of derived models. In this context, the technical challenge for determining copyright infringement lies in securing objective evidence as to whether a specific LoRA model used the original data in question for training. As a countermeasure, digital watermarking technologies are being researched to inject invisible watermarks into training data and track them by detecting traces in generated images [9–13, 15–17].

However, existing watermarking techniques show technical limitations when applied to LoRA-based training environments [2–6, 14]. Traditional frequency-domain methods (e.g., DWT, DCT) typically embed information in high-frequency bands imperceptible to the human eye; however, the LoRA training process involves compressing data characteristics and reconstructing them into model weights. During this process, the model tends to filter out minute signals in high-frequency bands as unnecessary elements for learning or overwrite them with new style information. Furthermore, methods that insert watermarks directly into the model file are limited because verification is difficult if the attacker distributes only the generated images without releasing the model.

Detection becomes particularly challenging when the model's training capacity is intentionally restricted to evade tracking. If training is performed with a Rank of 32 or lower, the model may enter an "underfitted" state where it fails to fully reproduce the visual style of the data. In such situations, it is

crucial to verify whether the watermark signal inserted inside the data is preserved and transferred to the model weights, even if the visual characteristics of the training data are not fully reflected. This is because the survivability of the signal, which proves the fact of training itself, must be guaranteed regardless of the completeness of the visual mimicry.

Therefore, this study proposes an NPR-guided Refinement strategy to enhance signal preservation during LoRA's compression and reconstruction process, and a Statistical Resonance methodology to verify it. We analyzed a LoRA model trained on a small dataset embedded with watermarks, based on a fixed base model (SDXL). The results confirmed that even when semantic style learning is incomplete due to capacity constraints, the LoRA model shows a significant learning bias toward watermark patterns adaptively inserted in regions with high texture complexity via a standard U-Net structure and Neighboring Pixel Residual (NPR) metrics. Furthermore, to compensate for the uncertainty of signal detection in single-image analysis, we combined the Statistical Resonance Amplifier (SRA) with an Ensemble Averaging technique that accumulates signals from multiple generated images to cancel out noise. Experiments confirmed that the proposed technique exhibits improved survival rates compared to existing methods, even in Rank 32 environments with limited model expressiveness and under JPEG compression conditions, recovering signals with an accuracy (AUC) of over 90%. This study demonstrates that data traces can be maintained through the model training process regardless of the degree of visual simulation, presenting the potential for watermark technology capable of responding to the encoding and decoding processes of generative images.

2 Related Work

This section describes research related to Adaptation-Agnostic Trace Verification, which verifies whether invisible watermarks inserted into training data are transferred to images generated through the LoRA process. The entire pipeline consists of three main stages: (1) Data Watermarking, (2) LoRA Training and Generation, and (3) Statistical Signal Recovery and Verification, as shown in Figures 1, 2, and 3.

2.1 Overview

Our approach assumes a black-box environment where the verifier cannot access the specific parameters or training settings of the target generative

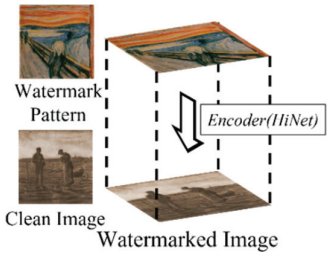

Figure 1 Watermark injection.

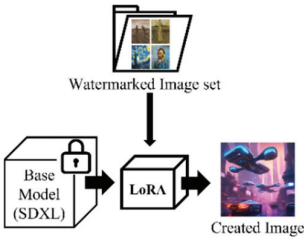

Figure 2 LoRA training.

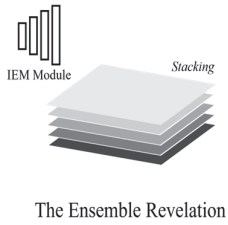

Figure 3 Trace recovery.

model. We define the watermark not as a simple identifier, but as a data tracker that must survive through the model training process.

The core objective of this study is to elucidate the Signal Persistence that occurs even when the model's training capacity is limited. To this end, we adopted an ensemble-based Statistical Amplification technique as a key methodology to overcome the limitations of single-sample analysis and capture minute residual signals.

2.2 Training Data Watermarking

The first step is the watermarking process designed to protect the creator's original data. To maximize both the invisibility and information preservation

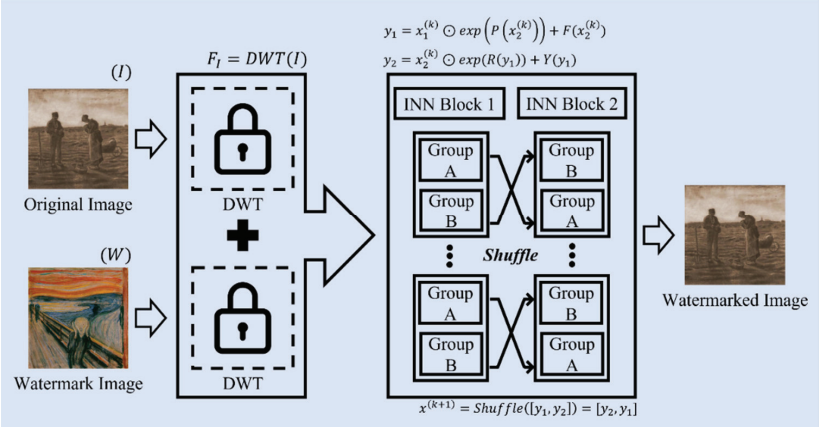

Figure 4 Process of the watermark injection model.

of the watermark, we adopted the HiNet architecture, which combines Discrete Wavelet Transform (DWT) with Invertible Neural Networks (INN).

2.2.1 Frequency decomposition and invertible embedding

The detailed architecture of the watermark injection network adopted in this study is illustrated in Figure 4. To ensure signal imperceptibility, we transform the original image into the frequency domain via DWT rather than processing it in the pixel space. The transformed feature maps are combined with the watermark and then intricately fused as they pass through INN blocks.

In particular, a Channel Shuffling module placed at the end of each block randomly mixes information. This prevents the loss of specific channels from leading to the collapse of the entire signal, serving as a mechanism to build resistance against model compression attacks such as LoRA.

The invertible blocks employ an Affine Coupling layer to transform inputs without information loss and utilize RRDB (Residual in Residual Dense Block) as a sub-network to precisely learn complex high-frequency patterns. Here, the functions (ρ, ϕ, η, ψ) represent RRDB-based sub-networks.

$$y_1 = x_1 \odot exp(\phi(x_2)) + p(x_2), \quad y_2 = x_2 \odot exp(\eta(y_1)) + \psi(y_1) \tag{1}$$

2.2.2 Channel shuffling

Specifically, to counteract model compression attacks such as LoRA, we introduced a Channel Shuffling operation at the end of each block. By

reversing the channel order of the tensor and evenly distributing information, this mechanism ensures that the overall watermark signal remains intact even if information in specific channels is lost.

2.3 Low-Rank Adaptation and Signal Transfer

The attacker fine-tunes the base model (SDXL) using the dataset X'. Although the attacker attempts to compress the model using LoRA, the previously distributed watermark signal leaves traces in the model's weight matrix ($W' = W_{base} + BA$) due to High-frequency Inductive Bias. The generated image x_{gen} is modeled as follows:

$$x_{gen} \approx x_{content} + \in \cdot w_{gt} + n_{gen} \tag{2}$$

2.4 Statistical Signal Recovery and Verification

To recover minute residual signals, we designed a three-stage pipeline consisting of SRA, HiNet, and Ensemble.

2.4.1 Residual learning-based signal amplification

The generated image w_{gen} contains visual content ($x_{content}$) that is significantly stronger than the watermark, thereby hindering detection. To effectively recover these minute watermark signals, we propose the SRA. Our choice of a U-Net architecture combined with ResBlocks and Global Residual Learning, rather than a simple feed-forward network, is mathematically justified as follows.

The generated image x_{gen} can be expressed as a linear combination of the high-energy content signal x_c and the extremely weak watermark signal $w_{\in}$.

$$x_{gen} = x_c + \lambda \cdot w_{\epsilon} \quad (where\ \|x_c\| \gg \|\lambda \cdot w_{\epsilon}\|) \tag{3}$$

If a general CNN $F(\cdot)$ is used to directly extract the watermark, the gradient of the loss function $\mathcal{L}$ becomes dominated by the prevailing content signal x_c. This leads to a Signal Vanishing problem, where the model treats the watermark $w_{\in}$ as meaningless high-frequency noise and filters it out.

To address this issue, we introduced Global Residual Learning into the SRA. The SRA is defined as the sum of an identity term, which passes the input image unchanged, and a term $R(\cdot)$ that estimates the residual.

$$x_{enhanced} = SRA(x_{gen}) = x_{gen} + R(x_{gen}) \tag{4}$$

Here, the target that network R must learn is not the entire image, but solely the amplification of the watermark signal. During backpropagation, the content component x_c is canceled out via the identity function. Consequently, the network can be optimized by focusing exclusively on the patterns of the minute signal $w_{\in}$.

$$\frac{\partial \mathcal{L}}{\partial \mathcal{R}} \propto \nabla(h_l, W_l) \tag{5}$$

Furthermore, to prevent signal attenuation in deep networks, we applied Local Residual Blocks. For the input h_l of the l-th layer, the output h_{l+1} is defined as follows:

$$h_{l+1} = h_l + \mathcal{F}(h_l, W_l) \tag{6}$$

By initializing the weights W_l with small values (scale $= 0.1$) during the initial training phase, we ensure that $\mathcal{F}(h_l) \approx 0$, and consequently, $h_{l+1} \approx h_l$. This ensures that the watermark signal, weakened by LoRA, is not distorted or lost due to non-linear transformations as it traverses deep layers, but is instead propagated intact to the final layer via the Gradient Highway.

In conclusion, this Dual Residual Structure of the SRA functions as a filter that effectively bypasses the content signal x_c and selectively amplifies only the watermark signal $w_{\in}$, even in scenarios where the frequency bands of the content and watermark are intermixed.

2.4.2 Watermark extraction

The amplified image $x_{enhanced}$ undergoes the Reverse Pass of the previously described HiNet. The pre-trained invertible blocks reassemble the distributed signals back into a single, intact watermark pattern $\hat{w}$ through reverse operations.

2.4.3 Ensemble averaging

To remove the generation noise n_{gen} still remaining in the single extracted signal $\widehat{w_i}$, ensemble averaging is performed. By averaging the signals extracted from N generated images, random noise converges to zero, and only the consistent watermark pattern is amplified and revealed.

$$\hat{w} = \frac{1}{N} \sum_{i=1}^{N} HiNet_{rev}(SRA(x_{gen}^{(i)})) \approx \in \cdot w_{gt} \tag{7}$$

The similarity between the final recovered signal $\overline{w}$ and the original w_{gt} is measured via NPR (Negative Pixel Ratio). Since NPR evaluates the

consistency of the sign rather than the magnitude of the signal, it provides robust detection capabilities even when the signal is attenuated in low-rank environments.

3 Methodology

3.1 Training the Watermark Injection Model

To generate the experimental dataset, we utilized the COCO 2017 Unlabeled dataset. From a total of 123,403 images, we randomly selected 10,000 samples and preprocessed them to the 1024×1024 resolution recommended by the SDXL model. Subsequently, we pre-trained the injection network and the SRA. The training process was structured in two phases, incorporating Differentiable Proxy Attacks to simulate the actual LoRA compression environment. Since LoRA updates weights in the form of $W + \Delta W$, the NPR-enhanced watermark pattern acts as a learning bias for ΔW, ensuring that the watermark pattern is preferentially learned even if the model fails to capture the semantic style.

3.1.1 Model architecture and settings

The HiNet for watermark embedding is constructed with 8 Invertible Blocks, where each block incorporates an RRDB sub-network and a Channel Shuffling module. The SRA, designed for signal amplification, employs a U-Net architecture that performs Residual Learning. It features a 3-stage ResBlock encoder-decoder to facilitate deep feature extraction.

3.1.2 Two-phase training strategy

Phase 1 (HiNet Training): first phase in Figure 5, we trained the HiNet independently, excluding the SRA. The objective is to take an original image

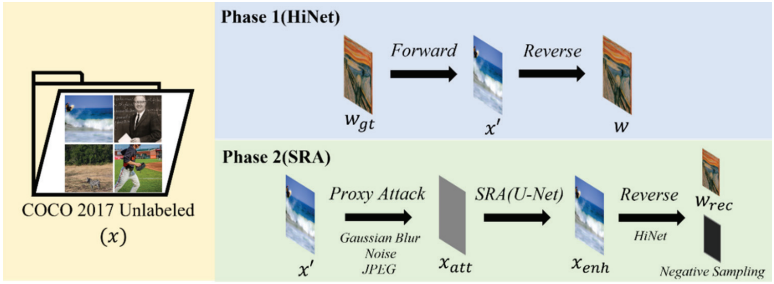

Figure 5 Two-phase training strategy.

x and a fixed watermark w_{gt} as inputs to generate a contaminated image x' – which exhibits no perceptible visual differences – and subsequently restore it back to w. For the loss function, we employed a combination of Perceptual Loss ($\mathcal{L}_{percep}$, based on VGG19) to maintain image quality and pixel-wise MSE Loss ($\mathcal{L}_{mse}$).

$$\mathcal{L}_{Phase1} = \lambda_1 \mathcal{L}_{mse}(x, x') + \lambda_2 \mathcal{L}_{percep}(x, x') + \lambda_3 \mathcal{L}_{mse}(w_{gt}, \hat{w}) \quad (8)$$

However, standard pixel-wise loss functions (MSE) treat all image regions uniformly. Consequently, they tend to embed watermark signals unnecessarily even in low-frequency (flat) regions, where information is prone to loss during the LoRA training process. To address this limitation, we introduced the NPR map, which adaptively regulates the watermark strength based on the local texture complexity of the image.

Specifically, the texture complexity $\mathcal{M}_{npr}$ at position (i, j) of the input image x is defined as the average difference between adjacent pixels.

$$\mathcal{M}_{npr}(x)_{i,j} = \frac{1}{4} \sum_{(m,n)\in\Omega_{i,j}} |x_{i,j} - x_{m,n}| \quad (9)$$

Here, $\Omega_{i,j}$ denotes the set of 4-connected neighboring pixels. This $\mathcal{M}_{npr}$ map yields high values in regions rich in edges or textures, while converging to zero in flat areas. Based on this property, we generated a weighting mask for watermark injection and designed the NPR loss function (L_{npr}) as follows:

$$L_{npr} = \left| (x - x_{stego}) \odot \frac{1}{1 + \alpha \cdot \mathcal{M}_{npr}(x)} \right|_2^2 \quad (10)$$

In this equation, x_{stego} represents the watermarked image, and α denotes the texture sensitivity constant. According to this formula, the loss penalty is relaxed in regions with high texture complexity (where the denominator increases), thereby allowing for stronger signal insertion; conversely, signal modifications are strictly suppressed in flat regions. Consequently, this strategy functions as a compass, guiding watermark signals into textured regions where they have a higher probability of being transferred to the LoRA weights.

Phase 2 (SRA Training and Robustness): In the second phase, with the weights of the trained HiNet encoder (H_{enc}) and decoder (H_{dec}) frozen, only the SRA (S_θ) is trained. The objective of this phase is to actively recover the

signal, ensuring that the watermark survives even in environments subject to attacks similar to LoRA fine-tuning. We constructed the training batches by mixing positive samples (containing watermarks) and negative samples (without watermarks). The total loss function L_{Phase2} is defined as follows:

$$L_{Phase2} = E_{pos}[L_{rec}] + \lambda_{neg} \cdot E_{neg}[L_{silent}] \tag{11}$$

Here, the first term L_{rec} guides the SRA to accurately extract the ground truth watermark w by correcting the image $x_{stego} = H_{enc}(x, w)$ that has been corrupted after passing through the proxy attack module $\mathcal{A}$.

$$L_{rec} = |w - H_{dec}(S_\theta(\mathcal{A}(x_{stego}))|_2^2 \tag{12}$$

The second term, L_{silent}, serves as a Silence Loss that suppresses the generation of false signals when a clean, non-watermarked image x_{clean} is provided as input. Its primary purpose is to prevent hallucination, a phenomenon where the SRA misinterprets random noise as a valid signal and erroneously amplifies it.

$$L_{silent} = |0 - H_{dec}(S_\theta(\mathcal{A}(x_{clean})))|_2^2 \tag{13}$$

In the above equation, 0 denotes a Zero Tensor consisting entirely of zero elements, and λ_{neg} represents the weight assigned to negative sampling. Through this training strategy, we optimized the SRA to effectively compensate for high-frequency losses induced by LoRA compression, while simultaneously equipping it with the discriminative capability to remain unresponsive to non-watermarked images.

3.1.3 Proxy attack simulation

Since the LoRA training process is computationally intensive and difficult to differentiate, we simulated the information compression effects of LoRA within the training loop by randomly combining Gaussian Blur ($\sigma \in [0.1, 2.0]$), Differentiable JPEG approximation, and Gaussian Noise. Consequently, the model was trained to generate robust signals capable of withstanding the LoRA compression environment.

3.1.4 Model training process

To maintain a low False Positive Rate (FPR) in real-world verification scenarios, we introduced a Random Negative Sampling strategy during the SRA training phase (Phase 2). Specifically, 20% of the training batch consisted of

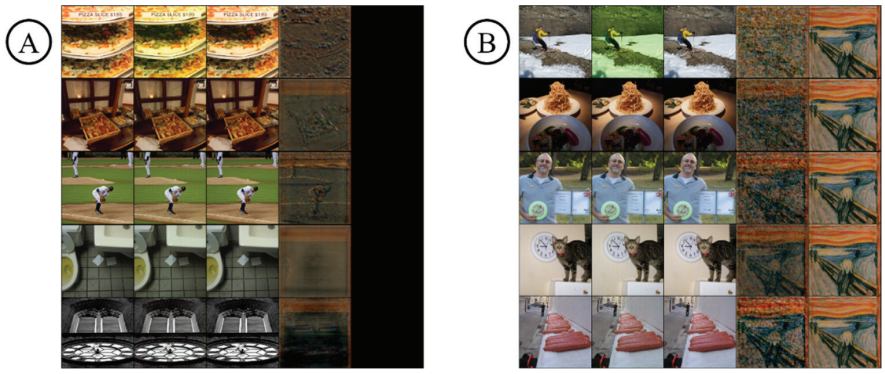

Figure 6 Model training process.

clean images without watermarks (Negative Samples), while the remaining 80% comprised watermarked images (Positive Samples).

The sequence of images in Figure 6 is displayed as follows: Input → SRA Result → Target → Recovered Watermark → Ground Truth Watermark. In case A (Negative Sample), the ground truth appears black because no watermark is present; in case B (Positive Sample), the correct ground truth watermark is displayed. The purpose of this imbalanced training allocation is to suppress Signal Hallucination in the model. If the model were trained solely on data containing watermarks, it would develop a tendency towards Overfitting, misinterpreting input noise as watermark patterns and attempting to forcibly extract them. By forcing the model to output a Zero Tensor for negative samples, we enhanced its discriminative capability to clearly distinguish between random noise and meaningful signals.

3.1.5 Quantitative performance of the trained model

To validate the quality of the finalized watermark model intended for the LoRA experiments in Section 4.2, we evaluated its invisibility and extraction accuracy using a test dataset of 1000 images. The results showed (Table 1) an average PSNR (Peak Signal to Noise Ratio) of 40.1081 dB and an SSIM of 0.9667 between the original image x and the watermarked image x'.

These figures exceed the 40 dB threshold, which is generally considered the standard for visual imperceptibility, thereby demonstrating that our watermark injection model does not compromise the visual quality of the images. This quantitative assessment ensures that our watermarking module is capable of generating high-quality poisoned data at a level

Table 1 Quantitative performance of the trained model

	PSNR	SSIM
Tree-Ring	31.85	0.90
HiDDeN	31.50	0.86
SpecGuard [7]	40.86	0.99
DIAGNOSIS	32.21	0.96
Coprguard	39.90	0.98
Suggest Model	40.10	0.96

Table 2 Environment used in the experiment

OS	Windows 11 24H2
CPU	AMD Ryzen 7 9800X3D
RAM	64GB
VGA	NVIDIA RTX 5090
Driver Version	576.88
CUDA Version	12.9
cuDNN	9.16.0
Python	3.10

imperceptible to attackers, effectively underpinning the reliability of the subsequent experiments.

4 Experimental Results

4.1 Experiments and Verification

To validate the efficacy of our proposed methodology in detecting copyright infringement within real-world creative environments, we designed a Style Mimicry scenario. The experimental environment is detailed in Table 2.

4.1.1 LoRA dataset preparation and poisoning

We selected the artworks of Vincent van Gogh as the subject of our experiment, as his unique artistic style renders his work a primary target for style mimicry. We curated 20 high-resolution images from his masterpieces to construct the training dataset X_{style}. This selection reflects the Few-shot Learning characteristic of LoRA, which enables high-quality style transfer using only a small number of data samples. Utilizing the watermark model (HiNet) prepared in Section 3.1, we invisibly embedded a fixed watermark

pattern w_{gt} into the original images. The resulting poisoned dataset X'_{style} is visually indistinguishable from the original (PSNR $>$ 40 dB); thus, we assume that the attacker unknowingly utilizes it for training.

4.1.2 LoRA training settings

The attacker fine-tunes the base model, SDXL 1.0, using the poisoned dataset X'_{style}. To ensure high alignment with real-world scenarios, we employed kohya_ss v25.2.1, the most widely used LoRA training script in the open-source community. To verify whether the watermark persists within the model capacity range typically adopted by actual users, we set the Rank to three distinct values: 32, 64, and 128. Ranks below R $=$ 32 lack style expressiveness, while those exceeding R $=$ 128 suffer from diminished efficiency; thus, this range represents the most common and realistic threat interval. The goal of this experiment is to detect the watermark signal internalized by the model itself, without relying on specific trigger words. Therefore, training was conducted for 1000 steps to induce the signal to propagate deeply into the model weights. The learning rate was set to 0.0001, and the bf16 data format was utilized. The training proceeded for a total of 50 epochs, configured to save safetensors files every 5 epochs.

4.1.3 Generation of verification data

We generated a large-scale verification dataset to conduct watermark detection experiments on each of the trained LoRA models (Ranks 32, 64, and 128). In this experiment, we took into account that real-world users variably adjust the LoRA application weight (Strength) to control the stylistic intensity of the generated results. Accordingly, we generated images by subdividing the LoRA Strength from 0.2 to 1.0 in increments of 0.2 for each Rank condition. This approach allows for a comprehensive verification of signal Robustness specifically, whether the watermark signal is fully preserved and transferred to the model weights even in low-strength intervals where LoRA intervention is minimal.

Figure 7 visualizes the image samples generated under these various Rank and Strength conditions. To demonstrate that the model does not merely output signals overfitted to specific prompts, we prepared 100 Random Prompts that are unrelated to the Van Gogh style. By generating five images per prompt using different seeds, we secured a total of 500 images per model, as shown in Figure 7. This provides a sufficient statistical sample for the ensemble analysis (N $= 1 \sim 100$) to be conducted in Section 4.2.

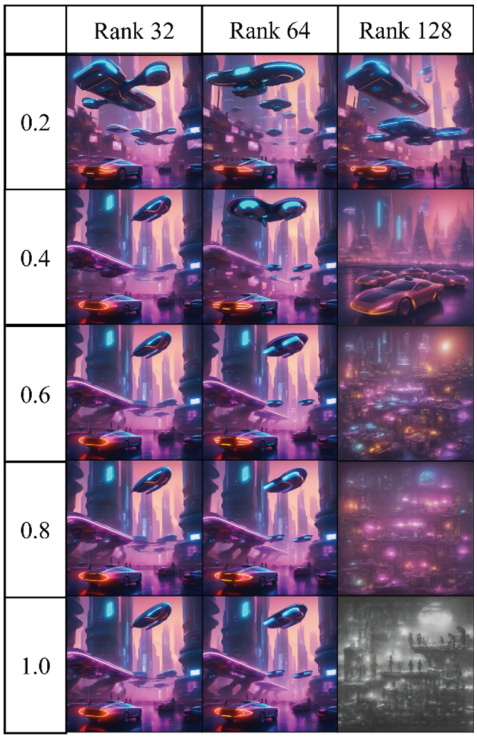

Figure 7 SDXL + LoRA created images.

Table 3 Ensemble-based signal amplification analysis based on the number of images

	N=1	N=5	N=10	N=20	N=50	N=100
Clean	0.480	0.492	0.591	0.512	0.536	0.493
Rank 32	0.533	0.598	0.639	0.627	0.796	0.891
Rank 64	0.519	0.620	0.680	0.794	0.913	0.958
Rank 128	0.464	0.734	0.835	0.910	0.982	0.999

4.2 Analysis of Ensemble-based Signal Amplification

Based on the total of 2000 generated images (500 for the Clean model and 500 for each Rank) secured in Section 4.1, we conducted an in-depth analysis of how variations in ensemble size (N) and LoRA capacity (Rank) impact watermark detection performance.

First, when attempting watermark extraction using only a single image ($N = 1$), the AUROC (Area Under the Receiver Operating Characteristic Curve) hovered around the random guessing level of 0.5 across all

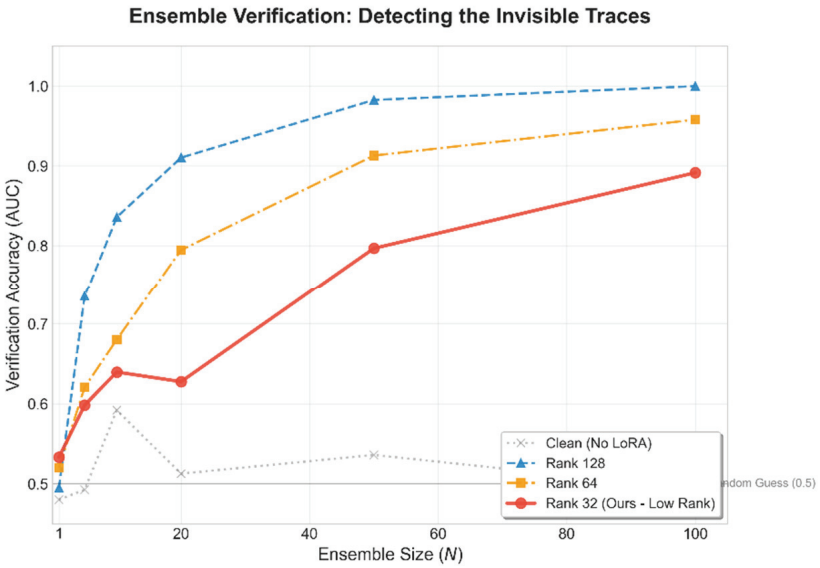

Figure 8 Ensemble verification: detecting the invisible traces.

experimental conditions, as presented in Table 3. This trend was consistent for the Clean model (0.480) as well as for Rank 32 (0.533), Rank 64 (0.519), and Rank 128 (0.464). This is because the high-variance stochastic noise, which inevitably occurs during the generation process of diffusion models, overwhelms the subtle watermark signal. Paradoxically, this low single-image detection rate suggests that our watermarking technique secures high invisibility, making it indiscernible to the naked eye or simple filtering-based attacks.

However, as the ensemble size N increased, the latent signal was distinctly amplified through the phenomenon of Statistical Resonance, as illustrated in Figure 8. Particularly in the case of Rank 128, which possesses sufficient model capacity, the performance showed a steep upward trend. It recorded an AUROC of 0.910 at $N = 20$, approaching the threshold of 0.95, reached 0.982 at $N = 50$, and finally achieved a detection performance of 0.999 at $N = 100$. Rank 64, representing an intermediate capacity, exhibited a similar trend, recording a high accuracy of 0.958 at $N = 100$. This is attributed to the Central Limit Theorem: during the ensemble process, random generation noise cancels out, reducing variance. In contrast, the watermark signal, which is fixedly internalized within the model weights, accumulates consistently, thereby maximizing the Signal-to-Noise Ratio (SNR).

Of particular note are the results under the Rank 32 condition, where the model's expressiveness is limited. While Rank 32 showed a somewhat gradual rise to 0.627 up to N = 20, performance improved significantly to 0.891 at N = 100, where a sufficient ensemble size was secured. This indicates that it has achieved a level of detection capability usable as meaningful evidence. This demonstrates that even with reduced model capacity, the watermark (composed of high-frequency components) is preferentially transferred to the model weights. Conversely, for the Clean model (without watermark insertion), the AUROC fluctuated irregularly between 0.493 and 0.591 even as N increased to 100, showing no specific upward trend. This confirms the high reliability of the proposed method, ensuring it does not generate False Positives for standard generated images lacking watermarks.

4.3 Analysis of Visual Quality and Stability of Generated Images

Just as crucial as the robustness of the watermark is the utility of the generative model. To analyze the impact of the poisoned dataset X' on image quality during actual LoRA training, we measured FID (Fréchet Inception Distance) scores under various conditions of Rank ($\mathrm{r} \in \{32, 64, 128\}$) and application strength ($\{\mathrm{Strength}\} \in [0.2, 1.0]$).

As demonstrated in Table 4, under the Rank 32 and Rank 64 conditions, FID scores exhibited a downward stabilization trend – decreasing from the low 60s to the 48∼50 range – as the LoRA Strength increased from 0.2 to 1.0. This indicates that the model is successfully capturing Van Gogh's artistic style without significant interference from the watermark noise. Notably, under the most common configuration of Rank 64 and Strength 1.0, the model recorded an FID of 50.27. This represents respectable generation quality within a Few-shot Learning environment, and no artifacts or distortions were observed during Qualitative Inspection.

In contrast, for Rank 128 – the configuration with the largest model capacity – we observed severe Overfitting, characterized by a drastic surge in the FID score to 288.10 as the Strength increased. This is a typical phenomenon associated with fine-tuning high-capacity models on a small-scale dataset (only 20 images); it stems from data scarcity rather than the

Table 4 FID scores based on various ranks and strengths

Strength	Rank 32	Rank 64	Rank 128
0.2	62.32	65.84	70.99
1.0	48.61	50.27	288.09

presence or absence of the watermark. These results imply that to obtain high-quality images, an attacker is effectively compelled to prefer Low-Rank settings, such as Rank 32 or 64. Therefore, the high watermark detection rates in Low-Rank intervals (Rank 32, 64), as demonstrated in the previous analysis in Section 4.2, support the conclusion that our method serves as a highly effective defense measure in realistic threat scenarios.

4.4 Comparison and Analysis with Existing Watermarking Techniques

To demonstrate the robustness of the proposed SRA strategy, we conducted a comparative experiment on survivability against the DWT-DCT method, a representative frequency-domain watermarking technique [1]. To ensure a fair comparison, we embedded DWT-DCT watermarks into the identical Van Gogh dataset (20 images) at a level where imperceptibility is maintained (alpha = 3.0). Subsequently, we trained the LoRA model under the same conditions used in our study (Rank 64). Furthermore, to simulate a real-world online distribution environment, we measured the detection rate after applying standard JPEG compression with a quality factor of 80 to the generated images.

Table 5 presents the results. Our SRA-based technique maintained a detection performance of nearly 95% even under the harsh conditions where LoRA model compression and JPEG lossy compression are applied simultaneously. In contrast, the detection accuracy of the DWT-DCT method plummeted to 40.0%, a figure lower than that of random guessing. The existing State-of-the-Art (SOTA) technique, Tree-Ring, relies on specific phases in the frequency domain; consequently, it is structurally vulnerable to phase shifts that occur during the LoRA reconstruction process.

This performance disparity stems from fundamental differences in the signal storage domains and optimization methodologies. Traditional techniques, such as DWT-DCT, uniformly embed information into Fixed High-frequency Bands. However, during LoRA's Low-Rank Reconstruction process, these signals are often treated as "meaningless noise," making them prone to being filtered out or overwritten by new stylistic features. In contrast,

Table 5 Comparison between DWT-DCT and the proposed technique (SRA)

	Target Model	Post-Processing	Detection Accuracy
DWT-DCT	LoRA (Rank 64)	JPEG	40.0%
SRA	LoRA (Rank 64)	JPEG	95.8%

our proposed method employs a standard U-Net architecture while guiding the training process based on the NPR metric. This approach allows for the adaptive internalization of signals into High Entropy Regions (areas with high texture complexity). Consequently, we achieved robust resistance against LoRA training and compression without requiring any modifications to the underlying model architecture.

5 Conclusion

This study proposed and validated an Adaptation-Agnostic Trace Verification method to technically resolve the copyright infringement issues arising from the Low-Rank Adaptation (LoRA) of diffusion models. To overcome the limitations of existing frequency-based watermarking – which is prone to being filtered out or lost during LoRA's compression and reconstruction processes – we introduced a two-phase training strategy. This approach adaptively internalizes signals into regions with high texture complexity, utilizing NPR as a basis. Through this, we confirmed the structural feasibility of transferring watermark signals into model weights, even in environments where the model's training capacity is extremely limited.

Experimental results indicated that while it was difficult to determine the presence of a watermark using single-image analysis due to stochastic noise in the generation process, we were able to secure a reliable level of detection performance by applying the Statistical Resonance technique established in this study. In particular, by canceling out uncorrelated noise through the ensemble of multiple generated images, we achieved significantly higher survivability and detection accuracy compared to the existing DWT-DCT method, even under underfitting conditions (Rank 32) and JPEG lossy compression. This empirically demonstrates that, independent of the visual completeness of style mimicry, the traces of training data are preserved within the model and subsequently transferred to the generated outputs.

This study holds significance in that it proposes a methodology capable of tracing training usage solely through generated images within a black-box environment, without requiring access to the target model's parameters. Future research should focus on advancing the model by enhancing its robustness against geometric transformation attacks, such as rotation and cropping, and by improving efficiency through a reduction in the number of samples required for verification. Furthermore, by exploring the potential for expansion beyond the image domain into video or audio generation models, this approach can evolve into a comprehensive

copyright protection technology capable of addressing multimodal generative environments.

Acknowledgement

This work was supported by the Software Copyright Research and Development Program funded by the Ministry of Culture, Sports and Tourism and managed by the Korea Institute of Culture Technology Evaluation and Planning (KCTEP) (Project Name: Development of Copyright Technology for OTT Contents Copyright Protection Technology Development and Application, Project Number: RS-2023-00225267, Contribution Rate: 100%).

References

[1] Liu, Zhenguang, et al. "Harnessing frequency spectrum insights for image copyright protection against diffusion models." *Proceedings of the Computer Vision and Pattern Recognition Conference* 2025.

[2] Lin, Dongdong, et al. "An efficient watermarking method for latent diffusion models via low-rank adaptation." *arXiv preprint arXiv:2410.20202* 2024.

[3] Cui, Yingqian, et al. "Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models." *ACM SIGKDD Explorations Newsletter* 26.2: 76–88, 2025.

[4] Cao, Huangsen, et al. "HyperDet: Generalizable detection of synthesized images by generating and merging a mixture of hyper LoRAs." *arXiv preprint arXiv:2410.06044* 2024.

[5] Wang, Zilan, et al. "Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models." *Proceedings of the Computer Vision and Pattern Recognition Conference* 2025.

[6] Lv, Peizhuo, et al. "LoRAGuard: An effective black-box watermarking approach for LoRAs." *arXiv preprint arXiv:2501.15478* 2025.

[7] Alam, Inzamamul, et al. "SpecGuard: Spectral projection-based advanced invisible watermarking." *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2025.

[8] Lee Dae-hee, AI Copyright Lawsuit Analysis: Sarah Andersen v. Stability AI, Korea Copyright Commission, Issue Report 2013-15.

[9] Zhang, Xuanyu, et al. "Editguard: Versatile image watermarking for tamper localization and copyright protection." *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024.

[10] Zhong, Haonan, et al. "Copyright protection and accountability of generative ai: Attack, watermarking and attribution." *Companion Proceedings of the ACM Web Conference* 2023.

[11] Nguyen, Phuc, et al. "Image copyright protection: A comprehensive survey of digital watermarking, deep learning, and blockchain approaches." *IEEE Open Journal of the Computer Society* 7:244–263, 2026.

[12] Jiang, Zhengyuan, et al. "Watermark-based attribution of AI-generated content." *arXiv preprint arXiv:2404.04254* 2024.

[13] Zhao, Xuandong, et al. "Invisible image watermarks are provably removable using generative AI." *Advances in Neural Information Processing Systems* 37:8643–8672, 2024.

[14] Zhang, Lijun, et al. "Robust image watermarking using stable diffusion." *arXiv preprint arXiv:2401.04247* 2024.

[15] Zhang, Lijun, et al. "Attack-resilient image watermarking using stable diffusion." *Advances in Neural Information Processing Systems* 37:38480–38507, 2024.

[16] Yang, Zijin, et al. "Gaussian shading: Provable performance-lossless image watermarking for diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2024.

[17] Fernandez, Pierre, et al. "The stable signature: Rooting watermarks in latent diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2023.

Biographies

Jinseok Kim received his bachelor's degree in 2017 and his master's degree in computer engineering from Soongsil University, Republic of Korea, in

2019. He worked at ICTWay Co. Ltd. from 2019 to 2022 and is currently pursuing a Ph.D. in computer science and engineering at Soongsil University. His research interests include networks, network security, artificial intelligence, and DRM.

Uijin Jang received her Ph.D. in Computer Science and Engineering from Soongsil University, Republic of Korea, in 2010. Since 2018, she has worked at the Spartan SW Education Center at Soongsil University. Her research focuses on AI-based copyright technologies and their practical applications.

Yongtae Shin received his Ph.D. in Computer Science from the University of Iowa, USA, in 1994. Since 1995, he has been serving as a Professor in the School of Computer Science and Engineering at Soongsil University, Republic of Korea. His research interests include computer networks, distributed computing, Internet protocols, and e-commerce technology.