# ARE WATERMARKS FOR DIFFUSION MODELS RADIOACTIVE?

**Jan Dubiński** *
Warsaw University of Technology, IDEAS NCBR
jan.dubinski.dokt@pw.edu.pl

**Michel Meintz** *
CISPA Helmholtz Center for Information Security
michel.meintz@cispa.de

**Franziska Boenisch**
CISPA Helmholtz Center for Information Security
boenisch@cispa.de

**Adam Dziedzic**
CISPA Helmholtz Center for Information Security
adam.dziedzic@cispa.de

## ABSTRACT

As generative artificial intelligence (AI) models become increasingly widespread, ensuring transparency and provenance in AI-generated content has become a critical challenge. Watermarking techniques have been proposed to embed imperceptible yet detectable signals in AI-generated images, enabling provenance tracking and copyright enforcement. However, a second party can repurpose images generated by an existing model to train their own diffusion model, potentially disregarding the ownership rights of the original model creator. Recent research in language models has explored the concept of watermark *radioactivity*, where embedded signals persist when training or fine-tuning a new model, enabling the detection of models trained on watermarked data. In this work, we investigate whether similar persistence occurs in diffusion models. Our findings reveal that none of the tested watermarking methods transfer their signal when used for fine-tuning a second model. This means that images generated by this new model exhibit detection results for the watermarks of the original model indistinguishable from random guessing. These results indicate that existing techniques are insufficient for ensuring watermark propagation through the model derivation chain and that novel approaches are needed to achieve effective and resilient watermark transfer in diffusion models.

## 1 INTRODUCTION

As AI-generated content becomes increasingly realistic, concerns over AI-driven disinformation and copyright infringement continue to grow. High-fidelity generative models, such as Stable Diffusion (Rombach et al., 2022), have made synthetic media creation more accessible, raising critical issues regarding misuse, ownership, and content authenticity (Crothers, 2022). In response, regulatory efforts have emphasized the need for greater transparency, urging AI developers to implement watermarking techniques for content attribution and provenance tracking (Biden, 2023; California State Legislature, 2024; European Union, 2024).

To address these concerns, various watermarking techniques have been developed to embed identifiable signals within AI-generated images (Fernandez et al., 2023; Wen et al., 2023; Gunn et al., 2024; Yang et al., 2024). These watermarks serve as a means of attributing images to a specific diffusion model. However, an unresolved challenge is whether these watermarks persist when AI-generated images are repurposed to train a new diffusion model. If the watermark does not transfer, it exposes a gap in current techniques, rendering them less effective in preventing model misuse and protecting ownership rights.

In the context of Large Language Models (Sander et al., 2024) explores the *radioactivity* of watermarked text—its ability to leave detectable traces when used to fine-tune another model. Their findings show that even when the watermark is not directly visible, its residual signal can persist in a fine-tuned model's outputs. The strength of this effect depends on the watermark's robustness, its prevalence in the training data, and the fine-tuning process. This unintended consequence makes it

possible to identify models trained on watermarked content, helping to trace data provenance, detect unauthorized use, and evaluate the influence of synthetic data on model behavior.

Inspired by these findings, we examine whether images generated by watermarked diffusion models transfer detectable traits when used to fine-tune subsequent generative models. We find that none of the tested watermarking techniques transfer in the case of latent diffusion models. Although certain watermarking methods, such as *Stable Signature* (Fernandez et al., 2023), *Tree-Ring* (Wen et al., 2023), and *PRC* (Gunn et al., 2024) watermarking, enable perfect watermark detection in images generated by the original model, applying the detection algorithm to images generated by a fine-tuned model results in detection performance no better than random guessing.

Our findings indicate that the inability of watermarking methods to persist through fine-tuning in diffusion models is closely related to their vulnerability to visual paraphrasing attacks Barman et al. (2024). Given these limitations, we conclude that developing transferable watermarking techniques for diffusion models requires new methodologies that ensure robust watermark propagation throughout the model derivation chain.

## 2 BACKGROUND

In text-to-image models, watermarks are designed to be imperceptible to the human eye yet detectable by a specialized detection algorithm. Image watermarks must possess several essential properties. *Undetectability* ensures that the watermark remains imperceptible to unauthorized parties. *Unforgeability* guarantees that an adversary cannot reproduce the same watermark. *Robustness* ensures that the watermark remains detectable even under adversarial attacks. In this work, we consider the following watermarking methods:

**A Recipe for Watermarking Diffusion Models.** Zhao et al. (2023) propose a method that utilizes an pretrained encoder-decoder structure to embed a predefined binary signature into the training data of an Elucidated Diffusion Model (EDM) (Karras et al., 2022). The diffusion model is then trained on this watermarked dataset, causing it to reproduce the watermark in its generated images. A watermark decoder is subsequently used to extract and verify the binary signature from these outputs. We refer to this approach as the *Recipe* method.

**Stable Signature.** The *Stable Signature* method (Fernandez et al., 2023) introduces a watermarking technique for Latent Diffusion Models (LDMs) (Rombach et al., 2022), leveraging the model's structure. The method fine-tunes the latent decoder to embed an invisible binary signature in every generated image. This signature can later be recovered using a pretrained watermark extractor.

**Tree-Ring.** Wen et al. (2023) propose embedding a structured pattern into the *initial noise vector* used in diffusion model sampling. By structuring these patterns in the Fourier space, the method achieves a high level of robustness against common image transformations, such as cropping, dilation, and rotation. Unlike the *Stable Signature* approach, which modifies the decoder, *Tree-Ring Watermarking* influences the sampling process itself, allowing for watermark detection by inverting the diffusion process and analyzing the retrieved noise vector.

**PRC Watermarking: P**seudo**R**andom **C**ode (PRC) watermarking (Gunn et al., 2024) utilizes the random generation of an *initial noise vector*. This noise vector is sampled based on a pseudorandom error-correcting code (Christ & Gunn, 2024), enabling watermark detection. Given an image, the original noise vector can be reconstructed, allowing the identification of the specific random code used to generate the image.

## 3 PROBLEM FORMULATION

Consider a diffusion model $DM1$, trained on a dataset $X_0$, that has been *watermarked* using a method $w$. This model is then used to generate a collection of images $X_1$, which inherently carry the watermark *i.e.,*, an appropriate detection algorithm, specific to $w$, can successfully identify the watermark signal in $X_1$. Next, $X_1$ is used as a training dataset for a second diffusion model, $DM2$. Our goal is to determine whether the watermark signal persists in the images $X_2$ generated by $DM2$. If the watermark remains detectable in $X_2$, we consider the watermarking method $w$ to be *radioactive*, indicating that the watermark transfers through generated data.

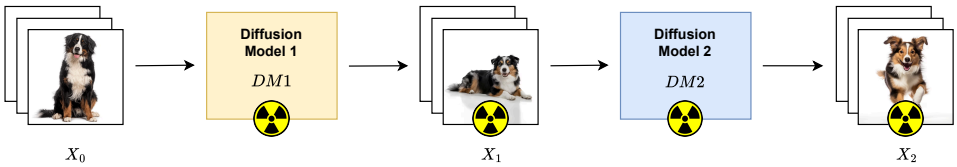



Figure 1: **Radioactivity in diffusion models.**

## 4 Experimental Evaluation

**Experimental Setup.** We evaluate the transferability of watermarks across two distinct DM types: one operating in image space and the other in latent space.

For our experiments on the image space DM, we employ an EDM model architecture (Karras et al., 2022) and the CIFAR-10 dataset (Krizhevsky, 2009). We investigate the *Recipe* framework as proposed by Zhao et al. (2023). The first model, referred to as *DM1*, is trained unconditionally on the watermarked images $X_0$. In accordance with the method, we embed the watermark directly into the training data $X_0$ of DM1. We then generate 50,000 images—matching the number of training samples used for DM1—to train a second model, *DM2*, using the same hyperparameters.

For the assesment on LDMs, we consider Stable Diffusion 2.1 (Rombach et al., 2022) and the MS COCO 2014 dataset (Lin et al., 2014). For the *Recipe*, we adopt a procedure analogous to the one used for image-space models described above. In the case of *Stable Signature*, *Tree-Ring*, and *PRC* watermarks, fine-tuning DM1 on $X_0$ with an embedded watermark is unnecessary, as the watermark in those methods is directly embedded in DM1's output. Consequently, we only fine-tune DM2 on the dataset $X_1$ generated by DM1, simulating full training, following the MS COCO captions.

In the case of LDMs, we also examine an extreme scenario where $DM2$ (and $DM1$ in the case of *Recipe*) is fine-tuned on a single image for an exceptionally high (4000) number of epochs to enforce the transfer of a specific watermark instance.

### 4.1 Results

#### 4.1.1 Insights into Image Space Diffusion Models

**Data-level watermarks transfer in the image space.** We find that in image-space DMs, the watermarking method *Recipe* (Zhao et al., 2023) successfully transfers from the output of DM1 to DM2. This aligns with expectations: when a watermark is present in the training data and propagates to the output of DM1, it should implicitly transfer to the output of DM2 as well, given that both models perform the same task. The EDM architecture facilitates this transfer, as images are directly generated and learned in image space. This direct reconstruction of pixel values, without abstraction into the latent space, enhances the retention of watermarks.

Table 1: **We find no transferability of watermarks in LDMs.** We evaluate two distinct types of diffusion models: one operating in image space (EDM) and the other in latent space (LDM). For the *Recipe* method, we also assess watermark detection in the training data of DM1. We analyze four different watermarking methods using identical model fine-tuning hyperparameters. The bitwise accuracy of the reconstructed watermark is reported for *Recipe*, *Stable Signature*, and *PRC*. When the bitwise accuracy reaches $50\%$, the watermark is indistinguishable from random noise and cannot be reliably detected. Since *Tree-Ring* does not rely on a bitwise key, we instead report TPR@FPR=1%, which measures the true positive rate at a 1% false positive rate for detecting watermarked images. In this case, TPR=1% corresponds to random guessing.

| Type of $M_1$ | Type of $M_2$ | Method | Metric | Output of $M_1$ | Output of $M_2$ |
|---|---|---|---|---|---|
| EDM | EDM | Recipe | TPR@FPR=1% | 82.5 | 82.2 |
| LDM | EDM | Tree-Ring | TPR@FPR=1% | 100 | 80.39 |
| LDM | LDM | Recipe | BA | 49.9 | - |
| LDM | LDM | Stable Signature | BA | 100 | 49.9 |
| LDM | LDM | PRC | BA | 100 | 50.0 |
| LDM | LDM | Tree-Ring | TPR@FPR=1% | 100 | 1 |

#### 4.1.2 INSIGHTS INTO LATENT DIFFUSION MODELS

Table 2: **Extensive fine-tuning on a single watermarked sample does not enforce watermark transfer.** Given the observed lack of watermark transferability in LDMs, we extend our analysis to an extreme scenario aimed at transferring a specific watermark instance. We report bitwise accuracy for *Recipe*, *Stable Signature*, and *PRC*, as well as TPR@FPR=1% for *Tree-Ring*.

| Type of $M_1$ | Type of $M_2$ | Method | Metric | Output of $M_1$ | Output of $M_2$ |
|---|---|---|---|---|---|
| LDM | LDM | Recipe | BA | 53.2 | 51.5 |
| LDM | LDM | Stable Signature | BA | 100 | 44.7 |
| LDM | LDM | PRC | BA | 100 | 51.2 |
| LDM | LDM | Tree-Ring | TPR@FPR=1% | 100 | 1 |

**We find no transferability of watermarks in LDMs.** Table 1 highlights the contrasting behaviors of EDM and LDM architectures in watermark transferability. Given these findings, we also conduct an experiment to analyze an extreme scenario where LDMs are extensively fine-tuned on a single instance of a watermarked image in an attempt to force watermark transfer. Table 2 shows that even under such conditions, the watermark does not transfer through fine-tuning, resulting in random detection outcomes.

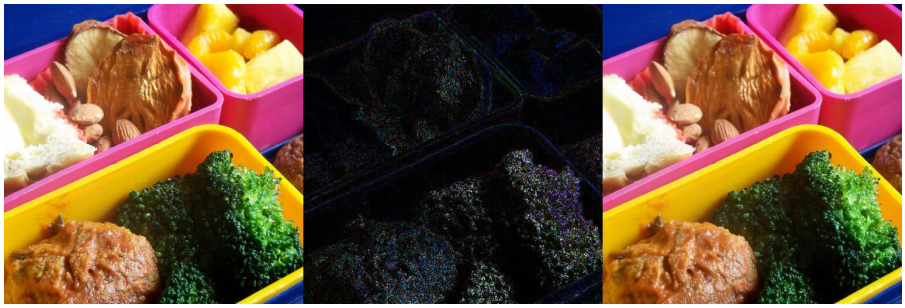


Figure 2: **The LDM autoencoder acts as a filter for the *Recipe* data-level watermark.** *Left*: original watermarked image with $100\%$ bitwise accuracy, *Middle*: difference $\times$ 2, *Right*: image after en-decoding by the SD2.1 autoencoder structure with $49.9\%$ bitwise watermark accuracy, *i.e.,* guessing.

**The autoencoder acts as a filter for data-level *Recipe* watermark in LDMs.** For the *Recipe* method, we observe that the watermark successfully transfers in EDMs but fails under full fine-tuning in the LDM architecture. We attribute this failure to the variational autoencoder (VAE) used in the Stable Diffusion 2.1 framework, which encodes images into latent space. We hypothesis that in the case of *Recipe* the autoencoder effectively acts as a filter, suppressing high-frequency details, including embedded watermarks. Figure 2 presents an example of a watermarked image used for fine-tuning SD2.1. We observe that, after passing through the VAE, the watermark's detection bitwise accuracy in the image drops to $49.9\%$.

**Watermarking methods explicitly designed for LDMs also show no signs of transferability.** The *Stable Signature*, *Tree-Ring*, and *PRC* watermarking techniques fail to transfer from DM1 to DM2. However, we find that for these methods, the watermark remains intact after being encoded into the latent space of SD2.1, suggesting that an alternative factor inhibits its transferability.

Figure 3 presents example images generated by DM1 and DM2 with the *Stable Signature* watermark. In Figure 3a, the watermark is clearly detectable, whereas in Figure 3b, it is no longer traceable. Additionally, the two images exhibit noticeable differences in composition and color. While the semantic structure remains consistent, the spatial layout differs significantly. This variation in image composition resembles a **visual paraphrasing attack** (Barman et al., 2024). Current watermarking techniques remain vulnerable to such transformations, highlighting a critical challenge in ensuring watermark robustness and transferability. Additional examples for other watermarking methods are presented in Figure 4 and Figure 5.

(a) DM1: bitwise accuracy 100%

(b) DM2: bitwise accuracy 49%

Figure 3: Comparison of the output of DM1 and DM2 and the *Stable Signature* watermark.

(a) DM1:bitwise accuracy 100%

(b) DM2:bitwise accuracy 51%

Figure 4: Comparison of the output of DM1 and DM2 and the *PRC* watermark.

## 5 Conclusion and Future Work

In this work, we assess the transferability of watermarks from one DM to another across four watermarking methods and two model architectures. Our results show that watermark transfer occurs only in one of the architectures—specifically, in image-space DMs. For the *Recipe* watermarking method, we attribute this behavior to the embedding process in the latent space of LDMs. Furthermore, we demonstrate that for watermarking methods designed for LDMs, fine-tuning a second DM on watermarked images behaves similarly to a visual paraphrasing attack, which inherently disrupts watermark integrity. These findings highlight the need for more robust watermarking approaches capable of withstanding such attacks.

An effective approach to transferable watermarking may involve embedding signals directly in the data space, ensuring they are strong enough to persist through the autoencoder, which naturally filters out high-frequency signals. Initial experiments with the method proposed by Wang et al. (2024), which detects unauthorized data usage by injecting imperceptible modifications into training data, show promising results: the signal successfully transfers through the fine-tuning process. However, this approach significantly degrades the generative performance of fine-tuned models, making it more suitable as a defensive mechanism against unauthorized data usage rather than a lossless watermarking solution. Future research could focus on designing watermarking methods that not only transfer through fine-tuning but also preserve the generative quality of DMs.

## ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) within the framework of the Weave Programme under the project titled "Protecting Creativity: On the Way to Safe Generative Models" with number 545047250. We also gratefully acknowledge support from the Initiative and Networking Fund of the Helmholtz Association in the framework of the Helmholtz AI project call under the name "PAFMIM", funding number ZT-I-PF-5-227. Responsibility for the content of this publication lies with the authors. This research was also supported by the Polish National Science Centre (NCN) within grant no. 2023/51/I/ST6/02854 and 2020/39/O/ST6/01478 and by Warsaw University of Technology within the Excellence Initiative Research University (IDUB) programme.

## REFERENCES

Niyar R Barman, Krish Sharma, Ashhar Aziz, Shashwat Bajpai, Shwetangshu Biswas, Vasu Sharma, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. The brittleness of ai-generated image watermarking techniques: Examining their robustness against visual paraphrasing attacks, 2024. URL https://arxiv.org/abs/2408.10446.

Joseph R. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, October 2023. Accessed: 2024-09-24.

California State Legislature. California assembly bill ab-3211 california digital content provenance standards, February 2024. URL https://legiscan.com/CA/text/AB3211/id/2984195. Accessed: 2024-09-24.

Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. *arXiv preprint arXiv:2402.09370*, 2024. URL https://arxiv.org/abs/2402.09370.

Brooke Crothers. Ai-generated deepfakes: A growing threat to democracy. *Forbes*, 2022. URL https://www.forbes.com/sites/brookecrothers/2022/10/15/ai-generated-deepfakes-a-growing-threat-to-democracy/.

European Union. Artificial intelligence act: Regulation (eu) 2024/1689 of the european parliament and of the council, June 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Accessed: 2024-09-24.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22466–22477, October 2023.

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. *arXiv preprint arXiv:2410.07369*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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. *European conference on computer vision (ECCV)*, pp. 740–755, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qGiZQb1Khm.

Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N. Metaxas, and Shiqing Ma. DIAGNOSIS: Detecting unauthorized data usages in text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=f8S3aLm0Vp.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 58047–58063. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b54d1757c190ba20dbc4f9e4a2f54149-Paper-Conference.pdf.

Zhiwei Yang, Mengnan Du, Zhiwei Jin, and Xia Hu. Gaussian shading: A new perspective for watermarking diffusion models. *arXiv preprint arXiv:2401.05678*, 2024.

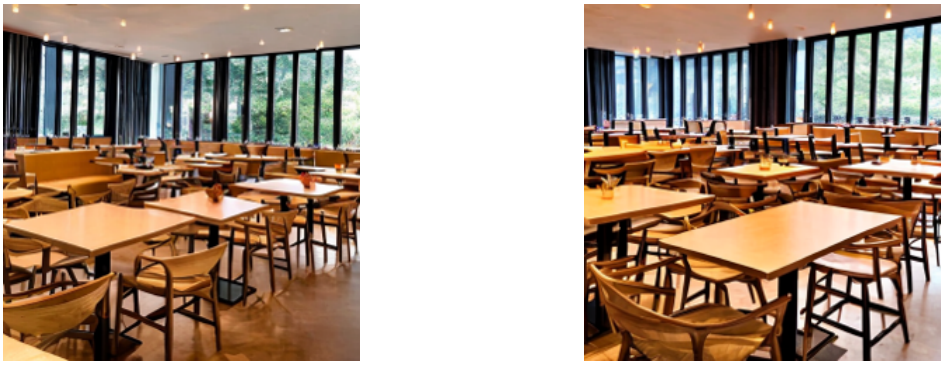
Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models, 2023.

## A Experiment details

The EDM model is trained unconditionally for 200 epochs with a batch size of 512 with other hyperparameters following Zhao et al. (2023).

In our experiments for LDMs, we perform full finetuning of the Stable Diffusion 2.1 (Rombach et al., 2022) model on the MSCOCO (Lin et al., 2014) or its synthetic counterpart regenerated with the diffusion model. We finetune all models for 10 epochs, with batch size = $4$ and learning rate = $1e-4$. We use 512×512 images.

## B Additional Figures



(a) DM1: TPR@FPR=1% 100%

(b) DM2: TPR@FPR=1% 1%

Figure 5: Comparison of the output of DM1 and DM2 and *the Tree-Ring* watermark.