RADIUM: RadioActive Decay of Image-Underlaid Marks

Michel Meintz, Louis Kerner, Maitri Vignesh Shah, Simon Hector,
Franziska Boenisch, Adam Dziedzic
CISPA Helmholtz Center for Information Security
{michel.meintz,louis.kerner,maitri.shah,simon.hector,boenisch,adam.dziedzic}@cispa.de

Abstract

Modern image generative models are able to produce photorealistic images. As those images become increasingly indistinguishable from real data and are published online, they are often scraped for subsequent training runs of new generative models. This practice of training on generated data has been shown to degrade model performance and cause model collapse. A possible mitigation lies in embedding radioactive watermarks into generated content. Radioactive watermarks are robust marks that are detectable in outputs of new models trained on watermarked data, enabling provenance tracing of generated content. In this work, we analyze the persistence of image watermarks across multiple training-generation runs. To do so, we introduce a novel statistical testing method RADIUM (RadioActive Decay of Image-Underlaid Marks) for reliable radioactivity detection across various watermarking methods. Using our RADIUM method, we observe disparate radioactivity across watermarking methods for image generative models. Only few watermarks remain detectable in subsequently trained models, while most decay severely, especially when the architecture of models differs. Our analysis highlights the critical need for more radioactive watermarking methods in the vision domain.

1 Introduction

The generation capabilities of large scale image generative models [7, 10, 27, 39, 44] have improved drastically, allowing them to generate high-quality images that are increasingly difficult to distinguish from natural, human-created data. These images are then published online, becoming part of the data that future models are trained on. As training these models requires large and diverse datasets [30], which are costly and time-consuming to collect and curate [2, 32], training data is often scraped from the internet [15, 32]. However, this scraped data contains not only natural images but also model-generated content and their quality makes it difficult to differentiate between them. This can create feedback loops where models are trained on their generated data leading to worse performance and ultimately to *model collapse* [1, 36, 43].

A promising solution to address these concerns is watermarking [3, 8, 17, 42]. Watermarking can help model owners to detect and track their generated data, enabling both the detection of misuse and prevention of training on their own generated data. Therefore, watermarks embed visually imperceptible signals into the generated images either during the generation process or after the image was generated. These imperceptible signals can then be identified by model owners, enabling them to detect their own generated images. While these watermarks are designed to be robust against removal attacks [24, 42], tracking their influence through the training of a new model requires an additional property: the watermark must remain detectable in the outputs of the model trained on watermarked data [17, 19]. This property is known as *radioactivity* [29, 31].

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

Radioactivity of watermarks for generative models has been analyzed in the language domain [31], demonstrating that fine-tuning a large language model (LLM) on watermarked text, results in the watermark being detectable in the output of the fine-tuned model. However, the extension of the radioactivity analysis to the image domain is non-trivial. Image generative models operate in high-dimensional continuous spaces, involving latent-space encoding or denoising processes that can destroy watermark signals. Existing work on radioactivity detection in the image domain [5] provided initial insights into the radioactivity of image watermarks, however, the existing evaluation lacked statistical guarantees.

In this work, we bridge this gap by extensively analyzing the radioactivity of image watermarking methods for different generation schemes under statistical guarantees. Therefore, we design a radioactivity test based on dataset inference [4, 6, 20, 25, 26, 28, 47], which aims to determine whether a collection of samples was part of a model's training data, by aggregating weak membership signals across the samples. We leverage the insights from dataset inference for our radioactivity test. Our method aggregates the watermark signals across many samples to detect their cumulative radioactivity. Beyond the previous dataset inference techniques, we further develop a new method for the radioactivity detection.

Our statistical testing method, *RADIUM* (*RadioActive Decay of Image-Underlaid Marks*), leverages recent advances in e-values to perform sequential, anytime-valid hypothesis testing [34]. We formulate radioactivity detection as a betting game played over a sequence of image pairs. At each step, the test places a *bet* against the null hypothesis that the suspect image contains no detectable watermark signal. As images are processed, evidence against the null is accumulated as statistical wealth, which is represented by an e-value: a nonnegative test statistic whose expectation is at most one under the null hypothesis. Consequently, large e-values provide evidence against the null, and rejecting the null whenever the e-value exceeds the value of $1/\alpha$, which yields a valid level-α test. The key advantage of this approach is that the evaluation is performed sequentially, and the detection procedure may be stopped at any point while providing valid statistical guarantees. In practice, this allows us to adaptively acquire additional samples when the evidence of radioactivity is inconclusive, or to terminate the test as soon as sufficient evidence of radioactivity has been accumulated.

In summary, we make the following contributions:

1. We introduce a new method *RADIUM* that combines dataset inference techniques with sequential testing to determine whether watermarks remain detectable after the watermarked images are used to train new generative models.
2. We develop an e-value-based statistical testing procedure for radioactivity detection, providing anytime-valid statistical guarantees across different image watermarking schemes.
3. We conduct extensive experiments across 7 watermarking methods and 2 model architectures, including diffusion and image autoregressive models: our results show how watermark signals decay over multiple rounds of model training and generation, with 4 out of the 7 tested watermarks exhibiting radioactive behavior.
4. We analyze factors that influence watermark radioactivity, providing practical insights for the design of image watermarks that remain detectable after new models are trained on them.

2 Background and Related Work

Diffusion Models. Diffusion Models (DMs) [14, 37] are trained by progressively adding noise to the data and then learning to reverse this process. In the forward diffusion process, Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1)$, is added to a clean image x, yielding the noisy sample $x_t \leftarrow \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\varepsilon$, where $t \in [0, T]$, is the diffusion timestep and $\alpha_t \in [0, 1]$ is a decaying parameter with $\alpha_0 = 1$ and $\alpha_T = 0$. The DM f_θ is then trained to predict the added noise ε, by minimizing $\mathcal{L}(x, t, \varepsilon; f_\theta) = \|\varepsilon - f_\theta(x_t, t)\|_2^2$. Latent diffusion models (LDMs) [27] improve over DMs, by performing the forward and reverse diffusion process in a latent space, rather than pixel space, significantly reducing computational complexity, making training scalable and more efficient. An encoder $\mathcal{E}$ encodes the input x to a lower dimensional latent representation $z = \mathcal{E}(x)$. The LDM diffusion loss is then formulated as $\mathcal{L}(z, t, \varepsilon; f_\theta) = \|\varepsilon - f_\theta(z_t, t)\|_2^2$.

Image Autoregressive Models. Image Autoregressive models (IARs) follow the next-token prediction paradigm from large language models (LLMs), treating images or image patches as sequences of discrete tokens, allowing fast and high quality image generation. Their similarity to LLMs allows them to benefit from the clear power-law scaling [39]. Some works rely on randomized inputs permuted into different factorization orders with annealing probability [44], where others proposed the next-scale prediction [39], making autoregressive image generation even faster and producing higher resolution images. The current state-of-the-art IAR, Infinity, predicts multiple bits per token, rather than predicting discrete tokens directly, allowing for a near infinite vocabulary size [10].

Watermarking. Watermarking embeds visually imperceptible signals into data, enabling the subsequent detection, attribution, and tracing of generated content. For images produced by generative models, watermarking can be applied at different stages of the generation pipeline. *In-generation* methods embed the signal during the sampling or generation process [8, 16, 17, 19, 42], whereas *post-processing* methods modify the image after it has been generated [3, 38, 46]. Our analysis focuses on several representative state-of-the-art watermarking schemes. We consider the post-processing watermarks: TrustMark [3], RivaGAN [46], and StegaStamp [38]. We further evaluate SIREN [23], a watermarking method explicitly designed for dataset tracing in DMs, which introduces an additional training loss to mitigate the limited learnability of prior watermarking approaches. We also experiment with in-generation methods: StableSignature [8], a diffusion-model watermark designed to have a limited impact on image quality, TreeRing [42], which is known for its robustness to standard image perturbations, and finally, we include BitMark [17], the only watermarking method currently available for the Infinity model. We provide additional information about the watermarks in appendix A.2.

Once an image has been watermarked, the embedded signal can be recovered or verified by an authorized detector, typically using a private key or message. To support reliable content tracing, watermarking schemes must remain robust to common image transformations and increasingly sophisticated removal attacks [24]. In general, image watermarking involves a three-way trade-off among *robustness*, *imperceptibility*, and *information capacity*. Robustness refers to the ability of the watermark to survive image augmentations and watermark removal attacks; imperceptibility captures the extent to which watermarked images remain indistinguishable from clean images; and information capacity denotes the amount of information encoded in the watermark, ranging from zero-bit detection to multi-bit messages. Improving one of these properties often comes at the expense of the others: for example, highly imperceptible watermarks may exhibit reduced robustness or capacity, whereas watermarks that have higher capacity might deteriorate more the visual quality and be less robust.

Watermark Radioactivity. Watermark radioactivity, *i.e.,* the persistence of watermarks through the training or fine-tuning process of a generative model is an additional robustness setting, that so far has had limited attention in the image domain. Recent studies in the language domain have demonstrated that training an LLM on watermarked data results in detectable traces in the outputs of the LLM [31]. Language watermarks change the prediction process of the model, by adapting its output probabilities to draw from a watermarked distribution. This distribution shift is then also detectable in the output of the second fine-tuned model. Sablayrolles et al. [29] show how to embed a radioactive signal into an image, such that a classifier trained on radioactive data is distinguishable from a classifier trained on clean data. Their method explicitly changes the models behavior, to trace the training data. A separate line of work with DIAGNOSIS [41], Recipe [48], and most recently SIREN [23], constructs watermarks for generative models, but with a fundamentally different goal: protecting natural training datasets by ensuring that any model trained on the watermarked data inherits a detectable signal. However they do not evaluate the radioactivity of existing in-generation watermarks to trace generated content. The closest prior work by Dubiński et al. [5], provides initial observations, but no statistical guarantees. We bridge this gap with our novel method that provides anytime-valid statistical detection of radioactivity across diverse watermarking schemes and models.

Dataset Inference (DI). DI was introduced as a framework for model ownership resolution first in supervised [25] and subsequently in the self-supervised [6] learning. The framework allows model owners to make a strong statistical claim if a given suspect model is a stolen copy of their own target (also referred to as a victim) model. To this end the behavior of the suspect model is compared to the behavior of the target model on samples of the training set. Compared to membership inference (MI) methods [35], which predict the membership of a single sample in a models training set, DI

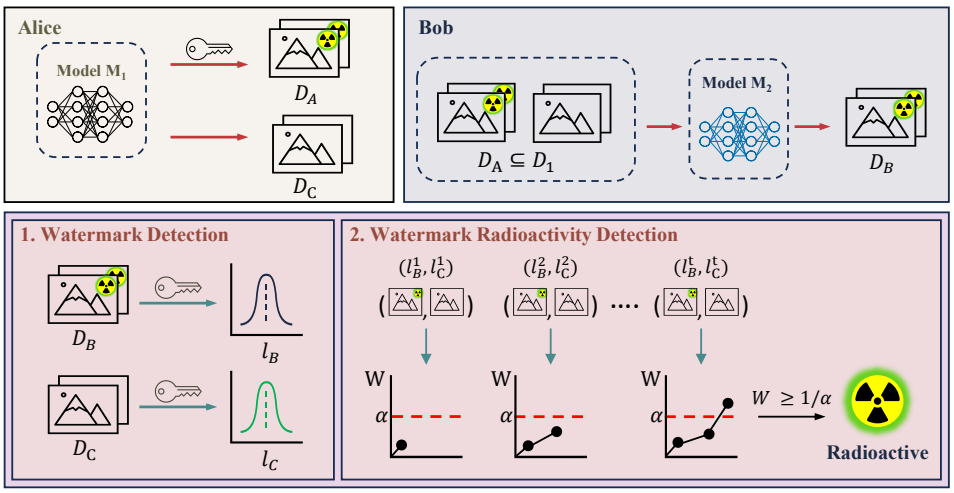


Figure 1: **Overview of our *RADIUM* method for radioactivity detection.** ***Alice*** creates model M_1. M_1 generates watermarked images D_A and clean images D_C. To train model M_2, ***Bob*** collects data D_1, which can contain a proportion of data D_A. M_2 then generates images D_B, which may carry traces of *Alice*'s watermark. *Alice* tests if *Bob* trained on her generated data: ***1. Watermark Detection.*** *Alice* uses her watermark detector on suspect images D_B generated from M_2 and on her clean images D_C generated from M_1 to obtain detection scores. ***2. Watermark Radioactivity Detection.*** *Alice* obtains the score distributions ℓ_B and ℓ_C. Our e-process tests whether D_B exhibits a significant watermark signal by accumulating wealth W. *Alice* detects radioactivity in D_B when W exceeds a given threshold $1/\alpha$.

enhances the membership signal by leveraging a collection of samples. However, this procedure faces a limitation in practice, namely that DI relies on statistical testing with p-values, which is limited to a fixed dataset size and predefined significance level. This causes issue as adding more data points, when the initial test fails, invalidates the statistical guarantees.

Sequential Testing by Betting. Sequential hypothesis tests [33, 40] frame the statistical tests as a betting game where data is drawn sequentially. Given a null hypothesis $H_0 : P \in \mathcal{P}_{null}$, which follows a distribution P and the alternative hypothesis $H_1 : P \in \mathcal{P}_{alt}$ as well as observations $Y_1, Y_2, \ldots,$ from space $\mathcal{Y}$, that are drawn i.i.d. according to P, a bettor can test H_0 by placing repeated bets on the observations Y_t, starting with initial wealth $W_0 = 1$. For testing, first a payoff function $S_t : \mathcal{Y} \to [0, \infty)$ is selected. Under the null, this function has an expected value of 1 and the future wealth after all observations is no larger than the current wealth. Then the next observation Y_t is revealed and the wealth changes by a factor of $S_t(Y_t)$. After t rounds, the wealth is described by $W_t = W_0 \prod_{i=1}^{t} S_i(Y_i)$. If H_1 is true, the wealth grows exponentially, otherwise the expected value of the wealth remains at W_0. Following [34], this process naturally leads to a sequential hypothesis test, where we reject the null if $W_t \geq 1/\alpha$, where α is the desired confidence level. Ville's inequality ensures that this test controls the type-I error at level α. This sequential test naturally fits into our detection method, where we can sequentially generate samples until we find sufficient evidence against the null or until we exhaust the budget allocated to the test.

3 Our Radioactivity Detection Method

3.1 Threat Model

Setting. Consider a model provider *Alice* who deploys a generative image model M_1. To trace the provenance of the content generated with M_1, *Alice* watermarks the outputs with a watermarking scheme $\mathcal{W}$ which has the detector $\mathcal{T}$. A second party, *Bob*, fine-tunes a model M_2 on a dataset D_1 that contains images generated by M_1. *Alice* wants to determine whether *Bob* has used any of her data to train M_2. We make the following assumptions:

Model Access: *Alice* has black-box access to M_2 and can generate images using prompts, but has no access to model weights or gradients.

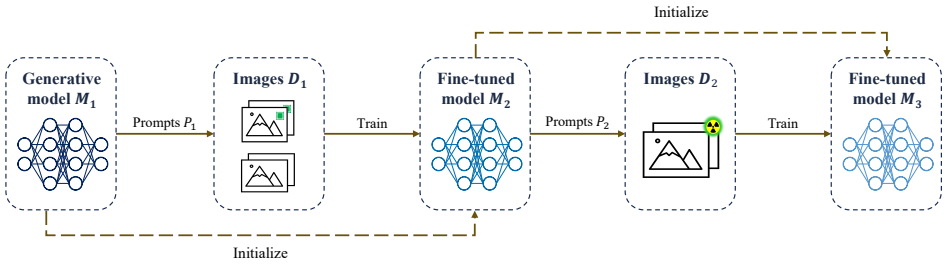


Figure 2: **Setup of the iterative model training mimicking the model collapse.** M_1 generates watermarked images D_1 with the prompts P_1. Then M_1 initializes the model M_2 and is trained on its own data D_1. The resulting model M_2 generates the radioactive dataset D_2 using the new prompts P_2. Iteratively the model M_i initializes the model M_{i+1} and is then trained on the data D_i.

Watermark Detection: *Alice* has access to her watermark detector $\mathcal{T}$, as well as to her own model M_1 and can generate watermarked and clean images with M_1. For the radioactivity detection, we leverage the detection scores of the detector, where higher scores mean stronger watermark detection. For example for BitMark we evaluate radioactivity using the z-score computed on the image. To provide outputs that are closer to being IID across different generative models and clean image distributions, we deploy whitening [8, 9], which computes biases in the detector on clean generated data and removes them via normalization. Therefore, we perform the PCA of the output of the watermark detectors on a set of 5k clean generated images. Thus we obtain the mean μ and eigendecomposition of the covariance matrix $\Sigma = U\Lambda U^T$. The whitening is then applied on the logit output of the detector, computing $y = \Lambda^{-\frac{1}{2}} U^T (x - \mu)$.

Prompt Knowledge: To detect traces of *Alice*'s watermark in *Bob*'s model, we consider two settings:

1. *supervised: Alice* knows the exact prompts of the watermarked images that *Bob* used to train M_2;
2. *unsupervised: Alice* has no knowledge about the prompts connected to the watermarked images that *Bob* used to train M_2.

3.2 Betting on Radioactivity

Overview. To help *Alice*, we sequentially generate clean images with *Alice*'s M_1 and (potentially) radioactive images with *Bob*'s M_2, using the same prompts. Then we apply the watermark detector $\mathcal{T}$ to obtain a detection score for both images. The goal is to determine whether the watermark scores of images generated by M_2 are significantly higher than those of the **clean** images generated by M_1. We therefore employ a statistical hypothesis testing by betting [34], where evidence is accumulated over round t, resulting in a wealth W_t corresponding to the significance of the test. We report significant detection when the e-value exceeds a predefined threshold. While similar results can be achieved with a t-test, this betting framework allows *Alice* to stop when sufficient evidence is gathered without committing to a fixed sample size. This can cut the required samples for detection from 1000s to less than 100.

Scoring procedure. Let $\mathcal{T}(x)$ denote a watermark score when evaluating an image x with the watermark detector $\mathcal{T}$. Higher values of $\mathcal{T}(x)$ indicate stronger evidence that the given image x contains the watermark. Depending on the watermark, $\mathcal{T}(x)$ may be a bit-wise accuracy, a z-score, or another detection statistic. Under the supervised setting, *Alice* has knowledge of the image-prompt pairs used by *Bob* during fine-tuning his model. *Alice* can generate test images using exactly these prompts, assuming that images generated from these prompts are more likely to carry the watermark signal. Under the unsupervised setting, *Alice* does not have knowledge of the prompts used by *Bob* during finetuning. Hence, she uses different prompts to generate image pairs.

Let $D_B = \{x_1, \ldots, x_{n_B}\}$ be the (potentially) radioactive images generated by the (Bob's) suspect model M_2 and $D_C = \{x'_1, \ldots, x'_{n_C}\}$ be the clean images generated by (Alice's) model M_1 without the watermarking scheme $\mathcal{W}$, where both x_i and x'_i were generated with the same prompt. We define the corresponding radioactive ℓ_B and clean ℓ_C scores as:

$$\ell_B := \{\mathcal{T}(x_i), \quad x_i \in D_B\}, \quad \ell_C := \{\mathcal{T}(x'_i), \quad x'_i \in D_C\}.$$

Let P_B and P_C denote the distributions of the radioactive ℓ_B and clean ℓ_C scores, respectively. To test whether the target model exhibits statistically significant watermark behavior, we use sequential testing by betting [34] and test the null hypothesis H_0 : *"The data was generated by a model that was trained without knowledge of the watermark."* (and also present the alternative hypothesis H_1):

$$H_0 : \ P_B \leq P_C, \qquad H_1 : \ P_B > P_C.$$

The test processes the score pairs (ℓ_B^t, ℓ_C^t), *i.e.,* observations, sequentially. At each step t, the payoff is defined as $S_t = 1 + \lambda_t u_t$, where $u_t \in \{-1, 0, 1\}$ is a bounded discrepancy score and λ_t is the wealth we bet on the observation.

Discrepancy metric. To compute u_t, we use the Kolmogorov-Smirnov (KS) witness function, estimated adaptively from all prior scores. This identifies the threshold, that maximally separates the two score distributions in the direction consistent with H_1. Under H_0 this discrepancy has zero conditional expectation given all prior observations, ensuring the validity of the test. The KS function then returns a discrepancy score $u_t \in \{-1, 0, 1\}$ based on whether the current observation agrees or disagrees with the null.

Betting strategy. For the betting fraction λ_t, we use the Online Newton Step (ONS) strategy [11]. ONS uses only the past observations to determine the betting fraction λ_t before seeing the current data point. It maintains a curvature accumulator that grows based on past observations. We provide more detailed information in Appendix A.1. A good betting strategy aims to grow wealth quickly, while controlling the risk of losing all wealth under the null.

Starting with the initial wealth $W_0 = 1$, the cumulative wealth is updated each round as $W_t = W_{t-1} \cdot (1 + \lambda_t u_t)$. This wealth can be interpreted as an *e-value* and represents the accumulated evidence against H_0. A key advantage of this process is the anytime-validity. Under H_0, the wealth process is a nonnegative supermartingale, so by Ville's inequality, rejecting H_0 when $W_t \geq 1/\alpha$ defines a valid α-level test. This enables *Alice* to monitor evidence continuously and stop when sufficient evidence is collected.

4 Experimental Evaluation

Experimental Setup. We evaluate the radioactivity of the in-generation watermarks: BitMark [17], TreeRing [42], and StableSignature [8]. We also test the post-processing watermarks: TrustMark [3], StegaStamp [38], RivaGAN [46], and SIREN [23]. For the evaluation we generate 2,000 image pairs of clean/watermarked images using MS-COCO Karpathy prompts. The post-processing watermarks are applied to clean images generated by the Infinity-2B model [10], Stable Diffusion 1.4 (SD1.4) and Stable Diffusion 2.1 (SD2.1) [27]. We fine-tune M_1, to obtain the second model M_2, on a total of 5,000 images to mimic model adaptation and report the hyperparameters in Appendix B.1. As threshold for the e-process we select the significance level $\alpha = 10^{-5}$.

4.1 Watermarks Exhibit Disparate Radioactivity

We report the main radioactivity detection results in Table 1. We fine-tune M_1 on a total of 5,000 self-generated images, of which a specific fraction is watermarked. We then generate images with M_2 using the training prompts (supervised) and disjoint prompts (unsupervised). For the evaluation, we sample these images generated by M_2 and their counterparts generated without the watermark by M_1 to infer them through the watermark detector $\mathcal{T}$, which yields the observation (ℓ_R^t, ℓ_C^t). We sample and bet on observations until either the wealth reaches the desired significance threshold $\alpha = 10^{-5}$ or we exceed the maximum number of 1000 steps. We report the number of samples required to reach $\alpha = 10^{-5}$ or the final e-value for the failed runs (that did not reject the H_0 after the 1000 steps). To prevent reporting false-positives, induced by fine-tuning on generated images, we also fine-tune an additional M_2 on non-watermarked images generated by M_1.

For all watermarks in the same model setting, we find that M_2 trained on clean images does not reach the significant detection. Our results show that BitMark, RivaGAN, StegaStamp, and SIREN watermarks can be detected when they make up 100% of the fine-tuning data for M_2 being Infinity-2B. However only BitMark remains detectable when constituting 1% of the fine-tuning data. We observe that the watermarks are radioactive for lower fractions of watermarked data on the Infinity-2B model compared to SD2.1. There are two potential factors that influence these differences, namely (1)

Table 1: **We find disparate radioactivity across models and watermarks.** We train M_2 on different fractions of watermarked data % generated by M_1. We report the stopping time, measured as the number of paired observations, at which the e-process crosses the rejection threshold $1/\alpha = 10^5$, corresponding to a significance level of $\alpha = 10^{-5}$. Faded cells report the final e-value for runs that did not reject H_0 within 1000 steps.

Watermark	Setting	Model	Fraction of Watermarked Data						
			0%	1%	5%	10%	25%	50%	100%
BitMark	Supervised	Inf-2B	0.28	87	66	53	52	56	37
	Unsupervised	Inf-2B	0.31	132	78	47	46	43	34
RivaGAN	Supervised	Inf-2B	0.07	0.73	0.52	0.84	201	74	43
	Unsupervised	Inf-2B	1.48	3.57	2.16	0.18	156	82	42
StegaStamp	Supervised	Inf-2B	0.05	0.89	0.18	0.16	5.89	249	35
	Unsupervised	Inf-2B	1.29	0.11	0.15	0.74	0.26	165	36
TrustMark	Supervised	Inf-2B	1.13	3.14	2.09	1.69	1.05	7.05	1.33
	Unsupervised	Inf-2B	5.06	1.13	5.06	5.06	5.06	2.53	9.61
SIREN	Supervised	Inf-2B	0.35	0.17	19.76	245.32	223	50	36
	Unsupervised	Inf-2B	0.1	0.08	0.1	0.28	0.12	65	38
TreeRing	Supervised	SD2.1	11.39	3.38	0.59	3.38	1.13	5.06	3.38
	Unsupervised	SD2.1	1.45	1.85	25.63	9.57	10.59	3.5	2.17
Stable Signature	Supervised	SD2.1	0.18	0.17	0.1	0.17	0.59	0.05	15.91
	Unsupervised	SD2.1	0.13	0.06	0.13	0.08	0.07	0.1	0.19
TrustMark	Supervised	SD2.1	0.63	1.38	0.11	0.22	0.19	0.06	0.14
	Unsupervised	SD2.1	0.19	0.03	0.1	0.12	0.29	0.12	0.06
StegaStamp	Supervised	SD2.1	0.43	0.06	0.29	0.08	0.14	0.09	37
	Unsupervised	SD2.1	0.19	0.06	0.13	0.09	0.72	0.09	42
RivaGAN	Supervised	SD2.1	1.05	147.08	0.31	130.91	78.29	1.02	212
	Unsupervised	SD2.1	0.3	0.35	0.46	0.55	0.04	0.18	0.72
SIREN	Supervised	SD2.1	0.12	1.16	0.17	0.38	0.06	0.05	95
	Unsupervised	SD2.1	0.31	0.31	0.36	0.14	0.07	0.6	208

different variational auto-encoders (VAEs) and (2) different training and generation processes. In Figure 6 we detect watermarks after they were encoded and decoded with the VAEs and find that the watermarks are stronger impacted by encoding with the VAE in SD2.1. The VAE in Infinity leaves fine-grained signals intact, which makes the watermark more easily learnable in the latent space. Furthermore, in the diffusion training process the images are iteratively noised, such that the model can learn to denoise the images. This noising-denoising process mimics a noising watermark removal attack, which can distort the watermark signal and result in limited radioactivity. We analyze two different knowledge assumptions concerning the prompts of the training images of M_2, the supervised setting, and the unsupervised setting. Our main results show that prompt access has limited impact on the detection of radioactivity, with the supervised and unsupervised setting agreeing in most cases. This prompt independence shows that radioactivity is not an artifact of memorization, where the mark is learned in combination to a specific prompt, but becomes a general capability of the model.

To further evaluate potential factors for radioactivity, we follow SIREN [23] and evaluate the impact of the watermark on the loss $\mathcal{L}$ of the generative model GM by computing:

$$\gamma = \mathcal{L}_{GM}(GM(x_{clean}|c)) - \mathcal{L}_{GM}(GM(x_{mark}|c)), \tag{1}$$

where x_{clean} is the clean image, x_{mark} its watermarked counterpart, and c the conditioning. SIREN maximizes this quantity γ, to adapt the watermark to specific features of the image and make it more learnable. We compute this quantity γ for the images generated by SD2.1 in Figure 4 and for the images generated by Infinity-2B in Figure 5. We find that watermarks that have a strong impact on γ are faster detectable in the generated images of M_2, whereas watermarks with $\gamma \approx 0$ are learned less. Specifically, TreeRing and StableSignature are hardly or not at all radioactive in SD2.1, whereas RivaGAN, StegaStamp, and SIREN are detectable with fewer samples. This analysis on Infinity-2B shows that Infinity's loss is impacted more by the watermark, which mirrors our results in Table 1, where watermarks are detected faster on Infinity-2B compared to SD2.1.

Table 2: **Radioactivity decay over multiple training and generation runs.** We use data generated by M_{i-1} to fine-tune M_i. Each row's bracketed percentage (%) gives the fraction of watermarked data used to train each model ($M_2 \dots M_6$). The reported values follow the same format as in Table 1.

Watermark	Setting	Model	M_2	M_3	M_4	M_5	M_6
BitMark (5%)	Supervised	Inf-2B	66	100	199	7.59	–
	Unsupervised	Inf-2B	78	143	236	0.95	–
BitMark (100%)	Supervised	Inf-2B	37	34	34	37	32
	Unsupervised	Inf-2B	34	33	34	32	33
StegaStamp (50%)	Supervised	Inf-2B	249	161	90	282	0.68
	Unsupervised	Inf-2B	165	154	157	228	433
StegaStamp (100%)	Supervised	Inf-2B	35	33	35	36	39
	Unsupervised	Inf-2B	36	33	31	41	34
SIREN (50%)	Supervised	Inf-2B	50	58	68	61	76
	Unsupervised	Inf-2B	65	60	78	73	72
SIREN (100%)	Supervised	Inf-2B	36	35	42	43	48
	Unsupervised	Inf-2B	38	41	39	51	56
StegaStamp (100%)	Supervised	SD2.1	37	71	124	8.52	–
	Unsupervised	SD2.1	42	70	109	7.77	–
SIREN (100%)	Supervised	SD2.1	95	7.35	–	–	–
	Unsupervised	SD2.1	208	14.74	–	–	–

4.2 Iterative Model Training

This section focuses on watermark radioactivity under the lens of model collapse [1], where models are iteratively trained on their own generated content. To formalize the setting, we define a set of prompts P_i that are used by model M_i to generate data D_i. We enforce that all prompt sets are pairwise distinct, to reduce the impact of the prompt on the detection of the radioactivity. Formally, we use M_i and P_i to generate dataset D_i and then fine-tune M_i on D_i to obtain M_{i+1}. We provide a visualization of the setup in Figure 2. Based on our findings in Table 1 we focus on BitMark, StegaStamp, and SIREN, since they showed higher levels of radioactivity. For $i > 2$ we train the model on 100% of the data generated by its' predecessor. Only M_2 is trained on $p\%$ of watermarked and $1 - p\%$ of clean data. Our results for the iterative model training in Table 2 show that the watermarks on Infinity-2B experience no or hardly any degradation across multiple fine-tuning iterations. On the other hand StegaStamp on SD2.1 decays gracefully over multiple iterations, whereas SIREN is no longer detectable in the M_3 model. SIREN shows clear differences in the behavior on Infinity-2B and SD2.1, which is likely related to the resistance of the watermark against encoding and decoding, analyzed in Figure 6. We provide examples of the output of the SD2.1 in Figure 11 and Infinity-2B in Figure 10.

4.3 Cross Architecture Analysis

The results from the two previous sections, show that some watermarks can be radioactive and enable content tracing even through multiple model iterations on the **same model architecture**. However in most case the model trained on the watermark images generated by M_1 will have a different architecture. Initially we find that some watermarks exhibit a distribution shift when evaluated on different model architectures. To guarantee the validity of our evaluation, we additionally assume access to M_2 which is obtained by training on 0% watermarked data. This failure case shows inherent limitations of existing watermarks. We further evaluate this in Appendix C.4. For the cross-architecture case we focus on settings where a larger model M_1 is used to generate data for a smaller model M_2. Our results in Table 3 indicate that the cross-architecture case is much harder than the same architecture case, where the watermarks are detected only when making up high fractions of the data and many samples are evaluated. To verify that our findings are not an artifact of the specific backbones we study, we additionally evaluate two substantially larger and more recent diffusion models, Stable Diffusion 3.5-medium and FLUX.1-dev with 12B parameters, and report the results in Appendix C.3. We observe the same qualitative behavior: SIREN and RivaGAN become detectable once they constitute 25% or more of the fine-tuning data, StegaStamp only under higher fractions,

Table 3: **Detection performance across model architectures.** We train M_2 on different fractions of watermarked data % generated by M_1, where M_1 and M_2 use different architectures. The reported values follow the same format as in Table 1.

Watermark	Setting	M_1	M_2	Fraction of Watermarked Data					
				1%	5%	10%	25%	50%	100%
BitMark	Supervised	Inf-2B	SD2.1	0.13	0.07	0.21	22.6	148	56
	Unsupervised	Inf-2B	SD2.1	0.16	0.31	0.05	1.48	173	59
	Supervised	Inf-2B	SD1.4	0.26	0.11	0.53	2.13	137	51
	Unsupervised	Inf-2B	SD1.4	0.19	0.23	0.12	184.92	134	54
RivaGAN	Supervised	Inf-2B	SD2.1	0.16	0.14	0.08	0.15	0.1	0.08
	Unsupervised	Inf-2B	SD2.1	0.58	0.13	0.13	0.59	0.13	0.24
	Supervised	Inf-2B	SD1.4	0.07	0.1	0.51	0.09	0.58	0.16
	Unsupervised	Inf-2B	SD1.4	0.71	0.4	0.09	0.48	0.17	0.04
	Supervised	SD2.1	SD1.4	0.13	0.72	0.07	0.25	0.06	0.62
	Unsupervised	SD2.1	SD1.4	0.38	0.23	0.11	0.95	0.28	0.07
StegaStamp	Supervised	Inf-2B	SD2.1	0.17	0.06	0.12	0.09	218	61
	Unsupervised	Inf-2B	SD2.1	0.05	0.44	0.05	0.15	97	57
	Supervised	Inf-2B	SD1.4	0.57	0.08	2.33	0.34	3.53	42
	Unsupervised	Inf-2B	SD1.4	0.15	0.08	0.05	0.64	0.33	44
	Supervised	SD2.1	SD1.4	0.24	0.23	0.33	0.07	0.13	21.01
	Unsupervised	SD2.1	SD1.4	0.13	0.3	0.06	0.11	0.15	7.42
TrustMark	Supervised	Inf-2B	SD2.1	0.28	0.76	1.6	0.32	1.75	61
	Unsupervised	Inf-2B	SD2.1	0.25	0.11	2.44	50.98	6.34	54
	Supervised	Inf-2B	SD1.4	0.83	0.27	0.07	0.07	0.03	0.04
	Unsupervised	Inf-2B	SD1.4	0.66	0.53	0.22	0.1	0.09	0.07
	Supervised	SD2.1	SD1.4	1.8	1.6	0.09	0.31	0.16	0.24
	Unsupervised	SD2.1	SD1.4	1.1	0.13	0.04	0.07	0.12	0.15
SIREN	Supervised	Inf-2B	SD2.1	0.08	0.26	0.9	345	154	65
	Unsupervised	Inf-2B	SD2.1	0.34	0.11	1.08	11.33	10.72	86
	Supervised	Inf-2B	SD1.4	0.54	0.35	0.63	7.98	188	90
	Unsupervised	Inf-2B	SD1.4	0.49	0.21	0.4	0.36	75.29	97
	Supervised	SD2.1	SD1.4	0.25	0.14	1.15	0.11	0.81	118
	Unsupervised	SD2.1	SD1.4	0.69	0.2	0.25	0.18	0.12	146

and TrustMark remains non-radioactive throughout. Notably, these larger diffusion models behave more similarly to Infinity-2B than to SD2.1, suggesting that model scale is itself a factor in how well a watermark is retained.

4.4 Radioactivity Removal Robustness

Given the knowledge about watermark radioactivity, *Bob*, who owns M_2, now wants to make sure that his model is not radioactive, to avoid *Alice* detecting he illegitimately trained on her data. To remove any radioactive traces, *Bob* collects carefully curated data, ensuring that the curated dataset contains no watermarked images. He then performs a cleaning step, by carefully fine-tuning the model on the clean data. Therefore, we analyze the models M_2 trained on different fractions of watermarked data. Our results in Table 9, show that the radioactivity of current watermarks is very fragile and decreased drastically after additional fine-tuning on clean data. Only BitMark, when using 100% of the fine-tuning data, is radioactive to M_3. We provide additional results for the radioactivity robustness under an adversary that applies image augmentations in Appendix C.2. We find that interjections before or during fine-tuning have the strongest negative effect on radioactivity, as they disrupt the watermark signal, before the model can learn it.

4.5 Hyperparameter Ablation

We analyze different hyperparameters that influence the radioactivity of watermarks in downstream models. Table 13 shows that a higher learning rate generally causes the e-process to stop earlier, consistent with the watermark being more strongly represented in the model's outputs. The number of fine-tuning epochs has a setup-dependent effect (Table 16): on the Infinity backbone, increasing the number of fine-tuning epochs markedly reduces detection, compared to a minor influence on the

SD2.1 model. In contrast, the effective batch size has little impact on radioactivity (Table 17) - the watermark signal is not washed out by computing the gradient on a batch. In Table 14 we analyze the impact of changing the strength of the watermark on the radioactivity. We focus on BitMark and SIREN and change their strength in proportion to their default values ($\beta = 1.5$ for SIREN and $\delta = 2$ for BitMark). The results show that a higher watermark strength reduces the number of samples required for significant detection for SIREN. Notably, a reduced watermark strength results in the watermark no longer being detectable.

4.6 Core Factors for Radioactive Watermarks

This section crystallizes the core factors making up a radioactive watermark and supports them with our experimental evidence. **(1) Robustness to the latent encoding is a prerequisite.** No watermark that is destroyed by the encoding-decoding cycle of the downstream model's VAE is radioactive in that model (Figure 6). This is the most predictive property we observe, and explains why StableSignature, RivaGAN, and SIREN are strongly weakened on SD2.1, whereas Infinity-2B leaves fine-grained signals intact. **(2) The mark must be reused across samples rather than tied to individual content.** A content-dependent signal appears to the downstream model as sample-specific noise, whereas a signal that reoccurs across the training set becomes a learnable regularity. This is especially apparent in our results with BitMark when attenuating the signal in Table 19. The same also shows when we increase the number of keys and the signals interfere with each other in StegaStamp (Table 15). **(3) The mark must survive the training process itself.** Augmentations applied before or during fine-tuning suppress radioactivity far more effectively than the same augmentations applied afterwards as we explore in Table 11. So a radioactive watermark must be robust to the standard augmentation pipeline, not only to post-hoc removal attacks.

5 Conclusions

We introduced a statistical method based on sequential betting that enables the analysis of the radioactivity of watermarks in downstream models. We investigated the radioactivity and radioactive decay of post-processing and in-generation watermarks spanning three different model architectures in the diffusion and autoregressive domain. Our results demonstrate that current watermarking techniques lack reliable detection in downstream models and can not guarantee long-term provenance tracing of their generated content. We show how watermarks that are robust against encoding and impact down-stream model loss consistently are more radioactive. Our ablation studies identify several factors influencing the radioactivity of image watermarks, such as the model architecture, learning rate and watermark strength. By providing a unified statistical method for radioactivity detection across different image watermarks, we enable watermark designers to evaluate watermark radioactivity through model derivatives.

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=ShjMHfmPs0`.

[2] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=Qf8uzIT1OK`.

[3] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *IEEE International Conference on Computer Vision (ICCV)*, October 2025.

[4] Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. Cdi: Copyrighted data identification in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18674–18684, 2025.

[5] Jan Dubiński, Michel Meintz, Franziska Boenisch, and Adam Dziedzic. Are watermarks for diffusion models radioactive? In *The 1st Workshop on GenAI Watermarking*, 2025. URL https://openreview.net/forum?id=gtXbVRMwQh.

[6] Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, and Nicolas Papernot. Dataset inference for self-supervised models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=CCBJf9xJo2X.

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[8] 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)*, pages 22466–22477, October 2023.

[9] Enoal Gesny, Eva Giboulot, Teddy Furon, and Vivien Chappelier. Guidance watermarking for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5ifzhjMCKq.

[10] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[11] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2–3):169–192, December 2007. ISSN 0885-6125. doi: 10.1007/s10994-007-5016-8. URL https://doi.org/10.1007/s10994-007-5016-8.

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[16] Nikola Jovanović, Ismail Labiad, Tomas Soucek, Martin Vechev, and Pierre Fernandez. Watermarking autoregressive image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=hVdD72iom4.

[17] Louis Kerner, Michel Meintz, Bihe Zhao, Franziska Boenisch, and Adam Dziedzic. Bitmark: Watermarking bitwise autoregressive image generative models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VSir0FzFnP.

[18] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

[19] Jan Kociszewski, Hubert Jastrzębski, Tymoteusz Stępkowski, Filip Manijak, Krzysztof Rojek, Franziska Boenisch, and Adam Dziedzic. SERUM: Simple, efficient, robust, and unifying marking for diffusion-based image generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AiBUm6iKBf.

[20] Antoni Kowalczuk, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Privacy attacks on image autoregressive models. In *Forty-Second International Conference on Machine Learning (ICML)*, 2025.

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

[22] P. Langley. Crafting papers on machine learning. In Pat Langley, editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

[23] Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, Jie Zhang, Run Wang, and Tianwei Zhang. Towards Reliable Verification of Unauthorized Data Usage in Personalized Text-to-Image Diffusion Models . In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2564–2582, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/SP61157.2025.00073. URL https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00073.

[24] Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mDKxlfraAn.

[25] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hvdKKV2yt7T.

[26] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[27] 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.

[28] Lorenzo Rossi, Bartłomiej Marek, Franziska Boenisch, and Adam Dziedzic. Natural identifiers for privacy and data audits in large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=doaAUf9Pi7.

[29] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Herve Jegou. Radioactive data: tracing through training. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine*

Learning Research, pages 8326–8335. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sablayrolles20a.html.

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

[31] 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.

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[33] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication, 2019. URL https://arxiv.org/abs/1903.06991.

[34] Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2024. doi: 10.1109/TIT.2023.3305867.

[35] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017. doi: 10.1109/SP.2017.41.

[36] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024.

[37] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.

[38] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020.

[39] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[40] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020. URL https://doi.org/10.1214/20-AOS2020.

[41] 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.

[42] 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, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 58047–58063. Curran Associates, Inc., 2023.

[43] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2113–2147, 2024.

[44] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation, 2024. URL https://arxiv.org/abs/2411.00776.

[45] Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DztaBt4wP5.

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

[47] Bihe Zhao, Pratyush Maini, Franziska Boenisch, and Adam Dziedzic. Unlocking post-hoc dataset inference with synthetic data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=a5Kgv47d2e.

[48] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

A Extended Background

A.1 Betting Strategy

The betting strategy determines the size of the bet to wager on the next sample, based on the already observed discrepancy between the two distributions. In our method we leverage the Online Newton Step (ONS) strategy.

Online Newton Step (ONS). First we use the discrepancy metric (in our case derived from Kolmogorov-Smirnov) to compute the bounded discrepancy score $u_t \in \{-1, 0, 1\}$. We then update a one-dimensional curvature accumulator A_t and calculate the subsequent betting amount λ_{t+1} for round $t + 1$:

$$z_t = \frac{u_t}{1 + \lambda_t u_t}, \quad A_t = A_{t-1} + z_t^2,$$

$$\lambda_{t+1} = \Pi_{[0,\,\lambda_{\max}]}\left(\lambda_t + \frac{2}{2 - \log 3} \cdot \frac{z_t}{A_t}\right),$$

with initializations $\lambda_1 = 0$ and $A_0 = 1$. Here $\Pi_{[0,\,\lambda_{\max}]}$ denotes projection onto the permissible, bounded stake interval. Restricting λ_t to $[0, \lambda_{\max}]$ rather than $[-\lambda_{\max}, \lambda_{\max}]$ enforces the one-sided nature of the test: the bettor only ever places stakes in the direction of H_1, consistent with the hypothesis that radioactive images produce stochastically higher watermark scores than clean images. While higher values of $\lambda_{\max}$ permit more aggressive betting, the upper bound is mathematically necessary to prevent the wealth process from reaching zero (ruin). We set $\lambda_{\max} = 0.5$ throughout. A good betting strategy aims to grow wealth quickly when the alternative hypothesis is true, while controlling the risk of losing all wealth under the null. ONS achieves this through its curvature accumulator A_t, which grows with the sum of squared past gradients and provides a data-dependent learning rate that achieves logarithmic regret without requiring prior knowledge of the problem difficulty.

A.2 Extended Related Work

Watermarking. The post-processing watermarks differ primarily in the space they are applied in and the robustness that they target. StegaStamp [38] is designed to embed signals that survive strong perturbation pipelines and can be detected in realistic physical detection scenarios. It incorporates a specific robustness training pipeline and trades visual fidelity for persistence. RivaGAN [46] was proposed for video and leverages an attention-based mechanism that hides messages in content-dependent regions. This design makes it robust against video-processing operations, in particular cropping and compression. Additionally, it is resolution-agnostic by construction, as its encoder has no bottleneck. SIREN [23] is a watermarking method designed to be learned by models fine-tuned on watermarked data. Therefore the watermark leverages a proxy student model and optimizes the watermark embedder, such that the watermark *decreases* the loss of the proxy model on the watermarked images, resulting in the watermark providing helpful features for the model and being learned more easily. Finally, TrustMark [3] targets broad robustness to common image edits, while preserving high image quality. To this end it incorporates a focal frequency loss (FFL) preserving the high frequency content of the images. Compared to RivaGAN and StegaStamp, which disrupt high-frequency content of the images, the FFL results in TrustMark being primarily applied in lower frequencies.

The evaluated in-generation watermarks differ substantially in the watermark embedding techniques. StableSignature [8] fine-tunes a latent-models decoder, so that every decoded image carries the watermark. This approach hardly affects the generation process, resulting in only minor changes to the generated images while staying robust under strong edits. TreeRing [42] is a watermark specifically designed for diffusion models. It applies a ring in the low frequencies of the initial noise in the diffusion process, which makes it robust against geometric perturbations, such as cropping, rotating, and flipping. However, its detection is highly expensive, as it requires a full DDIM-inversion from the image space back to the noise space. Finally, BitMark [17] is a watermark specifically designed for bit-wise generative models. These models produce images by generating bit-strings of varying length, sampling from a vocabulary consisting of bitstrings. BitMark influences the sampling process by separating the vocabulary into a green list and a red list of bitstrings with length n and biases the generation process such that bitstrings from the green list are sampled more frequently.

A.3 Image and Text Radioactivity

Text Watermarking. Existing work on LLM radioactivity [31] detects watermarks through a principled analysis of token-level generation in LLMs. This line of work focuses on watermarking methods that modify the decoding procedure [18]. Specifically, LLM decoding can be viewed as taking a context sequence $(x^{(-C)}, \ldots, x^{(-1)}) \in \mathcal{V}^C$, where $\mathcal{V}$ denotes the model vocabulary and C the context length. Given this context, the LLM outputs a logit vector $l \in \mathbb{R}^{|\mathcal{V}|}$, which is converted into next-token probabilities via $p = \mathrm{softmax}(l) \in [0,1]^{|\mathcal{V}|}$. The next token $x^{(0)}$ is then sampled from this distribution. Watermarking schemes influence this generation process using a secret key: the key and the current context are provided to a hash function, whose output serves as a seed s for a random generator. This seed determines how the watermark biases the next-token distribution, for example by increasing the sampling probabilities of tokens in a designated green list [18]. For detection, the generated text is re-tokenized, the corresponding contexts are reconstructed, and a watermark score is assigned to each token.

Image vs. Text Radioactivity. Sander et al. [31] make two key design choices in their radioactivity detection scheme. First, to improve detection, they introduce a filter ϕ that restricts the analyzed token stream to k-grams likely to have appeared during training. Tokens generated by M_2 are scored only when their context is not filtered out by ϕ. Second, they apply token de-duplication: a token is scored only if its context has not been observed previously during detection. Together, these two design choices yield faster and unbiased detection. However, these insights do not directly extend to the image domain. Unlike text watermarking methods that modify the sampling or decoding procedure [18], most image watermarking methods do not influence the generative model's sampling process directly, with a few exceptions [16, 17]. Instead, image watermarks are typically applied either before generation, e.g., to the initial noise in DMs [19, 42], or after generation, e.g., to generated outputs [3, 38]. Furthermore, image watermarks lack an inherent sequential context analogous to text k-grams, making it unclear how to de-duplicate watermark signals or remove context-induced biases in a comparable way. As a result, radioactivity in the image domain differs substantially from radioactivity in the text domain. In addition, image-specific transformations, such as changes in resolution, compression, cropping, or other augmentations, can further degrade watermark signals and negatively affect radioactivity detection.

B Experimental Setup

B.1 Hyperparameters

Table 4: **Hyperparameters used for model training.**

	Optimizer	Learning Rate	Batch Size	Epochs
Stable Diffusion 2.1	Adam	1×10^{-5}	4	5
Stable Diffusion 1.4	AdamW	1×10^{-5}	4	5
Stable Diffusion 3.5	AdamW	5×10^{-6}	4	5
Infinity-2B	AdamW	6×10^{-4}	4	2
FLUX.1-dev	AdamW	5×10^{-6}	4	5

For Stable Diffusion 2.1, we used Adam optimizer with a learning rate of 1×10^{-5} and fine-tune for 5 epochs with batch size of 4, following CLiD [45]. We fine-tune Infinity-2B [10] following their official implementation, using the AdamW optimizer with a learning rate of 6×10^{-4} for 2 epochs. This learning rate is lower than the default configuration (6×10^{-3}), as we adjusted for our specific fine-tuning task. The input resolution is set to 1024×1024. To fit the model on our hardware, we use a per-GPU batch size of 2 with 4 steps of gradient accumulation resulting in an effective batch size of 8.

Finally, we adjusted the noise strength parameter. While the default configuration uses noise strength of $\sigma = 0.3$, we noticed this degraded our image generation quality. To address this, we ran tests fine-tuning Infinity-2B for 1 epoch for the noise strengths 0.1, 0.2 and 0.3 (see Figure 3). After visually comparing the outputs, we selected 0.1 as it produced best images among the three settings.

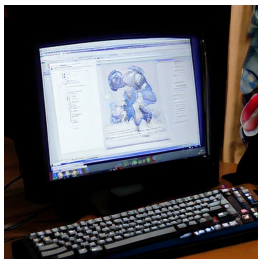
Table 5: **Image Generation Cost.** We compute the speed for image generation for single image batch sizes for the two generative backbones. We report the generation in seconds / image averaged across 1k images.

Model	SD2.1	Inf-2B
Generation	1.38	1.97

Table 6: **Watermark Detection Cost.** We compute the speed for watermark detection for single image batch sizes for the different watermarking schemes. We report the detection speed in seconds / image averaged across 1k images.

Model	StegaStamp	TrustMark	RivaGAN	SIREN	StableSignature	TreeRing	BitMark
SD2.1	0.006	0.008	0.009	0.009	0.009	1.198	0.134
Inf-2B	0.01	0.013	0.028	0.02	0.009	1.198	0.134

(a) $\sigma = 0.1$ image.



(b) $\sigma = 0.2$ image.

(c) $\sigma = 0.3$ image.

Figure 3: **Effect of Noise Strength on Infinity-2B.** Prompt: *"A young boy standing in front of a computer keyboard."*

B.2 Environment

Our experiments are performed on Ubuntu 22.04 with NVIDIA A100 Graphics Cards with 40GB and 80GB of memory. To run our experiments we used CUDA Version 12.5 and Python 3.12.4.

B.3 Computational Cost

The radioactivity analysis relies on multiple different components, most importantly: data generation, watermark detection and the e-process. Running the e-process when all 1k samples are generated and detected requires only 0.2 seconds. However, the image generation and watermark detection cost do not depend on *RADIUM*, but on the generative model and watermark scheme. Therefore we provide the generation and detection time in seconds per image in Table 5 and Table 6.

B.4 Discussion and Limitations

Training large-scale generative models, such as Infinity-2B [10] or Stable Diffusion 2.1 [27], from scratch is prohibitively expensive, especially for many training-generation runs or dataset compositions. We therefore do not perform full pretraining experiments at this scale. Instead, we study a fine-tuning setting as a proof of concept, following prior analyses of model collapse [36] and LLM radioactivity [31]. This setting allows us to systematically evaluate watermark radioactivity across multiple training-generation runs while keeping the experimental setup computationally feasible.

This choice, however, limits the extent to which our results transfer to large-scale pretraining regimes. Fine-tuning may preserve, amplify, or attenuate watermark signals differently from training from scratch, depending on the amount of synthetic data, the strength of the watermark, the model architecture, and the optimization procedure. Similarly, our conclusions are based on the set of watermarking methods, model architectures, and hyperparameters considered in our experiments. Other watermarking schemes or parameters may exhibit different levels of persistence.

Table 7: **Detection performance with 10k samples.** Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	Setting	Model	Fraction of Watermarked Data			
			0%	5%	25%	100%
BitMark	Supervised	Inf-2B	0.0358	143	53	34
	Unsupervised	Inf-2B	0.0438	208	51	33
StegaStamp	Supervised	SD2.1	0.099	0.0225	0.110	46
	Unsupervised	SD2.1	0.79	0.135	0.105	38

Table 8: **Detection performance for SD2.1 trained from scratch.** We train a randomly initialized M_2 on different fractions of watermarked data % generated by the randomly initialized and then trained M_1. Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	Setting	Model	Fraction of Watermarked Data
			100%
StegaStamp	Supervised	SD2.1	37
	Unsupervised	SD2.1	40
SIREN	Supervised	SD2.1	1.71
	Unsupervised	SD2.1	0.156

To explore the behavior of radioactivity on a larger scale and when evaluated from pre-training, we perform two ablations. (1) We increase the number of training samples by two, evaluating the radioactivity of the models when trained on 10k samples and (2) we initialize SD2.1 from scratch and first train on clean natural MS-COCO data for 5 epochs to obtain M_1 and then train a new randomly initialized model on the watermarked outputs from M_1. We report our results in Table 8 and Table 7 respectively. We find that a larger fine-tuning set has limited impact on the observed radioactivity. We believe that our core principles and insights that govern radioactivity also transfer to downstream retraining or web-scale data reuse. Specifically, important factors such as VAE robustness as a necessary condition for radioactivity, will also hold. Our insights that cross-architecture robustness is strictly harder than same-architecture transfer, holds for larger schemes, as the radioactivity is not only data, but also architecture dependent. Furthermore, as our results provide an empirical upper bound on the radioactivity in a controlled setting, watermarks that are not radioactive in this setting will likely also not be radioactive under realistic use cases.

C Additional Results

C.1 Watermark Analysis

Loss Analysis. We analyze the impact of the watermark on the generative model by computing:

$$\tau = \mathcal{L}_{GM}(GM(x_{clean}|c)) - \mathcal{L}_{GM}(GM(x_{watermarked}|c)),$$

where GM is the generative backbone, x is the image in either $clean$ or $watermarked$ version and c is the conditioning. For the Infinity-2B model the $\mathcal{L}_{GM}$ is the bit-wise cross-entropy loss averaged across all scales. For SD2.1 we compute the average loss across 20 different diffusion timesteps. We provide the loss plots for the images generated by SD2.1 in Figure 4 and for the images generated by Infinity-2B in Figure 5. We find that SD2.1 exhibits a lower impact on the loss from the different watermarks, whereas Infinity-2B exhibits higher loss for the watermarked images compared to the clean images. Unsurprisingly the post-processing watermarks have an overall lower impact on the loss and as we found in Table 1 they are also harder to detect in downstream models.

VAE-Robustness. For the watermarks to be learnable by the generative model, the watermark has to persist both in the latent space of that model and be detectable after decoding. The generative models

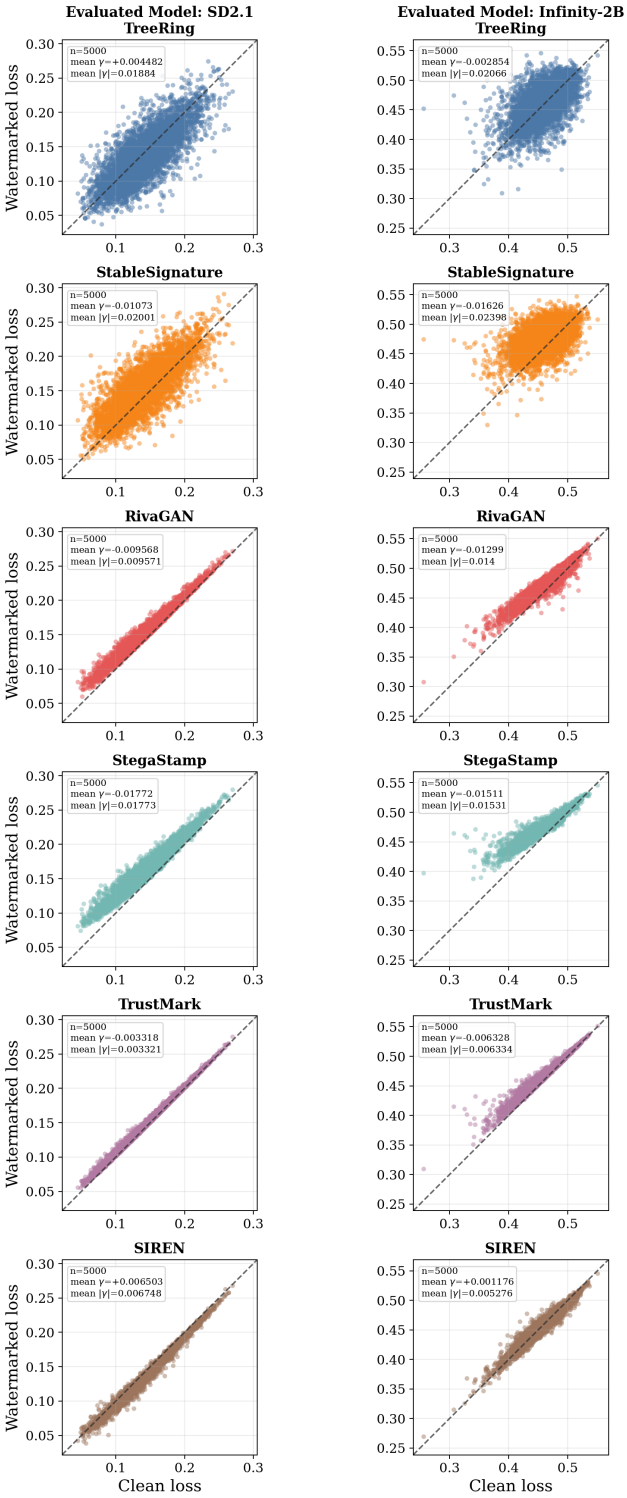


Figure 4: **The evaluation of watermark loss from SIREN for SD2.1 generated images.** We compute the loss difference on SD2.1 and Infinity-2B for 5000 pairs of clean and watermarked generated images.

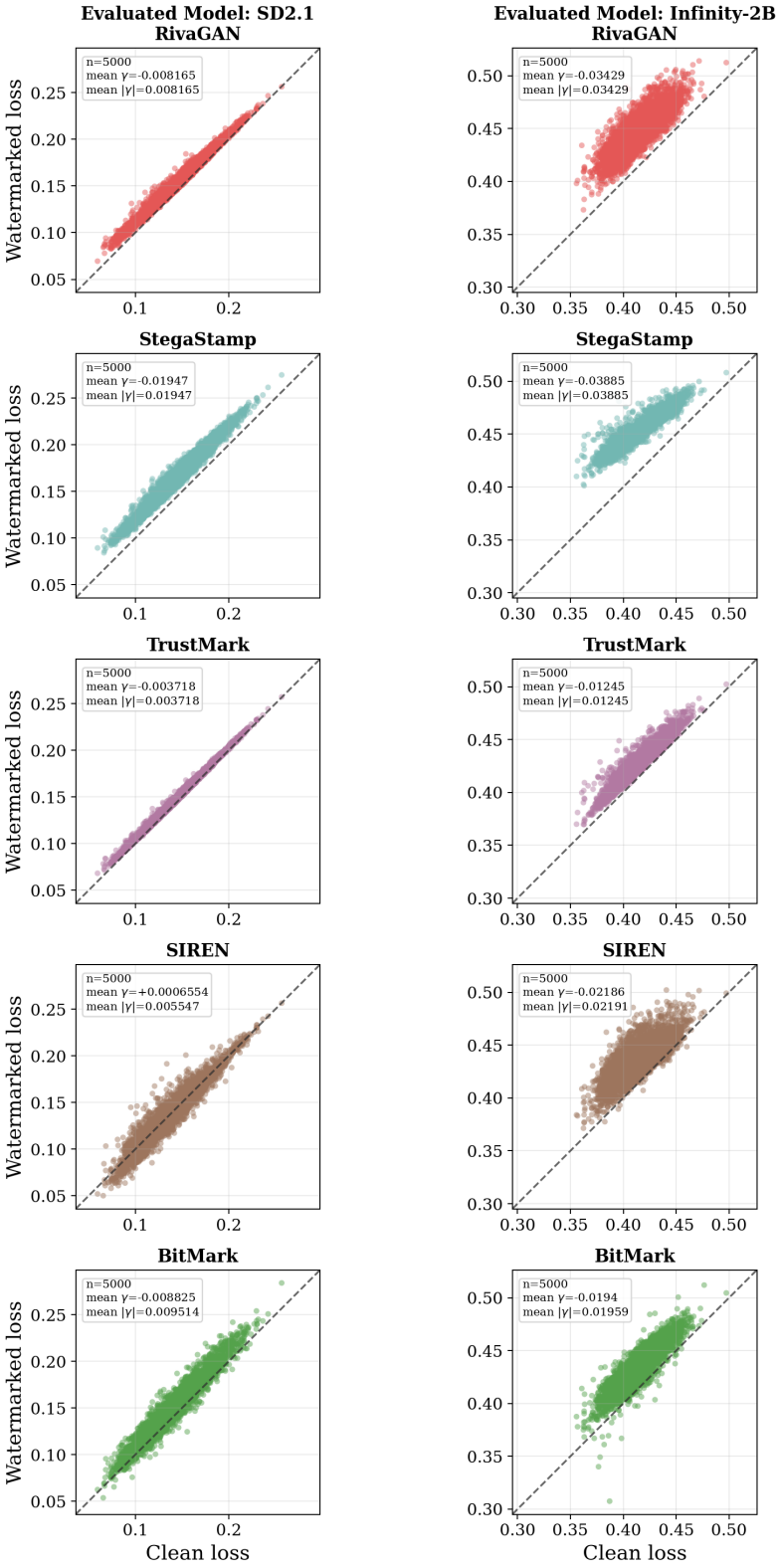


Figure 5: **The evaluation of watermark loss from SIREN for Infinity-2B generated images.** We compute the loss difference on SD2.1 and Infinity-2B for 5000 pairs of clean and watermarked generated images.

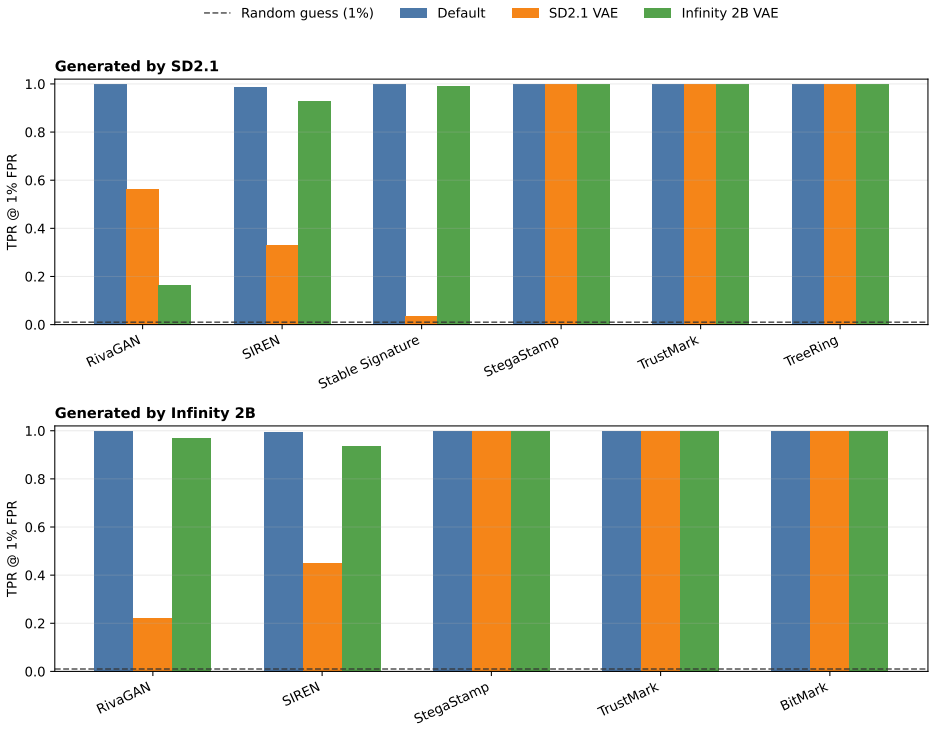


Figure 6: **Watermark robustness against VAE encoding-decoding.**

that we analyze operate in a latent space created by a VAE consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. To analyze the robustness of the watermark, we take 5000 watermarked samples, obtain their reconstruction via $\mathcal{D}(\mathcal{E}(x)) = x'$ and finally detect the watermark on x'. We plot the TPR@1%FPR of the watermark detection of x' in Figure 6 and find that StableSignature, RivaGAN and SIREN when embedded on SD2.1 images and reconstructed by the SD2.1-VAE are strongly weakened.

C.2 Radioactivity Ablations

Radioactivity Removal Robustness. In Table 9 we analyze the robustness of radioactivity in the same model setting against an additional fine-tuning step with clean data. Therefore we use clean generated data and the same hyperparameters parameters that we use in the iterative fine-tuning setting. For most watermarks we find that, even when making up 100% of the original fine-tuning data, they are removed after one additional cleaning step. Only BitMark can still be detected, as it influences the generation scheme of Infinity much stronger, due to it being embedded at every bit position across all scales and bits.

Adversarial Robustness. We additionally consider an adversary, that suspects *Alice* to embed a radioactive watermark signal, and that tries to remove it through augmentations. As the adversary must preserve utility as well as fast inference, we consider common image augmentations such as cropping to 90%, JPG compression to 75% quality, and horizontal flipping. We consider three different attack stages, namely (1) on the M_1 generated images before fine-tuning, (2) during fine-tuning and (3) on the M_2 generated images. We provide the results in Table 10, Table 11 and Table 12 respectively. We find that augmentations before or during fine-tuning are more effective than augmentations on the generated data after fine-tuning. Fine-tuning on the watermarked data enables the model to reproduce the watermark signal, whereas fine-tuning on the attacked images reduces the potential signal that can be picked up. This makes attacks before or during fine-tuning much more effective.

Learning Rate. The learning rate influences how strongly the model adapts to the training data and consequently impacts the strength at which the watermark is learned. In our experiments in

Table 9: **Clean fine-tuning removes radioactive traces in nearly all settings.** M_1 is the pretrained base. M_2 is fine-tuned on data generated by M_1, with the watermark fraction shown in brackets (%); M_3 further fine-tunes M_2 on clean data. For evaluation of this same model setting we leverage clean data from M_1. The reported values follow the same format as in Table 1.

Watermark	Setting	Model	M_2	M_3
BitMark (5%)	Supervised	Inf-2B	66	0.36
	Unsupervised	Inf-2B	78	0.16
BitMark (100%)	Supervised	Inf-2B	37	187
	Unsupervised	Inf-2B	34	194
StegaStamp (100%)	Supervised	Inf-2B	35	2.45
	Unsupervised	Inf-2B	36	27.73
SIREN (100%)	Supervised	Inf-2B	36	0.17
	Unsupervised	Inf-2B	38	0.04
StegaStamp (100%)	Supervised	SD2.1	37	0.57
	Unsupervised	SD2.1	42	0.71
SIREN (100%)	Supervised	SD2.1	95	1.69
	Unsupervised	SD2.1	208	0.08

Table 10: **Pre-training Robustness.** We apply augmentations on the data generated by M_1 before fine-tuning M_2. Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	Model	Attack	Setting	Fraction of Watermarked Data			
				5%	10%	25%	100%
BitMark	Inf-2B	Clean	Supervised	74	50	35	32
			Unsupervised	90	48	35	34
		Combined	Supervised	0.164	0.0688	635	75
			Unsupervised	0.321	0.907	0.0153	83

Table 13, we fix all hyperparameters and settings and fine-tune M_2 varying only the learning rate. We find that a higher learning rate results in radioactivity being detected with fewer samples up to a setting-specific maximum: beyond this point overall model utility degrades (see Figure 7).

Watermark Strength. The watermark strength controls the magnitude of the embedded signal and in turn the imperceptibility and robustness. In SIREN [23], the watermark encoder $\mathcal{E}_w$ first determines the watermark w based on an image x, with $w = \mathcal{E}_w$. This watermark is then finally embedded onto the image via $x_w = x + \beta \cdot w = x + \beta \cdot \mathcal{E}_w$, where β is the parameter that controls the watermark strength. BitMark [17] embeds the watermark at generation time when the next bit is predicted. It embeds a green bit list, e.g. $\{01, 10\}$, into the generated bit stream by slightly nudging the next predicted bit to take a specific value based on the previous bit. This is done by adding a constant δ to

Table 11: **During-training Robustness under Horizontal Flipping.** Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	M1	M2	Flip	Fraction of Watermarked Data		
				25%	50%	100%
RivaGAN	Inf-2B	Inf-2B	○	226	86	45
			•	0.32	191	69
BitMark	Inf-2B	Inf-2B	○	35	41	34
			•	35	45	34

Table 12: **Post-training Robustness.** We apply the attacks on the images generated by M_2. Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	Model	Attack	Setting	Fraction of Watermarked Data					
				1%	5%	10%	25%	50%	100%
BitMark	Inf-2B	Clean	Supervised	97	67	47	36	48	34
			Unsupervised	118	88	55	38	54	34
		Crop	Supervised	2.51e+04	364.804	353	52	112	34
			Unsupervised	3.934	78.839	687	55	93	34
		Horizontal Flip	Supervised	1.642	3.689	0.473	74	168	36
			Unsupervised	2.327	1.316	0.667	83	167	38
		JPG	Supervised	148	96	58	37	50	34
			Unsupervised	188	126	66	37	52	34
		Combined	Supervised	1.965	0.232	0.096	261	401	68
			Unsupervised	1.359	0.115	0.431	383	570	70

Table 13: **Larger learning rates increase radioactivity, up to a regime-specific threshold. M_2 is fine-tuned on data generated by M_1, with the watermark fraction shown in brackets (%) and vary only the learning rate.** The reported values follow the same format as in Table 1.

Watermark	Setting	Model	Learning Rate						
			6×10^{-6}	1×10^{-6}	1×10^{-5}	6×10^{-5}	1×10^{-4}	6×10^{-4}	6×10^{-3}
BitMark (5%)	Supervised	Inf-2B	–	225	88	–	112	66	1.14
	Unsupervised	Inf-2B	–	259	134	–	103	78	0.68
SIREN (50%)	Supervised	Inf-2B	–	194	56	–	70	50	69
	Unsupervised	Inf-2B	–	1.49	168	–	154	65	52
SIREN (100%)	Supervised	SD2.1	85	–	95	96	85	34	–
	Unsupervised	SD2.1	116	–	208	120	101	36	–

the logit of the next green bit, e.g. if the previous bit was 1, then the bias will be added to the logit predicting 0. The constant δ influences the strength of the watermark. Per default we use $\beta = 1.5$ and $\delta = 2$. In Table 14 we vary the watermark embedding strength from our default value and find that higher watermark strength can result in stronger radioactivity, as long as the image quality is not impacted negatively. Reducing the strength, on the other hand, can result in the watermark being no longer detectable in the downstream model at all.

Key Reuse. In all of our experiments, we actively reuse the same key in all images to amplify the signal of the watermark. To evaluate the impact of using multiple keys, we embed the StegaStamp watermark with Key A on 50% of the training images and 50% on the other training images, resulting in 100% watermarked data, but less overall watermark information. Our results in Table 15 show that while the watermark remains radioactive and detectable, more samples are required for detection. The per key signal is reduced.

Table 14: **Watermark strength impacts radioactivity.** We fine-tune M_2 at each watermark's default training configuration and vary the watermark's strength parameters. The reported values follow the same format as in Table 1.

Watermark	Setting	Model	Watermark Strength (% of default)			
			25%	50%	100% (default)	150%
BitMark (5%)	Supervised	Inf-2B	3	50.26	66	104
	Unsupervised	Inf-2B	0.13	0.39	78	175
SIREN (50%)	Supervised	Inf-2B	63.08	94	50	159
	Unsupervised	Inf-2B	2.1	123	65	34.43
SIREN (100%)	Supervised	SD2.1	3.18	163.02	95	68
	Unsupervised	SD2.1	0.11	7.21	208	67

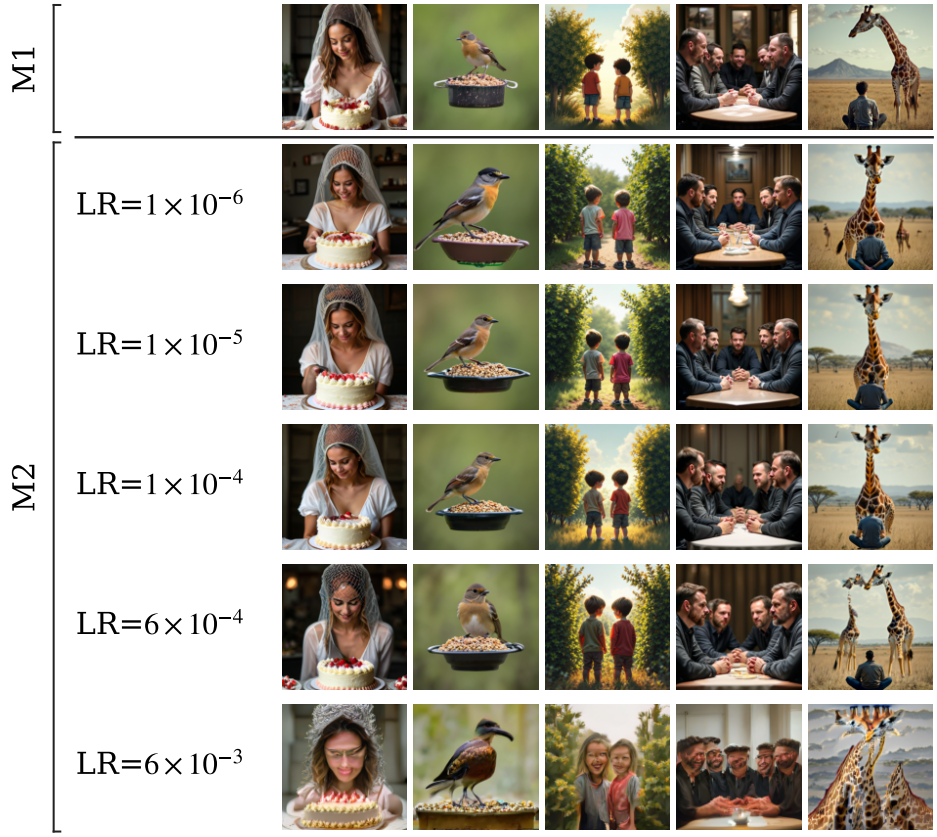


Figure 7: **An excessively high learning rate degrades model utility and weakens the watermark signal,** making radioactivity harder to detect.

Table 15: **Reusing the same key increases radioactivity.** We embed two distinct keys A and B onto half of the training data each and compute the radioactivity. We report the stopping time following Table 1.

Watermark	Model	Setting	100% Sup	100% Unsup
StegaStamp	SD2.1	Single Key	46	56
		Dual Key (Key A / Key B)	90 / 97	101 / 95
StegaStamp	Inf-2B	Single Key	33	34
		Dual Key (Key A / Key B)	40 / 39	44 / 43

Table 16: **The optimal number of fine-tuning epochs is setup-specific, over-training erodes the watermark.** We fine-tune M_2 with each watermark's default training configuration and vary the number of epochs. The reported values follow the same format as in Table 1.

Watermark	Setting	Model	Number Of Epochs					
			1	2	4	5	10	20
BitMark (5%)	Supervised	Inf-2B	88	66	136.12	–	–	–
	Unsupervised	Inf-2B	100	78	0.13	–	–	–
SIREN (50%)	Supervised	Inf-2B	121	50	155	–	–	–
	Unsupervised	Inf-2B	145	65	202	–	–	–
SIREN (100%)	Supervised	SD2.1	94	87	–	95	96	101
	Unsupervised	SD2.1	126	137	–	208	180	184

Table 17: **Ablation over effective batch size.** We fine-tune M_2 with each watermark's default training configuration and vary the effective batch size. The reported values follow the same format as Table 1.

Watermark	Setting	Model	Effective Batch Size			
			1	4	16	64
BitMark (5%)	Supervised	Inf-2B	66	66	186	67
	Unsupervised	Inf-2B	131	78	205	56
SIREN (50%)	Supervised	Inf-2B	55	50	142	68
	Unsupervised	Inf-2B	89	65	156	147
SIREN (100%)	Supervised	SD2.1	94	95	105	107
	Unsupervised	SD2.1	108	208	144	119

Fine-tuning Epochs. We study the effect of the number of fine-tuning epochs on the models radioactivity in Table 16. The results show that the epochs have very limited impact on the samples required to stop the e-process.

Batch Size. We analyze the impact of the training batch size on the watermark radioactivity. Larger batch sizes average the model gradients across more samples potentially diluting the watermark signal during training. However our results in Table 17 suggest that the watermark radioactivity is mostly robust to differing batch sizes. This indicates that the watermark signal is sufficiently detectable across the batch of samples.

Lower Watermarking Fractions. A central question for watermarking-based radioactivity detection is how small a fraction of watermarked data is sufficient to leave a detectable trace in the downstream model. To probe this regime, we extend the BitMark sweep on Inf-2B to fractions below the 1% setting reported in Table 1. As shown in Table 18, the number of samples required to stop the e-process grows sharply as the fraction decreases, and below 1% watermark fraction in the training data the signal is no longer significant under either supervised or unsupervised detection, neither at 0.1% (i.e., 5 images) nor at 0.02% (i.e., 1 image). This indicates a practical lower bound, for BitMark in this setting, of roughly 1% watermarked training data (i.e., 50 images) for reliable downstream detection.

Bit Capacity. To evaluate the impact of the bitcapacity on the radioactivity, we perform additional experiments using BitMark. By default, BitMark leverages a 2 bit pattern, embedding dense informa-

Table 18: **Extended BitMark watermark fraction ablation.** Continuation of Table 1, reporting BitMark detection at lower watermarked-data fractions.

Watermark	Setting	Model	Fraction of Watermarked Data								
			0%	0.02%	0.1%	1%	5%	10%	25%	50%	100%
BitMark	Supervised	Inf-2B	0.28	0.15	0.19	87	66	53	52	56	37
	Unsupervised	Inf-2B	0.31	0.05	0.24	132	78	47	46	43	34

Table 19: **Detection performance for different BitMark payload sizes.** Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$).

Watermark	Setting	Model	Target Model		
			M_2	M_3	M_4
BitMark (2-bit)	Supervised	Inf-2B	74	164	165
	Unsupervised	Inf-2B	90	82	236
BitMark (3-bit)	Supervised	Inf-2B	86	375	454
	Unsupervised	Inf-2B	141	193	786
BitMark (4-bit)	Supervised	Inf-2B	215	10200.0	2670.0
	Unsupervised	Inf-2B	396	630	3600.0

Table 20: Radioactivity detection without access to the secret detector. We train a ResNet-50 on paired clean and watermarked data. We report the results under the *RADIUM* setting in (). The reporting scheme follows Table 1.

Watermark	**Model**	**25%**	**50%**	**100%**
SIREN	Inf-2B	271 (223)	53 (50)	37 (36)
SIREN	SD2.1	616 (0.06)	544 (0.05)	51 (95)
StegaStamp	Inf-2B	244 (5.89)	107 (249)	32 (35)
StegaStamp	SD2.1	2.55 (0.14)	8788.36 (0.09)	35 (37)
BitMark	Inf-2B	99 (52)	91 (56)	31 (37)

tion across the image. We extend this pattern to 3 and 4 bits respectively, resulting in more embedded information, however it becomes less dense due to the increase in the context window size. We report the results in Table 19, following Table 2, fine-tuning M_{i+1} with 5% watermarked data. As also observed by Sander et al. [31], a larger context window drastically reduces the radioactivity of the watermark.

Detection Without Key. While *RADIUM* is designed for watermark owners with access to the watermark detector, we additionally evaluate the detection performance without access to the correct watermarking key or detector. To overcome this limitation, we assume an upper bound setting, where we have access to clean and watermarked data from the same distribution. We then train a ResNet-50 to differentiate between the clean and watermarked data, returning a score of 1 for watermarked images and 0 for clean images. Our results in Table 20 show that under these strong assumptions a blind detector can replace the original watermark detection scheme and detect the radioactive traces in the M_2 model. We find that the blind detector partially improves over the watermark owner detection, as the blind detector is trained on a single key, whereas the watermark owner can detect arbitrary keys.

Inheritance vs. Watermark. Our fine-tuning setting in both the single-hop as well as multi-hop cases initialize M_{i+1} by its predecessor M_i. This opens the question, wether radioactivity in the case of $i \geq 2$ is a property of the model weights, or a property of the data generated by M_i. To evaluate this, we train M_3, but instead of initializing the model with its prior M_2 (fine-tuned on watermarked data from M_1), we initialize it with the original model architecture M_1. Our results in Table 21 show that for both model and watermark combinations, the initialization with M_1 has limited impact on the radioactivity of the watermarks (in 100% watermarked data used for fine-tuning: slightly for Inf-2B and more for SD2.1). This is to be expected, as the model now no longer carries the initial signal it obtained from the originally watermarked data, but still learns the weaker signal of the generations from M_2. SIREN becomes detectable, one potential reason is the limited perturbations of the images due to re-initialization with non-fine-tuned weights. The results show that while inheritance through initialization is one factor of radioactivity, the core impact stems from the watermark itself.

Table 21: **Radioactivity is a property of the data and the model. We initialize M_3 with the original clean weights M_1 and fine-tune on data generated by M_2. We report the baseline results from Table 2 in "()".**

Watermark	**Model**	**Setting**	M_3 **100%**
StegaStamp	SD2.1	Supervised	42 (37)
StegaStamp	SD2.1	Unsupervised	61 (42)
SIREN	SD2.1	Supervised	203 (7.35)
SIREN	SD2.1	Unsupervised	456 (14.74)
BitMark	Inf-2B	Supervised	34 (34)
BitMark	Inf-2B	Unsupervised	33 (33)
StegaStamp	Inf-2B	Supervised	35 (33)
StegaStamp	Inf-2B	Unsupervised	33 (33)

Table 22: **Detection performance for SD3.5.** We train M_2 on different fractions of watermarked data % generated by M_1. Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	Setting	M_1	M_2	Fraction of Watermarked Data			
				0%	25%	50%	100%
StegaStamp	Supervised	SD3.5	SD3.5	0.11	0.01	0.12	99
	Unsupervised	SD3.5	SD3.5	0.10	0.03	0.01	169
RivaGAN	Supervised	SD3.5	SD3.5	0.10	880	492	135
	Unsupervised	SD3.5	SD3.5	0.02	1117.07	799	179
SIREN	Supervised	SD3.5	SD3.5	0.10	249	139	46
	Unsupervised	SD3.5	SD3.5	0.01	222	116	38
TrustMark	Supervised	SD3.5	SD3.5	0.06	0.04	0.26	0.05
	Unsupervised	SD3.5	SD3.5	0.01	0.02	0.02	0.02
BitMark	Supervised	Inf-2B	SD3.5	0.10	137	48	38
	Unsupervised	Inf-2B	SD3.5	0.30	511	62	35

C.3 Larger Models

We additionally explore the radioactivity of image watermark on larger diffusion model architectures, namely Stable Diffusion 3.5-medium with 2B parameters and FLUX.1-dev [21] with 12B parameters. We report our results in Table 22 and Table 23, following the convention of Table 3 and provide the hyperparameters in Table 4. Our results show similar behavior between the two large model architectures, with SIREN and RivaGAN being detectable when making up 25% or more of the training data and StegaStamp being radioactive, but only under higher data fractions. TrustMark remains non-radioactive, as we already found for Infinity-2B and SD2.1. Comparing the results for the larger DMs and SD2.1 from Table 1, we find that the behavior of the larger models more closely aligns to the behavior of Infinity-2B, suggesting that model size could be a factor influencing the radioactivity of watermarks.

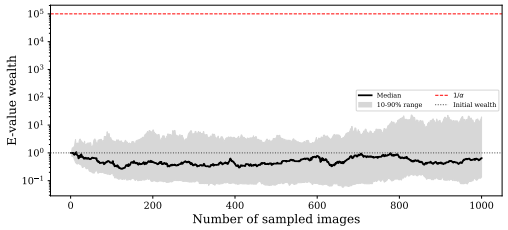
C.4 Watermark Control

Do we remove this section, given that we deploy whitening or do we still incorporate it to analyze what happens without whitening?

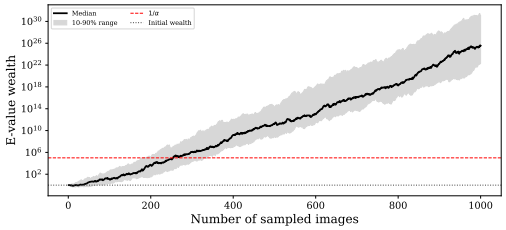
In this section we analyze the assumption that the watermarks behave the same across clean models. Therefore we perform a hypothesis test using e-processes. We find that for two watermarks, SIREN [23] and BitMark [17], this assumption is broken, when the architecture between M_1 and M_2 differ. Therefore we use the model M_2 that was trained on only **clean generated data** and compute the watermark scores and compare these against the clean data generated by M_1.

Table 23: **Detection performance for FLUX.1-dev.** We train M_2 on different fractions of watermarked data % generated by M_1. Cells report the stopping time (in number of paired observations) at which the e-process crossed the rejection threshold $1/\alpha = 10^5$ (significance level $\alpha = 10^{-5}$); faded cells report the final e-value of runs that did not reject.

Watermark	Setting	M_1	M_2	Fraction of Watermarked Data			
				0%	25%	50%	100%
StegaStamp	Supervised	FLUX.1	FLUX.1	0.05	0.39	79	34
	Unsupervised	FLUX.1	FLUX.1	0.03	0.34	103	55
RivaGAN	Supervised	FLUX.1	FLUX.1	0.01	861	701	56
	Unsupervised	FLUX.1	FLUX.1	2.02	565	368	58
SIREN	Supervised	FLUX.1	FLUX.1	0.06	464	77	39
	Unsupervised	FLUX.1	FLUX.1	0.07	786	96	39
TrustMark	Supervised	FLUX.1	FLUX.1	0.04	0.03	0.02	0.72
	Unsupervised	FLUX.1	FLUX.1	0.02	0.12	0.14	0.04

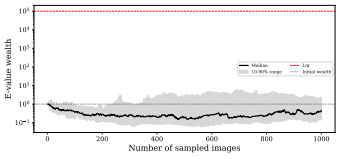


(a) M_1: Infinity-2B; M_2: Infinity-2B, Watermark: BitMark

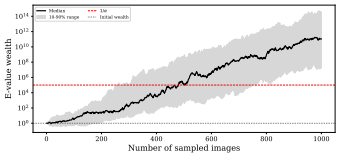


(b) M_1: Infinity-2B; M_2: SD2.1, Watermark: BitMark

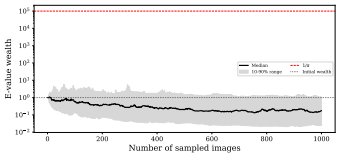
Figure 8: **Bias in BitMark.** We find that BitMark is slightly biased on the tested cross-architecture distribution. While the per-sample test is still insignificant for SD2.1 generated samples, aggregating this bias can lead to false positive detection with our test. TODO: We should fix this figure



(a) M_1: Infinity-2B; M_2: Infinity-2B, Watermark: SIREN



(b) M_1: Infinity-2B; M_2: SD2.1, Watermark: SIREN



(c) M_1: SD2.1; M_2: SD2.1, Watermark: SIREN

Figure 9: **Bias in SIREN.** We find that SIREN is slightly biased on the tested cross-architecture distribution. While the per-sample test is still insignificant for SD2.1 generated samples, aggregating this bias can lead to false positive detection with our test.

Table 24: FID and CLIP scores for BitMark and StegaStamp for different watermark fractions in M_2. Lower FID and higher CLIP scores are better.

Watermark	Metric	M1 Clean	M1 Watermarked	0%	1%	5%	10%	25%	100%
BitMark (Inf-2B)	FID ↓	40.68	38.10	44.91	45.98	44.53	42.92	43.41	42.17
	CLIP ↑	31.29	31.30	31.37	31.43	31.41	31.38	31.42	31.42
StegaStamp (SD2.1)	FID ↓	25.60	27.04	34.56	35.30	34.74	34.78	35.16	35.01
	CLIP ↑	31.37	31.45	31.33	31.41	31.47	31.48	31.42	31.46

Table 25: FID and CLIP scores for different BitMark and StegaStamp in the iterative training. Lower FID and higher CLIP scores are better.

Watermark	Metric	M1 Clean	M1 Watermarked	M2	M3	M4	M5	M6
StegaStamp (Inf-2B)	FID ↓	40.68	39.98	44.20	45.22	46.06	47.69	52.14
	CLIP ↑	31.29	31.17	31.32	31.46	31.47	31.48	31.39
BitMark (Inf-2B)	FID ↓	40.68	38.10	42.17	43.68	43.79	44.47	46.64
	CLIP ↑	31.29	31.30	31.42	31.57	31.75	31.69	31.58
StegaStamp (SD2.1)	FID ↓	25.76	27.04	35.01	61.40	110.86	172.65	214.96
	CLIP ↑	31.37	31.45	31.46	30.81	29.02	26.34	23.61

C.5 Visual Quality Evaluation

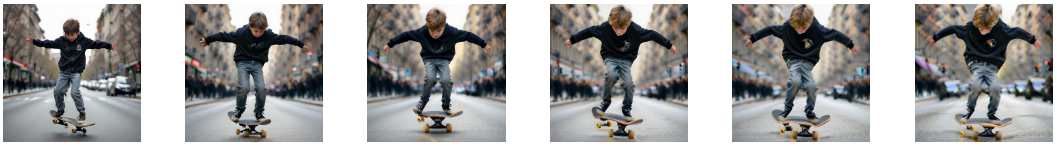
C.5.1 Model Quality

We evaluate the impact of training on watermarked data on model quality under FID [13] and CLIP score [12]. Therefore we generate 5k MS-COCO images, using validation prompts not seen during training and compute the CLIP score, as well as the FID against the matched 5k natural MS-COCO images. Our results in Table 24 show, that training on watermarked data reduces the quality of M_2 by reinforcing existing model biases. However, detectability in M_2 is not solely driven by image quality. Rather image quality and detectability are driven by a confounder, training on generated content. We further evaluate the image quality across multiple model iterations in Table 25. Our results show that especially for SD2.1 image quality degrades drastically with higher model iterations, whereas in Infinity quality degrades more gracefully.

C.5.2 Example Images

We provide examples of the generated images of the models in the iterative training experiment. Therefore we generate an image with all trained models M_i for Infinity-2B trained on BitMark in Figure 10 and for SD2.1 trained on the SIREN watermark in Figure 11. Our results show that with progressing iterations the image quality drops, as the errors and perturbations of the generated images propagate through the model. Since M_{i+1} is the model M_i trained on its own data, errors propagate and can not be fixed, resulting in drastic decay of model quality.

D Add



(a) M_1 (b) M_2 (c) M_3 (d) M_4 (e) M_5 (f) M_6

Figure 10: **Impact of multiple training-generation runs on image quality.** We train consecutive models $M_2 \dots M_6$ on 100% of generated data that were watermarked with the radioactive BitMark.

(a) M_1 (b) M_2 (c) M_3

Figure 11: **Impact of the radioactive decay for the SIREN watermark on image quality.** Each successive model is trained on 100% of the data of the preceding model. Training SD2.1 on radioactive SIREN images results in a destruction of the watermark and degradation of the image quality.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose a new method to measure the radioactivity of existing image watermarks.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a discussion of the limitations in Appendix B.4.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: -

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the code to reproduce our results and the paper contains detailed information about the experimental setup.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide full information about the experimental setup and the code in the supplementary material. Additionally we provide anonymized code here.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the information of hyperparameters, how they were chosen and an ablation thereof in the paper.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our results are based on statistical tests that provide information about the detection significance.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

Justification: We provide information about the compute environment, however given the amount and variety of experiments estimating the full compute and time for each experiments exactly is not possible.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully follow the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the impact of our work in the conclusion.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release new models or new datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the owners of all existing assets that we use and we respect their licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide code for reproduction of our experiments and the usage of our framework.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: -

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: -

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:-