RETHINKING METRICS FOR PLUG-AND-PLAY IN-PROCESSING IMAGE WATERMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

In-processing watermarking schemes have made it possible to add a truly invisible watermark to images generated by image-generative models. Tree Ring Watermarking(Wen et al. (2023))introduced a plug-in scheme that embeds watermarks by modifying the Gaussian sampling space of diffusion models. However, since these watermarks influence the generative process, the resulting watermarked image is perceptually different from a non-watermarked image. This poses a drawback in task-specific image generation, where minimal perceptual distortion is crucial. A truly plug-and-play scheme should minimize perceptual changes in the watermarked image while maintaining its effectiveness. In this study, we extend prior work on watermarking the Gaussian sampling space, focusing on quantifying the resulting changes in the generated image.

1 INTRODUCTION

The rapid advancements in text-conditioned image generation models such as Stable Diffusion and Midjourney and their extensive use in media, entertainment, gaming, and fashion have led to the need to copyright the generated images. Earlier digital image watermarking techniques involved converting images to the frequency domain through operations like Discrete Wave Transform (Xia et al. (1998)), Discrete Cosine Transform (Al-Haj (2007)), Singular Value Decomposition (Al-Haj (2007)), Fast Fourier Transform embedding watermarks (Pun (2006)) in the high-frequency space. However, such watermarks in the frequency domain were highly susceptible to attacks involving jpeg compression.

In-processing watermarking schemes were later introduced to modify or manipulate the image generation process, embedding the watermark directly into the generated image. One such generative model, Stable Diffusion (Rombach et al. (2021)) (with DDIM (Song et al. (2022)) sampling) works by iteratively denoising a noisy image using the textual description provided in a user prompt. The forward diffusion process in DDIM progressively adds noise until the image becomes pure white noise, requiring the backward diffusion process to start from an initial vector sampled from a Gaussian distribution for optimal reconstruction. Tree-Ring watermarking(Wen et al. (2023)) leverage this step by modifying the frequency domain of the initial sample vector to embed a watermark, leading to perceptible changes in the content of the generated image without introducing any artifacts.

The authors evaluated these changes using Fréchet Inception Distance (FID), which measures the distributional difference between watermarked images and real reference images, and CLIP Score, which measures how well the watermarked image aligns with the given prompt. A watermarking scheme for image generation can be considered truly plug-and-play only if it does not affect the perceptual features of the generated images. Beyond measuring how close a watermarked image is to a real image, a 'plug-in'ness metric should also be used to quantify the perceptual difference between the watermarked and non-watermarked images. (Figure 1) . In this study, we discuss two different types of metrics to quantify 'plug-in'-ness or the plug-and-play characteristics of a watermarking scheme, specifically from the perspective of TreeRing watermarks.

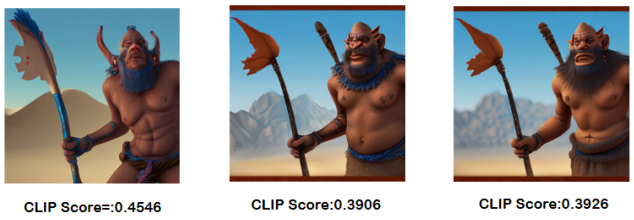

Figure 1: Left: Image generated with Tree-Ring$_{Ring}$, Center: Image generated without any watermarking, Right: Image generated with Spatial$_{rand}$. The left image has the highest CLIP score, indicating the strongest alignment between the generated image and the prompt. However, to the naked eye, it appears perceptually different from the vanilla image. The right image, despite having a lower CLIP score, is perceptually more similar to the vanilla image.

2 Proposed Methodology

The latent sampling space of DDIMs follows a Gaussian distribution. Any modification to the latent vector alters this distribution, which subsequently impacts the generated image. In theory, a smaller deviation in the sampling distribution should result in a smaller perceptual change in the generated image. However, the backward diffusion process may exhibit differential sensitivity to different types of deviations, even when the magnitude of the deviation (as measured by a chosen metric) remains constant within the sampling space. To summarize, the observed changes in the watermarked image can be divided into two distinct degrees, and we introduce two metrics to quantify them:

1. For the lower degree **Mathematical deviation**(md_w) captures the first-order effect of a watermarking scheme by quantifying the deviation introduced in the underlying sampling distribution of the diffusion model. md_w must be designed to provide a mathematically grounded measure of how a watermark alters the distribution, focusing on shifts in statistical properties.
2. For the higher degree **Perceptual difference**(pd_w) captures the second-order effect by measuring the perceptual impact of the watermark in the generated image. pd_w must incorporate human visual perception into the analysis, addressing how the watermarking process affects the image's visual fidelity and introduces distortions that can be perceived by human observers.

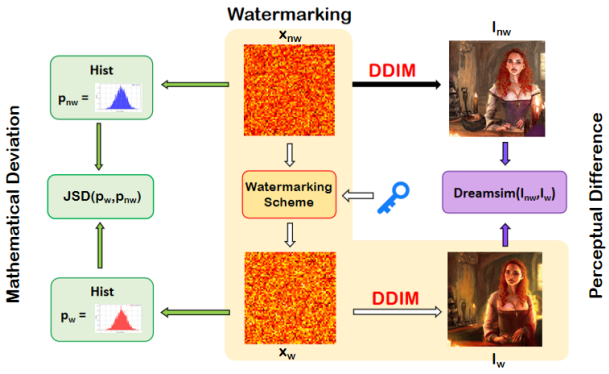

Figure 2: The Proposed evaluation framework to measure the plug-and-play characteristics for watermarking schemes operating on the sampling space of diffusion models.

Mathematical deviation(md_w): For the mathematical deviation, we quantify the change in the sampling distribution using the Jensen-Shannon (JS) divergence Menéndez et al. (1997). The JS divergence is a smoothed and symmetric variant of the Kullback-Leibler (KL) divergence. Unlike KL divergence, JS divergence is bounded, making it a more stable metric for evaluating distributional shifts in the sampling space Arjovsky et al. (2017). Specifically, we compute the divergence

between the probability distribution of the sampling space before (P_{nw}) and after injecting the watermark (P_w), as defined by:

$$JS(P_w \parallel P_{nw}) = \frac{1}{2}KL(P_w \parallel M) + \frac{1}{2}KL(P_{nw} \parallel M)$$

$$\text{where } M = \frac{1}{2}(P_w + P_{nw})$$

$$\text{and } KL(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

To estimate the probability distributions P_w and P_{nw}, we generate histograms representing the frequency vs intensity of pixel values over 1000 sampled latent vectors and calculate JSD across 6 different watermarking patterns (P_w) in Table 1. Please refer to Appendix B for the detailed watermarking patterns.

Pattern Injected	JS Divergence
Tree-Ring$_{ring}$	5.38×10^{-3}
Tree-Ring$_{zero}$	4.47×10^{-3}
Tree-Ring$_{rand}$	6.40×10^{-3}
Spatial$_{ring}$	8.76×10^{-4}
Spatial$_{zero}$	1.58×10^{-3}
Spatial$_{rand}$	8.39×10^{-5}

Table 1: We find the Jensen-Shannon divergence between the original distribution and watermarked distribution across 6 watermarking pattern- Tree-Ring$_{ring}$, Tree-Ring$_{zero}$, Tree-Ring$_{rand}$ which operate on the frequency domain,and Spatial$_{ring}$ Spatial$_{rand}$ and Spatial$_{zero}$ which directly operate on the spatial domain.

Pattern Injected	DreamSim	Clip Score (Vanilla: 0.3607)
Tree-Ring$_{ring}$	0.1935	0.3617
Tree-Ring$_{zero}$	0.1841	0.3576
Tree-Ring$_{rand}$	0.1958	0.3550
Spatial$_{ring}$	0.1965	0.3586
Spatial$_{zero}$	0.1041	0.3607
Spatial$_{rand}$	0.0986	0.3612

Table 2: For DreamSim, a higher score indicates that the watermarked image is more perceptually different, while a lower score means it is more similar to a Non watermarked image. In contrast, for CLIP score, a higher value signifies stronger alignment between the watermarked image and the prompt, whereas a lower score indicates weaker alignment.

Perceptual Difference(pd_w): For measuring perceptual difference, we use DreamSim (Fu et al. (2023)), which learns a similarity metric for embeddings of ViT-L/14 based on human judgments, instead of hand-crafted distance functions. The authors demonstrated that DreamSim outperforms other perceptual metrics, such as LPIPS (Heusel et al. (2017)) and SSIM (Wang et al. (2004)), in capturing human-perceived similarity. Another approach would be to use Fréchet Inception Distance (FID) (Rombach et al. (2021)) to compare the distributions of watermarked and non-watermarked images. However, FID has significant drawbacks (Jayasumana et al. (2024)) - it is sample inefficient, requiring a large number of samples to converge to a stable value, and it is biased toward the model used for feature extraction. In contrast, DreamSim is an instance-based metric that evaluates image pairs independently, without the need for distributional comparisons. We compare DreamSim with CLIP Score (Radford et al. (2021)), a widely used instance-based metric to evaluate perceptual quality of watermarking Schemes in Image generative models. We find the average DreamSim Score and CLIP score across 6 different watermarking patterns for 100 images in Table 2.

3 RESULTS

For mathematical deviation, in Table 1 we observe that watermarking schemes operating in the frequency domain-Tree-Ring$_{ring}$, Tree-Ring$_{zero}$, Tree-Ring$_{rand}$ induce greater changes to the sampling distribution compared to watermarking schemes in the spatial domain-Spatial$_{ring}$ Spatial$_{rand}$ and Spatial$_{zero}$. Among spatial watermarking patterns, Spatial$_{rand}$, which uses a fixed random Gaussian noise as a key, results in the least JS divergence from the original distribution. This is consistent with the fact that replacing a portion of a Gaussian distribution with another independent Gaussian (with the same mean and variance) results in another Gaussian with the same mean and variance.

Figure 3: Generated image for spatial ring watermark. We observe high distortions in the middle of the image after watermarking leading to a higher Perceptual difference.

For perceptual difference, in Table 2 we find that Spatial$_{rand}$ and Spatial$_{zero}$ exhibit greater perceptual similarity to the non-watermarked image compared to other schemes. Additionally, the CLIP score across all models remains close to the original CLIP score, indicating that each scheme preserves text correlation. However, the difference in DreamSim with respect to the non-watermarked image is significantly more pronounced. Figure 1 further highlights how the CLIP score can be misleading in the context of watermarking, as it may increase even when a watermark introduces significant changes to the content of the image. This suggests that DreamSim serves as a more effective metric for evaluating the impact of a watermarking scheme, compared to CLIP score.

Interestingly, the Spatial Ring pattern had a JS divergence of $8.76 * 10^{-4}$ indicating a relatively low mathematical deviation, yet it resulted in a DreamSim score of $1.965 * 10^{-1}$ indicating a significantly greater perceptual difference. This discrepancy may be explained by the hypothesis that a well defined ring pattern in the sampling space has a strong influence on the backward diffusion process, leading to distortions(Figure 3) or large perceptual changes in the generated image.

4 CONCLUSION AND FUTURE WORK

We identified the need for metrics to assess the plug-and-play characteristics of watermarking schemes operating on the sampling distribution of diffusion models. To address this, we proposed two distinct metrics: Mathematical Deviation (md_w) to capture the first-order effect on the sampling distribution, and Perceptual Difference (pd_w) to measure the second-order effect on the generated image. The combination of these metrics offers a comprehensive evaluation, balancing mathematical rigor with perceptual relevance. This framework can be extended to evaluate current and future works in invisible image watermarking, and we look forward to exploring different combinations of md_w and pd_w in future studies.

REFERENCES

Ali Al-Haj. Combined dwt-dct digital image watermarking. *Journal of Computer Science*, 3(9), 09 2007. doi: 10.3844/jcssp.2007.740.746.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. URL https://arxiv.org/abs/1701.07875.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. URL https://arxiv.org/abs/2306.09344.

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 (eds.), *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.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024. URL https://arxiv.org/abs/2401.09603.

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. ISSN 0016-0032. doi: https://doi.org/10.1016/S0016-0032(96)00063-4. URL https://www.sciencedirect.com/science/article/pii/S0016003296000634.

Chi-man Pun. A novel dft-based digital watermarking system for images. In *2006 8th international Conference on Signal Processing*, volume 2, 2006. doi: 10.1109/ICOSP.2006.345581.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023. URL https://arxiv.org/abs/2305.20030.

Xiang-Gen Xia, Charles G. Boncelet, and Gonzalo R. Arce. Wavelet transform based watermark for digital images. *Opt. Express*, 3(12):497–511, Dec 1998. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-3-12-497.

A Experimental setting

We adopt the same experimental setup as used in the Tree Ring watermarking (Wen et al., 2023), employing Stable Diffusion-v2, a state-of-the-art open-source latent text-to-image diffusion model. Image generation is performed with 50 inference steps, using a default guidance scale of 7.5 and an empty prompt for DDIM inversion to simulate the scenario where the image prompt is unknown during detection. The watermark radius r is set to 10. The prompts are sourced from the Hugging Face dataset Gustavosta/Stable-Diffusion-Prompts. All experiments are conducted on a single NVIDIA RTX 2060 GPU. The code is built upon the implementation from Yuxin Wen's Tree-Ring Watermark repository.

B Watermark Patterns

We use 6 watermarking Patterns: Tree-Ring$_{ring}$, Tree-Ring$_{zero}$, Tree-Ring$_{rand}$ which were introduced in the original paper and additionally we add Spatial$_{ring}$ Spatial$_{rand}$ and Spatial$_{zero}$ for a better comparison. Across all the patterns the mask is designed as a circular region. In Tree-Ring$_{ring}$,

Tree-Ring$_{zero}$, Tree-Ring$_{rand}$ the noise vector is transformed to the frequency domain before replacing the pixels of it inside the circular mask with the key. In Tree-Ring$_{ring}$, Tree-Ring$_{zero}$, Tree-Ring$_{rand}$ the noise vector is transformed to the frequency domain before replacing the pixels of it inside the circular mask with the key.

- Tree-Ring$_{zero}$, Spatial$_{zero}$: The key is initialized as an array of zeros, which guarantees invariance under shifts, cropping, and dilations.
- Tree-Ring$_{rand}$, Spatial$_{rand}$: The key is sampled from a fixed Gaussian distribution, preserving the i.i.d. Gaussian property of the original Fourier modes in the noise array.
- Tree-Ring$_{ring}$, Spatial$_{ring}$: The key consists of multiple concentric rings, each with a constant value, ensuring rotational invariance of the watermark.

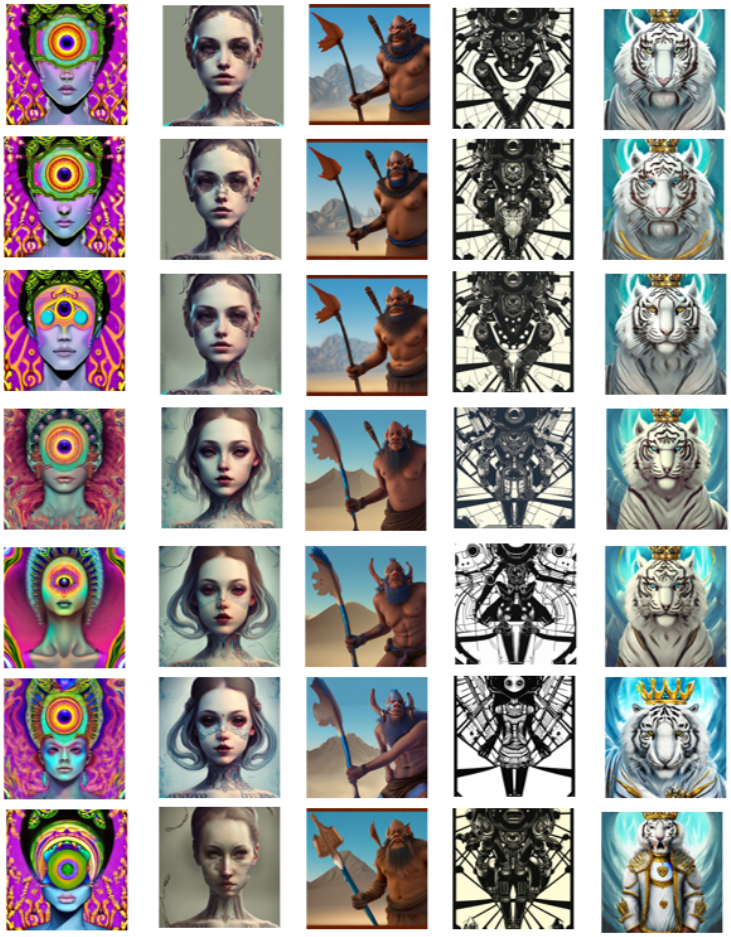

Figure 4: Provide more examples of generated images, arranged from top to bottom in the following order: Non-watermarked, Spatial$_{rand}$, Spatial$_{zero}$, Tree-Ring$_{zero}$, Tree-Ring$_{ring}$, Tree-Ring$_{rand}$, and Spatial$_{ring}$. This sequence follows an ascending order of their perceptual difference.