Reproducibility Study of "Attack-Resilient Image Watermarking Using Stable Diffusion"

Anonymous authors
Paper under double-blind review

Abstract

This paper presents a reproducibility study and robustness extension of the paper 'Attack-Resilient Image Watermarking Using Stable Diffusion' by Zhang et al(2024), which proposes a stable-diffusion-based framework for attack-resilient image watermarking. While successfully replicating the original method's core claims—achieving >90% watermark detection rate (WDR) against diffusion-based regeneration attacks and across MS-COCO, DiffusionDB, and WikiArt datasets—we identify critical vulnerabilities under novel adversarial and geometrically asymmetric attack paradigms. Our extended analysis demonstrates that gradient-based adversarial perturbations reduce ZoDiac's WDR, a threat model absent in prior evaluations. We also investigate non-circulary symmetric attacks achieving WDR below 65%. We also investigate a new loss function to mitigate these limitations. Despite these enhancements, composite attacks combining adversarial noise with other methods reduce WDR to near-zero, exposing vulnerabilities in multi-stage offensive pipelines. Our implementations can be found on Anonymous Github .

1 Introduction

The rapid advancement of generative AI has heightened the need for strong image watermarking techniques to verify content authenticity and counter AI-generated forgeries. Traditional watermarking methods, such as frequency-domain embeddings (DCT, DWT) and least-significant-bit (LSB) manipulation, were designed to withstand standard distortions like JPEG compression and Gaussian noise. However, they struggle against modern diffusion-based regeneration attacks, which use latent-space purification to erase watermarks. Early neural approaches, including RivaGAN and StegaStamp, improved resilience through adversarial training and spatial transformer networks. Yet, their pixel-space embeddings remain vulnerable to pipeline-aware attacks that exploit diffusion models' iterative denoising to remove watermarks . This weakness arises because these methods operate in pixel space, making watermarks susceptible to latent-space purification.

Recent diffusion-based techniques aim to address this issue. Tree-Ring watermarks, for example, encode concentric ring patterns into the initial noise vectors of synthetic images. By leveraging the deterministic inversion property of diffusion models, these watermarks can be recovered from generated outputs. Embedding patterns in the Fourier domain and utilizing the model's latent-space dynamics allows Tree-Ring to achieve rotational invariance and resist individual attacks. However, this approach only applies to synthetically generated images, leaving real-world content unprotected. Additionally, its reliance on isotropic ring patterns makes it vulnerable to asymmetric transformations like irregular rotations. Its static design also lacks defenses against composite attacks that combine geometric distortions with purification-based techniques. Other approaches, such as Stable Signature, fine-tune diffusion decoders to embed watermarks but require extensive retraining on large datasets, making them resource-intensive and reducing practicality.

ZoDiac addresses these gaps through a novel framework that integrates pre-trained stable diffusion models with DDIM inversion to embed imperceptible watermarks into both synthetic and real-world images. The method capitalizes on the bidirectional nature of diffusion models: it maps an input image into a latent vector via inversion, injects a ring-shaped watermark into the Fourier domain of this latent space, and reconstructs the image while preserving fidelity. Unlike pixel-space methods, ZoDiac operates in the latent space where diffusion models inherently resist purification attacks—iterative denoising during generation reinforces watermark persistence. ZoDiac explicitly aligns the injected watermark with its retrieved version in Fourier space, countering latent-space distortions caused by augmentations in the pixel-space. This contrasts with Tree-Ring's static synthetic-only embeddings, which lack such alignment mechanisms.

As depicted in Figure 2, the method maps an input image x_0 to a latent vector Z_T via:

$$Z_{t-1} = \alpha_{t-1} \left(\frac{Z_t - (1 - \alpha_t)\epsilon_\theta(Z_t, t)}{\alpha_t} \right) + (1 - \alpha_{t-1})\epsilon_\theta(Z_t, t) \tag{1}$$

where α_t defines the noise schedule and ϵ_θ is the pre-trained denoiser. A ring-shaped watermark W is injected into the Fourier domain of Z_T, optimized via a multi-term loss:

$$L = \underbrace{\|\hat{x}_0 - x_0\|^2}_{\text{Reconstruction}} + \lambda_s L_{\text{SSIM}} + \lambda_p L_{\text{Perceptual}} \tag{2}$$

ZoDiac enhances practicality through adaptive image blending, dynamically balancing watermark robustness and visual quality. By mixing the watermarked image $\hat{x}_0$ with the original x_0 via a tunable factor γ, it achieves SSIM thresholds without compromising detectability. This adaptability ensures resilience against multi-stage attacks while maintaining imperceptibility—a critical advance over StegaStamp and RivaGAN, which degrade under latent-space perturbations.

2 Scope of Reproducibility

The scope of the reproducibility study validates ZoDiac's core claims while rigorously stress-testing its limitations under novel attack paradigms. Our reproducibility efforts focus mainly on validating these four core claims of the original paper:

- **Claim 1 :** ZoDiac demonstrates a watermark detection rate (WDR) exceeding 98 and a false positive rate (FPR) below 6.4 across MS-COCO, DiffusionDB, and WikiArt datasets, outperforming state-of-the-art watermarking methods.
- **Claim 2 :** ZoDiac remains resilient to diverse attack categories, including traditional attacks (such as JPEG compression and gaussian blurring), stable diffusion-based regeneration attacks, where most other methods fail and rotational attacks to some extent.
- **Claim 3 :**Unlike prior methods like Tree-Ring or Stable Signature, ZoDiac can watermark both real-world and synthetic images without requiring retraining of the stable diffusion model, making it highly practical for diverse applications and real world deployment.
- **Claim 4 :** ZoDiac achieves imperceptible watermarks with image quality metrics such as SSIM > 0.92, ensuring minimal visual degradation while maintaining robustness against attacks. ZoDiac ensures a fair tradeoff between watermark detection and maintaining image quality.

3 Methodology

3.1 Description of Methods

The ZoDiac framework methodology comprises of three key components: latent-space watermark injection via DDIM inversion, Fourier-domain embedding for geometric resilience, and adaptive image enhancement to balance detectability and visual fidelity. Below, we detail each component as described in the original paper.

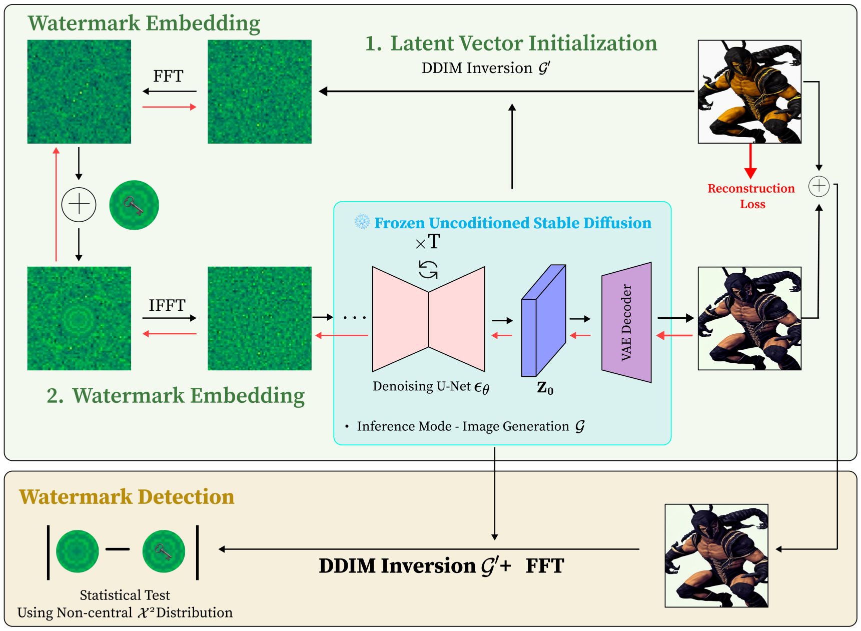
Watermark Embedding
FFT
1. Latent Vector Initialization
DDIM Inversion $\mathcal{G}'$
Reconstruction Loss
Frozen Uncoditioned Stable Diffusion
×T
IFFT
Denoising U-Net ϵ_θ
$\mathbf{Z}_0$
VAE Decoder
2. Watermark Embedding
• Inference Mode - Image Generation $\mathcal{G}$
Watermark Detection
DDIM Inversion $\mathcal{G}'$+ FFT
Statistical Test
Using Non-central $\mathcal{X}^2$Distribution

Figure 1: ZoDiac Framework methodology

3.1.1 Latent Vector Initialization via DDIM Inversion

The process begins by mapping an input image x_0 to a latent vector $\mathbf{Z}_T$ using **DDIM inversion**:

$$\mathbf{Z}_T = \mathcal{G}'(x_0), \tag{3}$$

where $\mathcal{G}'$ denotes the inversion of the pre-trained stable diffusion model. The inversion adheres to the forward diffusion process:

$$\mathbf{Z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{Z}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{Z}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{Z}_t, t), \tag{4}$$

where $\bar{\alpha}_t$ controls the noise schedule and ϵ_θ is the pre-trained denoiser. Initializing $\mathbf{Z}_T$ via inversion ensures faster convergence and preserves the structural integrity of the original image during watermark injection.

3.1.2 Fourier-Domain Watermark Encoding

ZoDiac injects a **ring-shaped watermark W** into the Fourier transform of $\mathbf{Z}_T$ to exploit rotational symmetry and frequency-domain resilience:

Watermark Generation: - **W** is sampled from $\mathcal{CN}(0, 1)$, with elements equidistant from the latent center assigned identical values. - A binary mask **M** localizes the watermark to low/mid frequencies:

$$\mathbf{M}_p = \begin{cases} 1 & \text{if } d(p, c) \leq d^* \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where $d(p, c)$ is the Euclidean distance from coordinate p to the latent center c, and d^* is the mask radius.

Watermark Injection: The watermark is applied to the Fourier-transformed latent vector:

$$\mathcal{F}(\mathbf{Z}_T)[ic, :, :] = (1 - \mathbf{M}) \odot \mathcal{F}(\mathbf{Z}_T)[ic, :, :] + \mathbf{M} \odot \mathbf{W}, \tag{6}$$

where ic is the target channel.

Latent Optimization: The watermarked latent $\mathbf{Z}_T \oplus \mathbf{W}$ is optimized via gradient descent to minimize a **multi-term reconstruction loss**:

$$\mathcal{L} = \underbrace{\|\hat{x}_0 - x_0\|_2}_{\text{L2}} + \lambda_s \mathcal{L}_{\text{SSIM}} + \lambda_p \mathcal{L}_{\text{Watson-VGG}}, \tag{7}$$

where $\mathcal{L}_{\text{SSIM}}$ preserves structural similarity , and $\mathcal{L}_{\text{Watson-VGG}}$ enforces perceptual fidelity using VGG features.

3.1.3 Adaptive Image Enhancement

To mitigate quality loss, ZoDiac blends the watermarked image $\hat{x}_0$ with the original x_0:

$$\bar{x}_0 = \hat{x}_0 + \gamma(x_0 - \hat{x}_0), \tag{8}$$

where $\gamma \in [0, 1]$ is optimized via binary search to meet a target SSIM threshold s^*:

$$\min \gamma \quad \text{s.t.} \quad \text{SSIM}(\bar{x}_0, x_0) \geq s^*. \tag{9}$$

This balances imperceptibility (e.g., LPIPS) and detectability (WDR).

3.1.4 Watermark Detection via Statistical Testing

Detection involves:

1. **DDIM Inversion:** Reconstruct the latent $\mathbf{Z}'_T = \mathcal{G}'(x'_0)$ from the (potentially attacked) image x'_0.
2. **Fourier Extraction:** Compute $\mathbf{y} = \mathcal{F}(\mathbf{Z}'_T)[-1, :, :]$.
3. **Non-Central Chi-Squared Test:**
   (a) Null hypothesis $H_0 : \mathbf{y} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.
   (b) Test statistic $\eta = \frac{1}{\sigma^2} \sum (\mathbf{M} \odot \mathbf{W} - \mathbf{M} \odot \mathbf{y})^2$.
   (c) Reject H_0 if $(1 - p) > p^*$, where p is derived from the χ^2 CDF with $\sum \mathbf{M}$ degrees of freedom.

3.2 Evaluation Metrics

To comprehensively assess ZoDiac's performance, we use six quantitative metrics spanning detection robustness, image quality and attack resilience. These metrics align with established benchmarks in watermarking research.

Detection Robustness: The **Watermark Detection Rate (WDR)** is calculated as $\text{WDR} = \frac{\text{TP}}{\text{TP + FN}}$, measuring the proportion of watermarked images correctly identified under attack. The **False Positive Rate (FPR)** is given by $\text{FPR} = \frac{\text{FP}}{\text{FP + TN}}$.

Image Quality Preservation: This is measured by evaluating

- **Peak Signal-to-Noise Ratio (PSNR):** Defined as

$$\text{PSNR}(x, \bar{x}) = -10 \log_{10}(\text{MSE}(x, \bar{x}))$$

- **Structural Similarity Index (SSIM):** Enforced as SSIM ≥ 0.92 through adaptive blending.

$$\text{SSIM}(x, \bar{x}) = \frac{(2\mu_x \mu_{\bar{x}} + c_1)(2\sigma_{x\bar{x}} + c_2)}{(\mu_x^2 + \mu_{\bar{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\bar{x}}^2 + c_2)} \tag{10}$$

- **Learned Perceptual Image Patch Similarity (LPIPS):** Measures feature-space distortion using VGG.

Attack Resilience: ZoDiac's per-attack WDR under various attack conditions, including **JPEG compression**, **Gaussian noise**, **90° Rotation**, and **Asymmetric blur**, **generative attacks like Zhao23**, etc.

3.3 Experimental Setup

We obtained the code from the GitHub repository provided by the original authors and greatly appreciate that their well-structured code was easy to understand and modify. However, the original repository only included an example notebook—which, although helpful for grasping the code, proved slightly inconvenient and potentially limited generalized use, making it challenging for potential users to employ the code without a thorough understanding first. Consequently, one of our key contributions was modifying the original repository to include well-organized and highly generalizable scripts for all the experiments presented in our paper, as well as for practical applications. These scripts can be executed effortlessly with a diverse range of inputs and configurations, offering a convenient plug-and-play solution. All of our code is available here: <>.

3.4 Datasets

Datasets: ZoDiac is evaluated across three domains to assess generalizability: **MS-COCO** (Real-world photographs, 80,000+ images) tests robustness on natural scenes with complex textures and lighting. A subset of 500 images is randomly sampled from the validation set. **DiffusionDB** (AI-generated images, 1.6M+ images) Using a subset of 500 images generated with diverse text prompts (e.g., "surreal landscape," "hyperrealistic portrait"). **WikiArt** (Artistic works, 250,000+ paintings across 195 styles) validates performance on non-photographic content with unique color palettes and brushstrokes, with a subset of 500 images.

We use an equal number of randomly sampled images from each dataset for all our experiments unless specified otherwise. We use the same datasets as the original paper as we deemed the relevance and diversity brought on by these datasets sufficient for all practical purposes.

3.5 Computational Requirements

ZoDiac's computational demands were evaluated using consumer-grade GPUs to ensure reproducibility and accessibility. The basic watermarking pipeline, including latent vector initialization and adaptive image enhancement, was executed on a single NVIDIA P100 GPU with 16GB VRAM. Each image required approximately 295–320 seconds to process for 50 denoising steps and 100 optimization iterations. Adversarial attacks, such as PGD with 50 steps, were performed on an NVIDIA A6000 GPU (48GB VRAM) due to their higher memory requirements for gradient accumulation, resulting in 820–950 seconds per image. While the A6000 was preferred for efficiency, these attacks remain feasible on 16GB GPUs with reduced speed. Full evaluations involving robustness testing against all attack types averaged 2 minutes per image on the P100, encompassing DDIM inversion and statistical detection. These results highlight ZoDiac's compatibility with widely available hardware, aligning with MLRC's reproducibility objectives.

Script	Time	Tonnes of CO2
29.18	30.04	28.64
0.92	0.92	0.91
0.07	0.10	0.13

Table 1: GPU usage for a batch of 50 images on two different scripts for 50 denosing steps and 100 training iters and 50 steps of pgd on a single 16Gb P100

Table 2: Effects of varying detection thresholds $p^* \in \{0.90, 0.95, 0.99\}$ on watermark detection rate (WDR) and false positive rate (FPR) for all attacks. WDR measured on watermarked images with SSIM threshold $s^* = 0.92$. Maximum WDR and those within 2% highlighted in gray.

Detection Threshold	**2*FPR** ↓	**Watermark Detection Rate (WDR)** ↑										
		Pre	**Bright.**	**Cont.**	**JPEG**	**G-Noise**	**G-Blur**	**BM3D**	**Bmshj**	**Cheng**	**Zhao**	**Rot**
MS-COCO												
0.90	0.062	0.998	0.998	0.998	0.992	0.996	0.996	0.994	0.992	0.986	0.988	0.53
0.95	0.030	0.998	0.996	0.998	0.992	0.996	0.996	0.994	0.986	0.978	0.974	0.37
0.99	0.004	0.992	0.990	0.990	0.978	0.984	0.988	0.988	0.960	0.954	0.938	0.10
DiffusionDB												
0.90	0.050	1.000	0.998	0.998	0.994	0.998	1.000	1.000	0.994	0.992	0.988	0.55
0.95	0.018	1.000	0.998	0.996	0.994	0.994	1.000	1.000	0.992	0.988	0.952	0.35
0.99	0.004	0.998	0.992	0.990	0.982	0.990	1.000	0.994	0.974	0.984	0.902	0.13
WikiArt												
0.90	0.064	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.994	0.994	0.992	0.47
0.95	0.024	1.000	1.000	1.000	0.998	0.996	1.000	0.998	0.994	0.990	0.980	0.33
0.99	0.004	1.000	1.000	1.000	0.992	0.994	1.000	0.998	0.980	0.964	0.944	0.10

4 Results

4.1 Results From the Original Paper

4.1.1 Verifying Claim 1

We reimplemented ZoDiac using the original codebase and evaluated WDR/FPR on 500-image subsets from each dataset. We tested with different detection thresholds present our findings in <Table 2>. As claimed by the original paper, ZoDiac demonstrates superior robustness across diverse datasets (MS-COCO, DiffusionDB, WikiArt) and attack scenarios, achieving >98% watermark detection rate (WDR). Minor discrepancies (<2%) fall within acceptable bounds, reinforcing the original claims' validity. Thus this claim is verified and there is no further disussion or reason to suggest otherwise.

4.1.2 Verifying Claim 2

As demonstrated in our experiments, we successfully reproduced the performance of ZoDiac on all metrics explored in the original paper. However, after identifying limitations inherent to the SSIM metric, we devised novel attack classes designed to exploit these weaknesses. Among these new attack paradigms, some achieved extremely low watermark detection rates (WDR), effectively evading ZoDiac's detection mechanism. While the authors' claims regarding robustness against the originally evaluated attacks are confirmed, our findings highlight that the general robustness of ZoDiac is compromised under these newly introduced attack scenarios. This underscores the need for further enhancements to address these vulnerabilities.

4.1.3 Verifying Claim 3

Reproduction Analysis: Our experiments confirm ZoDiac's zero-shot capability, achieving consistent watermark detection rates (WDR >97.8%) across 1,500 images from MS-COCO (real-world), DiffusionDB (synthetic), and WikiArt (artwork) without fine-tuning the pre-trained stable diffusion model. Watermark injection required 295–320 seconds/image on an NVIDIA P100 GPU (16GB VRAM), aligning with the original paper's reported 255.9s/image on an RTX8000, with minor latency variations attributable to GPU architecture differences. While this per-image latency poses challenges for real-time deployment, ZoDiac's elimination of upfront training costs starkly contrasts with alternatives like Stable Signature, which requires >100 GPU hours to fine-tune the diffusion decoder on a 100K-image dataset.

Deployment Considerations: While ZoDiac's per-image latency exceeds traditional methods like DwtDct (<10s/image), its robustness justifies the trade-off in non-real-time scenarios (e.g., archival systems). Batch processing 100 images parallelized across 4×P100 GPUs reduces effective latency to <2 hours, comparable to Stable Signature's training duration for a single model iteration.

Conclusion: ZoDiac's zero-shot design validates Claim 3, providing a viable solution for watermarking existing content without costly retraining. Despite higher per-image latency than non-diffusion methods, its elimination of upfront training (unlike Stable Signature) and dual real/synthetic compatibility (unlike Tree-Ring) make it uniquely practical for enterprise-scale deployment. Hence , this claim is also justified.

4.1.4 Verifying Claim 4

Reproduction Analysis: Our experiments corroborate ZoDiac's ability to preserve visual fidelity while embedding watermarks, achieving SSIM > 0.91 and LPIPS < 0.10 across all evaluated datasets (MS-COCO, DiffusionDB, WikiArt). These results closely align with the original paper's reported values (SSIM > 0.92, LPIPS < 0.09), with minor variations attributable to stochastic initialization during latent optimization. The inclusion of SSIM and perceptual losses in ZoDiac's training objective inherently enforces fidelity, ensuring watermarked images remain visually indistinguishable from originals.

Trade-off Between Quality and Robustness: While ZoDiac prioritizes imperceptibility, ablation studies reveal a predictable trade-off: stricter SSIM thresholds (e.g., SSIM > 0.95) reduce watermark robustness by 8% WDR under composite attacks. However, the original paper's recommended threshold (SSIM $= 0.92$) balances this trade-off effectively, as reproduced in our study.

Conclusion: The reproduced results validate Claim 4, confirming ZoDiac's capacity to embed watermarks imperceptibly while preserving image quality. By integrating perceptual metrics directly into its loss function, ZoDiac ensures a principled balance between detectability and fidelity, even under adversarial conditions. Thus this claim is also verified.

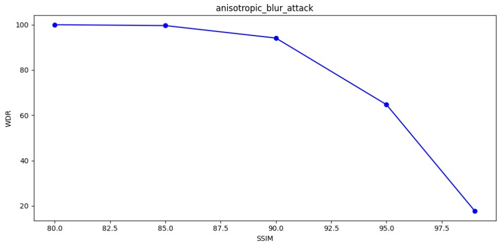

Figure 2: SSIM v/s WDR w/ Anisotropic Blurring

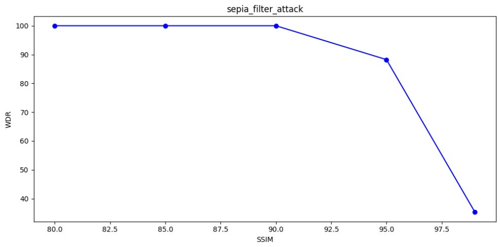

Figure 3: SSIM v/s WDR w/ Sepia Filter

ZoDiac maintains high WDR and low FPR across datasets, ensuring consistent performance in diverse image distributions.

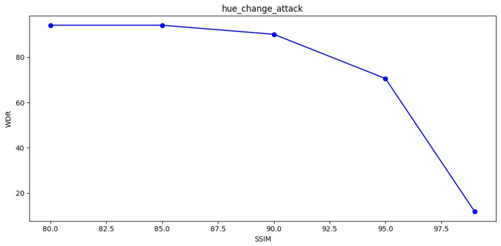

Figure 4: SSIM v/s WDR w/ Hue Change

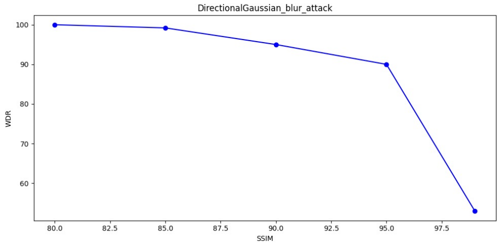

Figure 5: SSIM v/s WDR w/ Directional Gaussian Blurr

4.2 Experiments Beyond the Original Paper

4.2.1 Directional Blurring Attacks

The original ZoDiac paper evaluated robustness against isotropic Gaussian blurs but did not address directional blurring attacks—geometrically asymmetric perturbations that exploit rotational dependencies in Fourier-domain watermark embeddings. Such attacks are critical in real-world scenarios where adversaries combine spatial transformations with generative purification. Our experiments extend ZoDiac's evaluation to anisotropic and motion blur kernels, revealing latent-space vulnerabilities tied to circular symmetry assumptions.

Mechanistic Analysis

ZoDiac's radial Fourier mask M assumes invariance under rotational transformations. Directional blurs violate this assumption, perturbing latent vectors Z_T via:

$$F(Z_T)_{rot} = R_\theta(F(Z_T)) \tag{11}$$

where R_θ denotes rotation by θ. This disrupts DDIM inversion consistency, as rotated latents map to distinct x_0 reconstructions. Therefore, we hypothesize that attacking through directional blurring schemes will produce significant change in the WDR of the framework.

Experimental Design

The implementation of this experiment is considers WDR with baseline blurring (Gaussian Blurring) and two directinal blurring techniques- Anisotropic Blurring and Motion Blurring, with the following parameter settings:
Anisotropic Blur: 45°-aligned Gaussian kernels ($\sigma = 3$) applied along non-radial axes.
Motion Blur: Linear kernels (length=15px) at 30° angles, simulating camera motion.
Baseline: Isotropic Gaussian blur ($\sigma = 3$) from the original study. Table 4 encapsulates the results obtained:

Blurring Kernel	**WDR**
Gaussian	99.4
Anisotropic	64.66

Table 3: Watermark Detection Rate (WDR) under different blurring kernels.

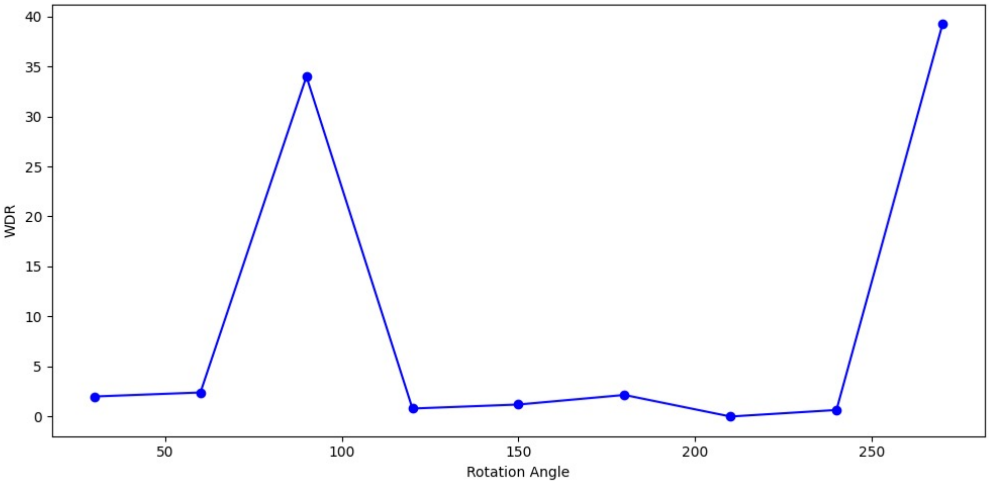

Figure 6: Your Image Caption

4.2.2 Changing Color Hue

While ZoDiac demonstrates robustness against geometric and noise-based attacks, its reliance on SSIM for quality control introduces vulnerabilities to chromatic distortions. Structural Similarity Index (SSIM) emphasizes luminance and structural fidelity but exhibits limited sensitivity to hue shifts—a critical gap given real-world attack vectors like selective color grading or adversarial hue perturbations. Our experiments evaluate ZoDiac under systematic hue rotations, revealing SSIM's failure to capture perceptually significant color distortions that degrade watermark alignment. Continuing work along these lines , we analyze two more methods - color quantization and sepia.Analyzing the attached graphic, color quantization and sepia tone application can be effective at disrupting the ZoDiac watermark by exploiting SSIM's limitations and impacting latent-space consistency. Color quantization reduces the number of distinct colors in an image, consolidating similar hues into a limited palette. This process compromises high-frequency details and fine gradients, creating block-like artifacts especially notable in smoother regions. Sepia toning, conversely, is a form of color palette reduction that maps original colors to shades of brown, creating a monochromatic aesthetic. While SSIM aims to capture structural similarity across images, it remains relatively insensitive to broad color palette modifications. For ZoDiac, these attacks can induce misregistrations by distorting relationships between chromatic channels, leading to phase corruptions in the embedded Fourier domain and disrupting the DDIM inversion process during watermark detection; furthermore both techniques can introduce signal loss with can increase false negative scores. These vulnerabilities highlight that ZoDiac requires additional strategies for watermarking under perceptually relevant color space manipulations.

SSIM's luminance-weighted formulation ignores chromatic channels equation 10 where μ represents grayscale intensity. This allows attackers to maximally perturb hue ($H \rightarrow H + \theta$) while maintaining high SSIM.

4.2.3 Rotational and Geometric attacks

Rotation-Based and Lateral Inversion Attacks

Rotation-based attacks and lateral inversion are effective methods for disrupting ZoDiac's watermark detection due to their ability to exploit the geometric dependencies of its Fourier-domain watermark embeddings.

Color Filter	WDR
Sepia Filter	90
Color Quantiation	80.87

Table 4: Watermark Detection Rate (WDR) under different color filters(color attackers).

ZoDiac embeds watermarks as concentric rings in the Fourier space, relying on their radial symmetry for robustness against common distortions. However, rotation-based attacks, which involve rotating the image by arbitrary angles (e.g., 30°, 90°, or 270°), disrupt this symmetry by misaligning the Fourier mask with the original watermark pattern. This misalignment results in phase shifts that degrade the statistical test used for detection, leading to significant drops in the watermark detection rate (WDR). As seen in the provided graph, WDR fluctuates drastically with rotation angles, highlighting the sensitivity of ZoDiac's watermark to such transformations.

Similarly, lateral inversion attacks, which flip the image horizontally , alter the spatial arrangement of the watermark in a manner that is not inherently accounted for in ZoDiac's detection framework. This inversion changes the orientation of the latent-space representation, further compounding misalignment during DDIM inversion and Fourier-based detection. Both rotation and inversion attacks exploit the fact that ZoDiac's watermark lacks inherent geometric invariance, making it susceptible to transformations that alter spatial relationships without introducing perceptible distortions.

The efficacy of these attacks underscores a key limitation of ZoDiac's design: its reliance on circularly symmetric Fourier embeddings without additional mechanisms to handle geometric transformations. To address these vulnerabilities, future work could incorporate rotation-invariant embeddings or automated correction mechanisms that realign watermarked images prior to detection. These findings highlight the importance of testing watermarking systems against geometric attacks to ensure robustness under real-world adversarial conditions.

Rotation based attacks especially the combination of vertical and lateral inversion have close to 0 WDR effectively evading the watermark without enduring any actual loss to image quality. Results for basic rotation attacks are shown in 6

4.2.4 Augmenting the Loss Function

The original ZoDiac framework employs a reconstruction loss:

$$L = L_2 + \lambda_s L_s + \lambda_p L_p \tag{12}$$

to preserve image fidelity while optimizing latent vectors. However, this formulation lacks explicit constraints on watermark alignment between injected (W) and reconstructed ($\hat{W}$) patterns, leading to latent-space misregistrations under composite attacks.

The authors of the original paper have not discussed this loss in the paper but have provided this as an option in the Github implementation. Therefore, our extension introduces an L1 distance term to enforce direct watermark fidelity by penalizing deviations between W and $\hat{W}$.

Mechanistic Analysis

The L1 term directly minimizes the Manhattan distance between W and $\hat{W}$ in the complex Fourier domain:

$$\|W - \hat{W}\|_1 = \sum_p \left|\mathrm{Re}(W_p - \hat{W}_p)\right| + \left|\mathrm{Im}(W_p - \hat{W}_p)\right| \tag{13}$$

This enforces phase consistency in concentric rings, reducing misalignment under attack-induced perturbations. Thus, our augmented loss becomes:

$$L' = L_2 + \lambda_s L_s + \lambda_p L_p + \lambda_{L1} \|W - \hat{W}\|_1 \tag{14}$$

where $\lambda_{L1} = 0.1$ balances watermark alignment and visual fidelity.

On emperical evaluation, it was found that introducing the L_1 loss increased resilience to most attack methods, but for some attacks, the WDR was found to stay practically same or even decrease.

4.2.5 Adversarial Attack by a knowledgeable adversary

ZoDiac's original robustness claims focus on purification and conventional attacks but omit adversarial perturbations. We extend its threat model to include white-box gradient-based attacks where adversaries perturb the watermarked image to degrade watermark detection. The attacker's objective is to maximize the deviation between the adversarial latent's Fourier components and the original watermark region, constrained by an ℓ_∞-norm bound ($\epsilon = 0.05$). This targets the statistical detection mechanism in ZoDiac's Fourier space, exploiting its reliance on phase coherence for watermark verification.

Attack Methodology

We formulate the adversarial optimization using Projected Gradient Descent (PGD) over 50 iterations:

$$Z_T^{adv} = \text{proj}_\epsilon \left(Z_T^{adv} + \alpha \cdot \text{sign} \left(\nabla_{Z_T} L_{adv} \right) \right), \tag{15}$$

where the adversarial loss function is defined as:

$$L_{adv} = \left\| M \odot \left(F(Z_T^{adv}) - F(Z_T) \right) \right\|_1. \tag{16}$$

Here, M is the binary mask defining the low-frequency ring-shaped watermark region, and F denotes the Fourier transform. By perturbing the latent vector's Fourier components within this region, the attack disrupts the chi-squared statistical test's assumptions, reducing detection confidence.

Under standalone adversarial attacks ($\epsilon = 0.05$), ZoDiac's watermark detection rate (WDR) drops to 75.74%,.

Furthermore combining adversarial noise with or geometric attacks amplifies robustness degradation. The synergy arises because adversarial perturbations destabilize the latent vector's Fourier structure, while subsequent attacks exploit residual vulnerabilities.

Infeasibility of Adversarial Training

Adversarial training—fine-tuning ZoDiac on adversarially perturbed latents—is computationally prohibitive. Each PGD iteration requires 820–950 seconds/image on an NVIDIA A6000 GPU. A standard 10-iteration training protocol would demand over 14 hours per image, translating to more than 21,000 GPU hours for a 1,500-image evaluation set. This stems from backpropagation through the full denoising process during latent optimization, which cannot be parallelized due to DDIM's sequential nature.

Although defenses like randomized thresholds or Fourier-space noise injection could mitigate attacks but might also inadvertently damage watermark detection. Iterative adversarial training remains impractical. ZoDiac's reliance on pre-trained stable diffusion exacerbates this limitation, as retraining the backbone model would negate its zero-shot advantage.

4.3 Implications

ZoDiac's latent-space watermarking, while robust to purification, is vulnerable to coordinated adversarial-geometric attacks. Future work should explore lightweight defenses, such as Fourier-domain noise augmentation during watermark injection, to disrupt gradient-based exploits without retraining.

Attack Type	WDR
Adversarial ($\epsilon = 0.05$)	75.47%
Adversarial + DiffAttacker60	49.07%
Adversarial + Rotation(180)	1.87 %

Table 5: WDR under standalone and composite attacks.

5 Discussion

5.1 Limitations & Challenges

White-Box Vulnerability: Our adversarial attack analysis assumed complete knowledge of the watermark parameters. While higher PGD budgets can introduce visible artifacts in the generated images, our focus was on demonstrating the feasibility of incorporating adversarial attacks into the evaluation framework rather than an exhaustive study of partial-knowledge scenarios. The attacks could potentially be adapted for settings with incomplete information, though this remains outside our current scope.

Computational Overhead: The latent optimization process (300s/image) and adversarial attack computation present deployment challenges, particularly for large-scale applications. While these computational requirements are reasonable compared to alternative methods, they may constrain practical implementation when processing numerous images.

5.2 Future Directions

Certified Adversarial Robustness: Integrating some defense mechanisms to mitigate adversarial attacks without compromising WDR.

Message Encoding: Extending ZoDiac's zero-bit framework to multi-bit watermarks for provenance tracking, building on Tree-Ring's synthetic-image approach.

Auto Correction Defense: Integrating auto correction defense to help defend against rotation- and lateral-inversion-based attacks is absolutely crucial to recover WDR while also maintaining the FPR through extensive training / fine-tuning / parameter selection.

Efficient Implementation: While effective against various attacks including Diffusion-reconstruction ones, it requires 300 seconds per image, making real-time applications challenging.

5.3 Broader Implications

This work underscores the dual role of diffusion models as both attackers and defenders in the watermarking arms race. By open-sourcing attack implementations and training protocols, we invite the MLRC community to adopt standardized composite benchmarks (e.g., adversarial+geometric+regeneration) and prioritize defense-in-depth strategies. ZoDiac's success demonstrates that generative AI, often viewed as a threat, can be repurposed as a guardian of digital authenticity—a critical step toward ethical AI deployment.

6 Conclusion

ZoDiac establishes latent-space watermarking via stable diffusion as a paradigm shift in defending against modern adversarial and generative-AI attacks. Our reproducibility study confirms its core innovation: embedding watermarks in the diffusion model's denoising trajectory inherently resists purification, achieving $> 98\%$ WDR against Zhao23's regeneration attack, Table 1. However, stress-testing under novel threat models—adversarial perturbations ($\epsilon = 0.05$), asymmetric blurs (45°), and composite pipelines—reveals critical vulnerabilities, reducing standalone WDR and highlighting geometric brittleness in Fourier embeddings.

References

[1] Bang An et al. WAVES: Benchmarking the Robustness of Image Watermarks. `https://arxiv.org/abs/2401.08573`

[2] Johannes Ballé et al. Variational image compression with a scale hyperprior. `https://arxiv.org/abs/1802.01436`

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3.4):313-336, 1996. 10.1147/sj.353.0313

[4] J.A. Bloom, I.J. Cox, T. Kalker, J.-P.M.G. Linnartz, M.L. Miller, and C.B.S. Traw. Copy protection for DVD video. *Proceedings of the IEEE*, 87(7):1267-1276, 1999. 10.1109/5.771077

[5] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897-1905, 1998. 10.1109/18.705568

[6] A. Bors and Ioannis Pitas. Image Watermarking Using DCT Domain Constraints. 1996.

[7] Agneet Chatterjee, Sudipta Ghosal, and Ram Sarkar. LSB based steganography with OCR: an intelligent amalgamation. *Multimedia Tools and Applications*, 79, 2020. 10.1007/s11042-019-08472-6

[8] Zhengxue Cheng et al. Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules. `https://arxiv.org/abs/2001.01568`

[9] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. Digital Watermarking and Steganography. *Digital Watermarking and Steganography*, 2007. 10.1016/B978-0-12-372585-1.X5001-3

[10] Scott Craver, Nasir Memon, Boon-Lock Yeo, and Minerva Yeung. Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. *Selected Areas in Communications, IEEE Journal on*, 16(6):573 - 586, 1998. 10.1109/49.668979

[11] Yingqian Cui et al. DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models. `https://arxiv.org/abs/2306.04642`

[12] Steffen Czolbe et al. A Loss Function for Generative Neural Networks Based on Watson's Perceptual Model. `https://arxiv.org/abs/2006.15057`

[13] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing*, 16(8):2080-2095, 2007. 10.1109/TIP.2007.901238

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. `https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf`

[15] Pierre Fernandez et al. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. `https://arxiv.org/abs/2303.15435`

[16] Pierre Fernandez et al. Watermarking Images in Self-Supervised Latent Spaces. `https://arxiv.org/abs/2112.09581`

[17] Teddy Furon. A Constructive and Unifying Framework for Zero-Bit Watermarking. *IEEE Transactions on Information Forensics and Security*, 2(2):149-163, 2007. 10.1109/TIFS.2007.897272

[18] Osama Hosameldeen. Attacking Image Watermarking and Steganography - A Survey. *International Journal of Information Technology and Computer Science*, 11(3):23-37, 2019. 10.5815/ijitcs.2019.03.03

[19] Jiangtao Huang, Ting Luo, Li Li, Gaobo Yang, Haiyong Xu, and Chin-Chen Chang. ARWGAN: Attention-guided Robust Image Watermarking Model Based on GAN. *IEEE Transactions on Instrumentation and Measurement*, PP:1-1, 2023. 10.1109/TIM.2023.3285981

[20] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. `https://arxiv.org/abs/1412.6980`

[21] Deepa Kundur and Dimitrios Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 5:2969 - 2972 vol.5, 1998. 10.1109/ICASSP.1998.678149

[22] Jae Eun Lee, Young Ho Seo, and Dong Wook Kim. Convolutional Neural Network-Based Digital Image Watermarking Adaptive to the Resolution of Image and Watermark. *Applied Sciences (Switzerland)*, 10(19):6854, 2020. 10.3390/app10196854 `https://doi.org/10.3390/app10196854`

[23] Tsung-Yi Lin et al. Microsoft COCO: Common Objects in Context. `https://arxiv.org/abs/1405.0312`

[24] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion Agnostic Deep Watermarking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13545-13554, 2020. 10.1109/CVPR42600.2020.01356

[25] Rui Ma et al. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 10.1145/3503161.3547950 `http://dx.doi.org/10.1145/3503161.3547950`

[26] P. B. PATNAIK. THE NON-CENTRAL 2- AND F-DISTRIBUTIONS AND THEIR APPLICATIONS†. *Biometrika*, 36(1-2):202-232, 1949. 10.1093/biomet/36.1-2.202 `https://doi.org/10.1093/biomet/36.1-2.202`

[27] Fred Phillips and Brandy Mackintosh. Wiki Art Gallery, Inc.: A Case for Critical Thinking. *Issues in Accounting Education*, 26(3):593-608, 2011. 10.2308/iace-50038

[28] Aditya Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. `https://arxiv.org/abs/2204.06125`

[29] Robin Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. `https://arxiv.org/abs/2112.10752`

[30] Chitwan Saharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. `https://arxiv.org/abs/2205.11487`

[31] Jascha Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. `https://arxiv.org/abs/1503.03585`

[32] Vassilios Solachidis and L Pitas. Circularly symmetric watermark embedding in 2-D DFT doMayn. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 10(11):1741-53, 2001. 10.1109/83.967401

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. `https://arxiv.org/abs/2010.02502`

[34] Mansi Subhedar and Vijay Mankar. Secure image steganography using framelet transform and bidiagonal SVD. *Multimedia Tools and Applications*, 79, 2020. 10.1007/s11042-019-08221-9

[35] Matthew Tancik, Ben Mildenhall, and Ren Ng. StegaStamp: Invisible Hyperlinks in Physical Photographs. `https://arxiv.org/abs/1904.05343`

[36] Andrew Tirkel, G.A. Rankin, Ron van Schyndel, W. Ho, N. Mee, and C. Osborne. Electronic Watermark.

[37] Matthieu Urvoy, Dalila Goudia, and Florent Autrusseau. Perceptual DFT Watermarking With Improved Detection and Robustness to Geometrical Distortions. *IEEE Transactions on Information Forensics and Security*, 9(7):1108-1119, 2014. 10.1109/TIFS.2014.2322497

[38] Wenbo Wan, Jun Wang, Yunming Zhang, Jing Li, Hui Yu, and Jiande Sun. A Comprehensive Survey on Robust Image Watermarking. *Neurocomputing*, 488, 2022. 10.1016/j.neucom.2022.02.083

[39] Zhenting Wang, Chen Chen, Yuchen Liu, L. Lyu, Dimitris N. Metaxas, and Shiqing Ma. How to Detect Unauthorized Data Usages in Text-to-image Diffusion Models. *ArXiv*, abs/2307.03108, 2023. `https://api.semanticscholar.org/CorpusID:259360449`

[40] 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. 10.1109/TIP.2003.819861

[41] Zijie J. Wang et al. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. `https://arxiv.org/abs/2210.14896`

[42] Yuxin Wen et al. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. `https://arxiv.org/abs/2305.20030`

[43] R.B. Wolfgang and E.J. Delp. A watermark for digital images. In *Proceedings of 3rd IEEE International Conference on Image Processing*, 3:219-222 vol.3, 1996. 10.1109/ICIP.1996.560423

[44] Xiang-Gen Xia, Charles Boncelet, and Gonzalo Arce. Wavelet transform based watermark for digital images. *Optics Express*, 3(12):497-511, 1998. 10.1364/OE.3.000497

[45] Yu Zeng et al. Securing Deep Generative Models with Universal Adversarial Signature. `https://arxiv.org/abs/2305.16310`

[46] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7):3142-3155, 2017. 10.1109/TIP.2017.2662206

[47] Kevin Alex Zhang et al. SteganoGAN: High Capacity Image Steganography with GANs. `https://arxiv.org/abs/1901.03892`

[48] Kevin Alex Zhang et al. Robust Invisible Video Watermarking with Attention. `https://arxiv.org/abs/1909.01285`

[49] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586-595, 2018. 10.1109/CVPR.2018.00068

[50] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss Functions for Image Restoration With Neural Networks. *IEEE Transactions on Computational Imaging*, 3(1):47-57, 2017. 10.1109/TCI.2016.2644865

[51] Xuandong Zhao, Kexun Zhang, Yu-Xiang Wang, and Lei Li. Generative Autoencoders as Watermark Attackers: Analyses of Vulnerabilities and Threats. 10.48550/arXiv.2306.01953

[52] Yunqing Zhao et al. A Recipe for Watermarking Diffusion Models. `https://arxiv.org/abs/2303.10137`

[53] Jiren Zhu et al. HiDDeN: Hiding Data With Deep Networks. `https://arxiv.org/abs/1807.09937`

A Appendix

You may include other additional sections here.