TRAJRING: TAMING TRAJECTORY DRIFT FOR HIGH-FIDELITY DIFFUSION WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion watermarking methods such as Tree-Ring enable provenance verification of diffusion-generated images by embedding watermark patterns in the initial latent noise. Despite their robustness to common image transformations, they inevitably compromise generation fidelity, causing watermarked outputs to deviate from their clean counterparts. Existing optimization-based approaches improve generation fidelity but incur substantial per-image computational overhead. To explore an efficient approach to preserving generation fidelity, we analyze paired clean and watermarked denoising trajectories. We identify watermark-induced trajectory drift: perturbations initially confined to the watermark region induce growing deviations in complementary frequencies during denoising, progressively driving the watermarked trajectory away from its clean counterpart. Motivated by this finding, we propose TrajRing, a training-free framework that suppresses this drift while preserving watermark detectability. At each denoising step, TrajRing performs trajectory-aware spectral projection, retaining watermark-bearing components while projecting complementary components onto a timestep-aligned clean reference trajectory. To complement this spectral constraint with spatial guidance, clean-reference query-key injection reuses queries and keys from the clean reference trajectory while retaining value representations from the watermarked branch. Across two benchmarks, TrajRing improves PSNR by up to 14.91 dB over Tree-Ring while maintaining average detection AUCs of at least 0.98 across clean and transformed images. Evaluations across multiple diffusion backbones further show consistent fidelity improvements. Without per-image optimization, TrajRing runs in 4.77 seconds per image, achieving 24$\times$ and 98$\times$ speedups over PT-Mark and ZoDiac, respectively.

1 INTRODUCTION

Diffusion models have enabled high-quality image synthesis at scale, making reliable provenance verification increasingly important for generated content Ho et al. (2020); Rombach et al. (2022); Dhariwal & Nichol (2021). Digital watermarking offers a practical approach by embedding identifiable signals that can later be recovered to verify provenance Zhang et al. (2019); Tancik et al. (2020); Zhu et al. (2018). In diffusion models, these signals can be embedded in the initial latent noise, integrating watermarking into the generation process without modifying the model architecture or post-processing the generated images Ci et al. (2024b). A representative example is Tree-Ring Watermarking Wen et al. (2023), which embeds a structured ring pattern into a compact Fourier region of the initial latent noise and detects it through diffusion inversion, achieving strong robustness to common image transformations.

Despite this robustness, modifying the initial latent noise can substantially alter the generated content, causing watermarked outputs to deviate from their clean counterparts under otherwise identical generation settings. Existing approaches based on per-image optimization can reduce these deviations, but incur substantial computational overhead. This raises a central question: *can we improve fidelity to clean generations while maintaining robust watermark detection, without costly per-image optimization?*

To investigate how such fidelity improvements can be achieved, we trace paired clean and watermarked denoising trajectories under matched generation settings. We find that perturbations initially

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

confined to the designated watermark region induce growing deviations in complementary frequencies during denoising. At the latent level, the watermarked trajectory also progressively departs from its clean counterpart. We term this phenomenon watermark-induced trajectory drift. This finding motivates retaining watermark-bearing components while controlling the deviations that develop outside the watermark region.

Based on this finding, we propose TrajRing, a training-free framework that uses a clean sampling pass to guide a second, watermarked pass. At each denoising step, trajectory-aware spectral projection retains the current watermarked components within the protected region and aligns complementary components with the clean reference. To complement this spectral constraint with spatial guidance, clean-reference query-key injection reuses queries and keys from the clean reference trajectory while retaining value representations from the watermarked branch. TrajRing improves fidelity to clean generations while maintaining robust watermark detection, without per-image optimization. It remains compatible with the original Tree-Ring inversion-based detector and introduces no additional watermark encoder or decoder.

Experiments on DiffusionDB and MS-COCO demonstrate PSNR improvements of 12.69 and 14.91 dB over Tree-Ring, respectively, while maintaining average detection AUCs of 0.98 and 0.99. With a runtime of 4.77 seconds per image, including reference sampling, TrajRing achieves approximately 24$\times$ and 98$\times$ speedups over PT-Mark and ZoDiac, respectively. Evaluations across multiple diffusion backbones further show consistent fidelity improvements. Trajectory-level analyses show that these fidelity gains are accompanied by reduced deviations in complementary frequencies and closer alignment with the clean reference trajectory.

Our main contributions are summarized as follows:

- We identify and characterize **watermark-induced trajectory drift**, showing that initially localized watermark perturbations induce growing deviations in complementary frequencies during denoising, progressively diverting the watermarked trajectory from its clean counterpart.
- We propose TrajRing, a training-free framework that combines trajectory-aware spectral projection with clean-reference query-key injection to suppress drift and preserve watermark detectability without per-image optimization.
- We demonstrate consistent improvements in fidelity to clean generations across datasets and diffusion backbones while maintaining robust watermark detection. Component ablations and trajectory analyses support the complementary roles of spectral projection and attention guidance.

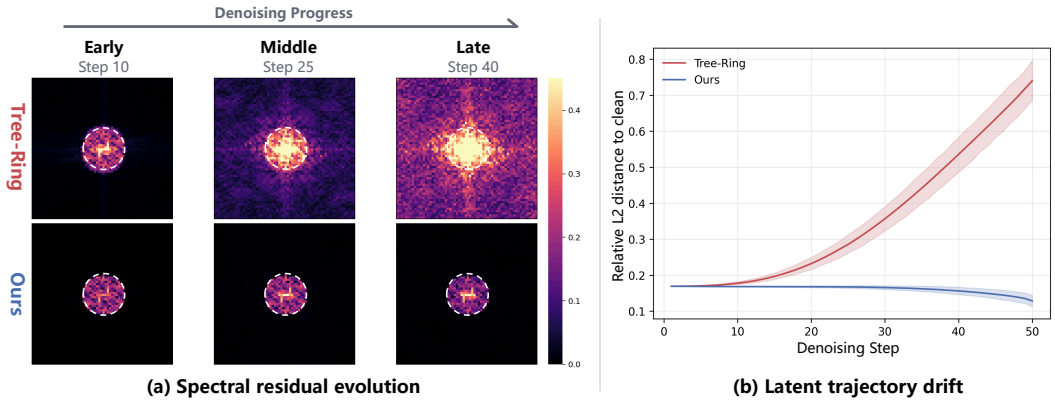
2 Related Work

2.1 Watermarking for Diffusion Models

Several diffusion watermarking methods require model fine-tuning or learned watermark components Fernandez et al. (2023); Feng et al. (2024); Wang et al. (2025); Meng et al. (2025); Wang et al. (2026a); Ci et al. (2024a); Min et al. (2024). Tree-Ring Wen et al. (2023) instead embeds spectral patterns in the initial noise and detects them through diffusion inversion without additional training. Gaussian Shading Yang et al. (2024) preserves the Gaussian latent distribution, although distribution preservation does not by itself ensure fidelity to a corresponding unwatermarked image. To improve content preservation, ROBIN Huang et al. (2024) optimizes watermark patterns and prompt guidance for intermediate-state embedding, while ZoDiac Zhang et al. (2024) optimizes image-specific latent representations. PT-Mark Wang et al. (2026b) aligns watermarked and original trajectories through per-step optimization of null-text embeddings. TrajRing shares the goal of content preservation but directly constrains complementary spectral deviations along a clean reference trajectory, avoiding per-image optimization.

2.2 Trajectory and Attention Control in Diffusion Models

Reference representations enable structural control during diffusion sampling Parmar et al. (2023); Couairon et al. (2022); Avrahami et al. (2023). Prompt-to-Prompt Hertz et al. (2022) controls cross-



(a) Spectral residual evolution **(b) Latent trajectory drift**

Figure 1: **Watermark-induced trajectory drift.** (a) Fourier residuals relative to paired clean latents; dashed contours mark the protected watermark region. (b) Relative latent ℓ_2 distance to the clean trajectory. TrajRing reduces deviations outside the watermark region and maintains closer alignment with the clean trajectory.

attention maps for text-guided editing, while Plug-and-Play Diffusion Features Tumanyan et al. (2023) injects reference spatial features and self-attention maps to preserve source structure. MasaCtrl Cao et al. (2023) and DragonDiffusion Mou et al. (2024) further exploit attention interactions or feature correspondence for content-consistent editing. TrajRing adapts reference-guided attention control to watermarking, where clean and watermarked branches share the same text prompt. It combines clean-reference query-key injection with trajectory-aware spectral projection to suppress watermark-induced drift while retaining watermark-bearing components.

3 Watermark-Induced Trajectory Drift

Latent diffusion models generate images by progressively denoising an initial latent under a text condition Rombach et al. (2022). Tree-Ring Wen et al. (2023) embeds a watermark by modifying a compact Fourier region of the initial latent noise. However, the resulting watermarked image can differ substantially from its clean counterpart. To understand how this deviation develops during generation, we trace the effect of watermark insertion throughout the denoising process.

Trajectory analysis. We compare paired clean and Tree-Ring generations initialized from the same noise sample before watermark insertion, using identical prompts and sampling settings. Figure 1(a) visualizes the average magnitude of their Fourier residuals at successive denoising steps; dashed contours mark the protected watermark region. For Tree-Ring, the residual is concentrated near the protected watermark region at early denoising steps, but deviations emerge and grow in complementary frequencies as denoising proceeds. Figure 1(b) further shows that Tree-Ring's relative latent ℓ_2 distance to the clean reference increases during denoising, indicating progressive departure from the clean trajectory. We term this phenomenon **watermark-induced trajectory drift**.

Design implication. These observations motivate controlling the deviations that develop outside the protected watermark region during denoising. Our design principle is to retain the current watermarked components within this region while aligning complementary components with a timestep-aligned clean reference trajectory.

4 TrajRing

4.1 Overview

As illustrated in Fig. 2, TrajRing uses a clean reference pass to guide a second, watermarked sampling pass. It combines trajectory-aware spectral projection with clean-reference query-key injection to suppress watermark-induced trajectory drift while preserving fidelity to the clean generation.

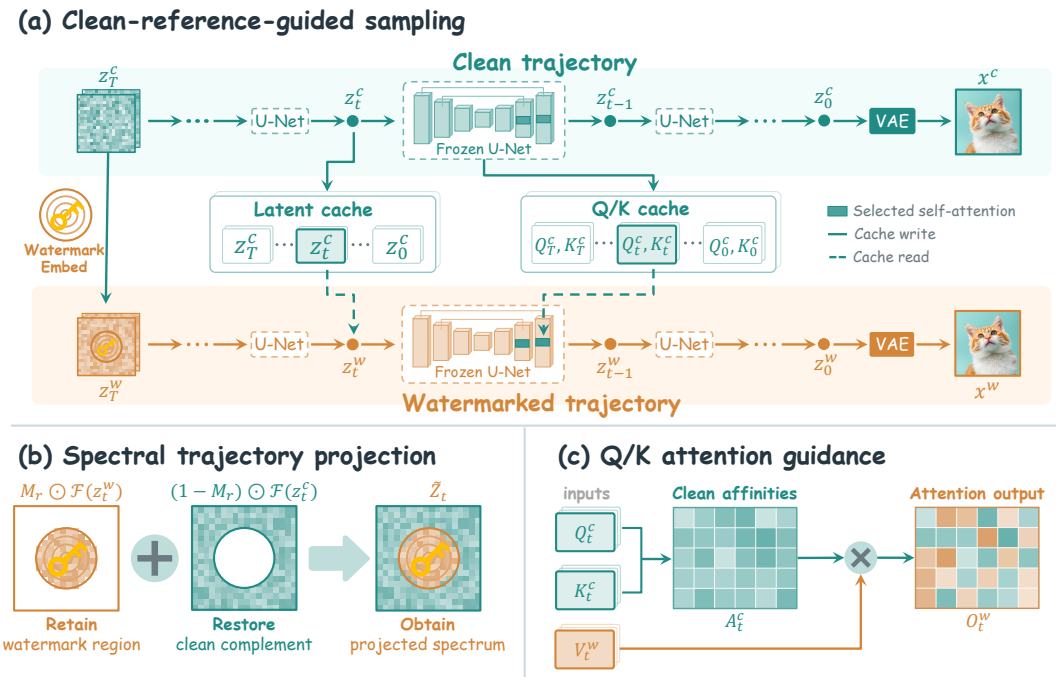


Figure 2: **Overview of TrajRing.** (a) A clean sampling pass caches latent states and self-attention queries and keys to guide a second, watermarked pass using the same frozen model. (b) Trajectory-aware spectral projection retains the current watermarked coefficients in the protected region and replaces complementary coefficients with the clean reference. (c) Clean-reference queries and keys determine attention affinities, which aggregate value representations from the watermarked branch.

Latent diffusion models (LDMs) Rombach et al. (2022) generate an image by progressively denoising an initial latent z_T under a condition c. We denote one denoising step as $z_{t-1} = D_\theta(z_t, t, c)$, where D_θ is the pretrained denoising model. After sampling, the final latent z_0 is decoded by a pretrained VAE decoder $\mathcal{D}$ to obtain the image $x = \mathcal{D}(z_0)$.

Tree-Ring Wen et al. (2023) embeds a structured watermark into the initial latent in the Fourier domain. Let $\mathcal{F}(\cdot)$ denote the Fourier transform, M_r a predefined binary mask specifying the watermark region, and W a private watermark pattern. Given an initial latent z_T, the watermarked initialization is constructed as

$$Z_T^w = (1 - M_r) \odot \mathcal{F}(z_T) + M_r \odot W, \qquad z_T^w = \mathcal{F}^{-1}(Z_T^w), \tag{1}$$

where $\odot$ denotes element-wise multiplication.

Given an initial latent z_T and text condition c, the clean pass follows the standard denoising process and records the latent state z_t^c together with query and key representations Q_t^c and K_t^c from selected self-attention layers at each denoising timestep. We denote these timestep-aligned references by $\mathcal{T}_{\text{clean}} = \{z_t^c, Q_t^c, K_t^c\}_t$.

The watermarked pass starts from z_T^w constructed in Eq. 1 and follows the same sampling configuration and text condition. At each step, spectral projection retains the watermarked components within the protected region while restoring complementary coefficients to their clean references. In parallel, query-key injection reuses the clean attention affinities while retaining the watermarked value representations. Together, these operations constrain latent drift and stabilize feature interactions during watermarked generation without model training or per-image optimization.

Watermark Verification. TrajRing retains the original inversion-based verification procedure of Tree-Ring Wen et al. (2023). Given a query image x', DDIM inversion Song et al. (2020) estimates its initial latent $\hat{z}_T$. The recovered spectrum $\mathcal{F}(\hat{z}_T)$ is compared with the private key W within the watermark mask M_r using the original Tree-Ring detection score and decision rule. Verification requires neither the clean reference trajectory nor an additional learned detector.

4.2 TRAJECTORY-GUIDED WATERMARKED SAMPLING

Given $\mathcal{T}_{\text{clean}}$, the watermarked pass combines trajectory-aware spectral projection and clean-reference query-key injection at each denoising step, as detailed below.

Trajectory-aware spectral projection. Let z_t^w and z_t^c denote the watermarked and clean latent states at timestep t, respectively. A direct replacement of z_t^w by z_t^c would eliminate the watermark-dependent variation. Instead, TrajRing retains the current watermarked Fourier components within the protected region and replaces complementary components with their clean-reference counterparts at the same timestep. Specifically, the projected spectrum is defined as

$$\widetilde{Z}_t = M_r \odot \mathcal{F}(z_t^w) + (1 - M_r) \odot \mathcal{F}(z_t^c), \qquad \widetilde{z}_t = \mathcal{F}^{-1}(\widetilde{Z}_t). \tag{2}$$

The first term retains the current watermarked components within the protected spectral region, whereas the second restores the complementary components to their timestep-aligned clean states. Repeated projection constrains the accumulation of deviations in complementary frequencies while allowing the coefficients within the protected region to evolve with the watermarked trajectory.

Clean-reference query-key injection. Spectral projection aligns complementary Fourier components with the clean reference, while watermark-dependent differences remain within the protected region. These differences may affect spatial interactions during denoising. To complement this spectral constraint with spatial guidance, we reuse queries and keys from the clean reference trajectory in selected self-attention layers while retaining value representations from the watermarked branch.

At timestep t, the query and key representations of the watermarked branch are replaced by their clean counterparts, $Q_t^w \leftarrow Q_t^c$ and $K_t^w \leftarrow K_t^c$, while the value representation V_t^w remains unchanged. The resulting attention output is

$$O_t^w = \text{Softmax}\left(\frac{Q_t^c (K_t^c)^\top}{\sqrt{d}}\right) V_t^w. \tag{3}$$

where d is the query/key dimension. The clean queries and keys determine spatial attention affinities, while V_t^w supplies feature content from the current watermarked branch. This design guides interactions among watermarked features without replacing their value representations with clean features.

After spectral projection, the projected latent $\widetilde{z}_t$ is passed through the denoising model, where the selected self-attention layers are guided by the corresponding clean query-key representations. The resulting state becomes z_{t-1}^w, and the same procedure is repeated at the next timestep.

5 EXPERIMENTS

We evaluate TrajRing's paired fidelity and watermark robustness in Sec. 5.2, computational efficiency in Sec. 5.3, and applicability across diffusion backbones in Sec. 5.4. We then examine component contributions and the effect of projection coverage in Sec. 5.5.

5.1 EXPERIMENTAL SETUP

Datasets and baselines. We evaluate on DiffusionDB Wang et al. (2023) and MS-COCO Lin et al. (2014), randomly sampling 5,000 prompts from each dataset. Paired clean and watermarked generations use matched prompts and sampling settings. For methods that watermark an existing image, the input image serves as the clean reference. We compare TrajRing with post-processing watermarking methods, including DwtDct, DwtDctSvd, RivaGAN Zhang et al. (2019), and StegaStamp Tancik et al. (2020), and diffusion-based methods, including AquaLoRA Feng et al. (2024), Gaussian Shading Yang et al. (2024), Tree-Ring Wen et al. (2023), ROBIN Huang et al. (2024), ZoDiac Zhang et al. (2024), and PT-Mark Wang et al. (2026b). We use official implementations where available and apply the same attack protocols to all methods.

Metrics and attacks. Paired fidelity is measured using PSNR, SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), and DINOv2 Oquab et al. (2023) cosine similarity between ℓ_2-normalized CLS embeddings. FID Heusel et al. (2017) additionally measures the distribution gap

Table 1: Paired fidelity and watermark robustness on DiffusionDB and MS-COCO. Fidelity is evaluated against unwatermarked references. Bold indicates the best results within each dataset

Dataset	Method	Image Fidelity					Watermark Robustness (AUC ↑)							
		PSNR ↑	SSIM ↑	DINOv2 ↑	FID ↓	LPIPS ↓	Clean	JPEG	Crop	Blur	Noise	Bright.	Rot.	Avg.
DiffusionDB	DwtDct	38.08	0.97	**0.99**	**1.34**	**0.01**	0.86	0.60	0.62	0.58	0.56	0.59	0.57	0.63
	DwtDctSvd	38.31	**0.98**	**0.99**	3.92	**0.01**	0.99	0.63	0.62	0.82	0.68	0.61	0.59	0.71
	RivaGAN	**40.37**	**0.98**	**0.99**	6.28	**0.01**	**1.00**	0.88	**0.99**	0.90	0.80	0.91	0.60	0.87
	StegaStamp	28.41	0.91	0.96	25.14	0.03	**1.00**	**1.00**	0.73	0.96	0.88	0.93	0.63	0.88
	AquaLoRA	9.68	0.28	0.66	64.80	0.62	0.93	0.88	0.84	0.91	0.79	0.85	0.70	0.84
	Gaussian Shading	10.11	0.32	0.67	44.90	0.60	**1.00**	**1.00**	0.71	0.99	0.89	**1.00**	0.66	0.89
	Tree-Ring	15.17	0.60	0.76	15.96	0.43	**1.00**	**1.00**	0.98	**1.00**	0.93	0.99	**0.97**	**0.98**
	ROBIN	23.91	0.77	0.92	26.30	0.12	0.99	0.99	**0.99**	**1.00**	**0.98**	0.99	0.95	**0.98**
	ZoDiac	25.88	0.94	0.98	12.90	0.05	**1.00**	**1.00**	0.97	**1.00**	**0.98**	**1.00**	0.67	0.95
	PT-Mark	27.55	0.95	0.98	10.95	0.03	**1.00**	**1.00**	0.97	**1.00**	0.97	**1.00**	0.95	**0.98**
	TrajRing	27.86	0.96	**0.99**	2.99	0.03	**1.00**	**1.00**	0.98	**1.00**	0.95	0.98	**0.97**	**0.98**
MS-COCO	DwtDct	39.95	0.98	**1.00**	**1.48**	**0.01**	0.87	0.58	0.60	0.57	0.55	0.57	0.55	0.61
	DwtDctSvd	39.18	**0.99**	0.99	3.42	**0.01**	**1.00**	0.61	0.60	0.80	0.66	0.59	0.58	0.69
	RivaGAN	**40.31**	0.98	0.99	5.12	0.02	**1.00**	0.86	0.98	0.88	0.78	0.90	0.57	0.85
	StegaStamp	28.13	0.92	0.97	17.80	0.03	**1.00**	**1.00**	0.71	0.95	0.86	0.92	0.60	0.86
	AquaLoRA	8.75	0.23	0.64	61.90	0.66	0.91	0.86	0.82	0.89	0.77	0.83	0.69	0.82
	Gaussian Shading	8.92	0.25	0.65	38.75	0.65	**1.00**	0.99	0.69	0.99	0.88	0.99	0.64	0.88
	Tree-Ring	12.64	0.52	0.73	16.78	0.50	**1.00**	**1.00**	0.97	**1.00**	0.97	0.99	**0.99**	**0.99**
	ROBIN	22.71	0.76	0.94	19.25	0.12	**1.00**	0.99	**0.99**	**1.00**	**0.98**	0.99	0.95	**0.99**
	ZoDiac	24.38	0.87	0.98	15.80	0.07	**1.00**	0.99	0.97	**1.00**	0.97	**1.00**	0.66	0.94
	PT-Mark	26.82	0.91	0.97	7.55	0.04	**1.00**	**1.00**	0.98	**1.00**	0.97	**1.00**	0.95	**0.99**
	TrajRing	27.55	0.96	0.99	2.28	0.02	**1.00**	**1.00**	0.97	**1.00**	**0.98**	0.98	0.97	**0.99**

between clean and watermarked image sets. Watermark robustness is evaluated using each method's original verification procedure. We report detection AUC without attacks and under six common transformations: JPEG compression, cropping, blur, noise, brightness adjustment, and rotation. Ablations additionally report average TPR at 1% FPR. Reconstruction and regeneration attacks are evaluated in Appendix A.2.

Implementation details. Unless otherwise stated, experiments use Stable Diffusion v2.1-base with FP16 inference, 512×512 resolution, 50-step DPM-Solver sampling Lu et al. (2022), and a classifier-free guidance scale of 7.5 Ho & Salimans (2022). The default TrajRing configuration embeds a radius-10 watermark in latent channel 3 (zero-based). Clean-reference Q/K features are injected into the nine self-attention layers in the U-Net upsampling blocks, while retaining watermarked V representations. Spectral projection uses the same watermark mask M_r as watermark embedding and verification. Further implementation details are provided in Appendix A.1.

5.2 Main Comparison

Paired fidelity preservation. Table 1 shows consistent fidelity improvements over Tree-Ring on both datasets. TrajRing increases PSNR by 12.69 dB on DiffusionDB and 14.91 dB on MS-COCO, while reducing LPIPS from 0.43 to 0.03 and from 0.50 to 0.02, respectively. The accompanying improvements in SSIM and DINOv2 similarity further support the preservation of reference structure and content. Among the evaluated diffusion-based methods, TrajRing achieves the highest PSNR and lowest FID on both datasets, indicating closer agreement with clean references at both the individual-image and distribution levels.

Robust watermark verification. These fidelity gains are achieved while maintaining average detection AUCs of 0.98 and 0.99 on DiffusionDB and MS-COCO, respectively, matching the reported averages of Tree-Ring. Performance varies across individual transformations: on MS-COCO, TrajRing increases noise AUC from 0.97 to 0.98, while rotation AUC decreases from 0.99 to 0.97. Post-processing baselines achieve higher paired PSNR but lower average detection AUC, showing that close pixel-level agreement alone does not ensure robust verification. Results under reconstruction and regeneration attacks are reported in Appendix A.2.

Qualitative comparison. Figure 3 compares watermarked outputs with their paired clean references. Tree-Ring introduces visible changes in flower arrangements, character structure, and portrait pose, producing pronounced structured residuals. ROBIN, ZoDiac, and PT-Mark preserve more of the reference composition, although differences in shape, color, and local details remain. TrajRing

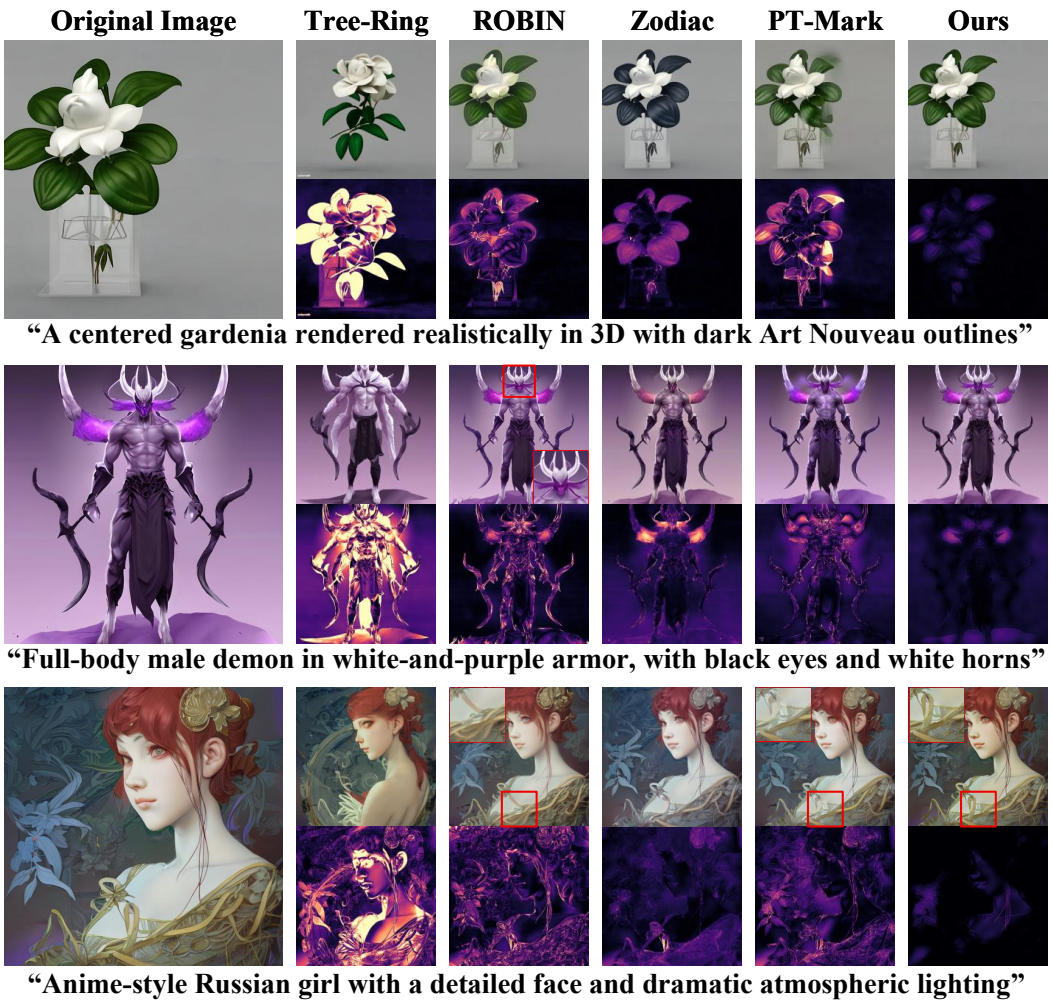


Figure 3: Qualitative comparison with paired clean generations. For each prompt, the leftmost column shows the unwatermarked reference. Each remaining column shows a watermarked output (top) and its absolute residual map relative to the reference (bottom). Insets magnify the regions outlined in red.

produces outputs closer to the clean references, with better-preserved fine details and weaker residual responses in these examples.

5.3 Efficiency

Table 2 compares offline optimization time and online runtime per image. Tree-Ring has the lowest online latency among the compared methods. ROBIN achieves low online latency by reusing an offline-optimized watermark and its associated prompt guidance signal. However, this optimization is specific to each watermark. In deployments that assign a distinct watermark to each user, separate offline optimization is therefore required, causing the aggregate setup cost to grow with the number of users. ZoDiac iteratively optimizes the initial latent, while PT-Mark optimizes timestep-specific unconditional embeddings. Both methods incur substantial per-image computational overhead.

Table 2: Runtime comparison. Offline time is in seconds, online runtime is in seconds per image.

Method	Offline opt.	Per-image opt.	Online
Tree-Ring	–	No	2.43
ROBIN	1133.07	No	2.85
ZoDiac	–	Yes	466.18
PT-Mark	–	Yes	114.09
TrajRing	–	No	4.77

TrajRing uses clean-reference guidance without offline or per-image optimization. Its two-pass procedure takes 4.77 s per image, including both clean-reference and watermarked sampling. Under the evaluated configurations, this corresponds to approximately 98× and 24× speedups over ZoDiac and PT-Mark, respectively.

5.4 Cross-Architecture Evaluation

To assess the applicability of TrajRing beyond SD 2.1, we extend the evaluation to SD 1.4, SDXL Podell et al. (2024), and PixArt-α Chen et al. (2024). As shown in Table 3, TrajRing consistently improves fidelity to paired clean references across all four backbones, increasing PSNR by 10.94–18.80 dB over Tree-Ring and reducing LPIPS to 0.02–0.05. These fidelity gains are accompanied by high watermark detection performance. Clean AUC remains 1.00 across all backbones, while average AUC under transformations ranges from 0.93 to 0.99. TrajRing achieves average AUC comparable to that of Tree-Ring on SD 1.4 and SD 2.1, with small decreases of 0.01 and 0.02 on SDXL and PixArt, respectively.

Table 3: Cross-architecture comparison of Tree-Ring and TrajRing.

Backbone	Method	PSNR↑	LPIPS↓	Clean AUC↑	Avg. AUC↑
SD 1.4	Tree-Ring	12.66	0.50	1.00	0.97
	TrajRing	**26.19**	**0.03**	1.00	0.97
SD 2.1	Tree-Ring	12.64	0.50	1.00	0.99
	TrajRing	**27.55**	**0.02**	1.00	0.99
SDXL	Tree-Ring	10.29	0.58	1.00	0.94
	TrajRing	**29.09**	**0.05**	1.00	0.93
PixArt-α	Tree-Ring	13.61	0.61	1.00	0.95
	TrajRing	**24.55**	**0.05**	1.00	0.93

5.5 Ablation and Mechanism Analysis

We examine how spectral projection and clean-reference attention guidance contribute to fidelity preservation and watermark robustness. We further vary projection coverage to investigate how the extent of trajectory correction affects the generated content.

5.5.1 Component Ablations

Complementary roles of projection and attention guidance. Table 4 evaluates spectral projection and Q/K injection individually and in combination, using Tree-Ring as the baseline. Both components independently improve paired fidelity, with projection alone providing the larger improvement. Adding Q/K injection to projection further increases PSNR from 25.03 to 27.55 dB and reduces LPIPS from 0.04 to 0.02. Average AUC remains comparable across variants, while the combined model raises average TPR from 75.17% for projection alone to 78.33%, approaching Tree-Ring's 78.83%.

Choice of attention components. To isolate the effect of attention injection, we disable spectral projection and replace different subsets of Q, K, and V with their clean-reference counterparts. Table 5 reveals a clear trade-off between reference matching and watermark robustness. Q/V and Q/K/V injection achieve the highest fidelity, but their average TPR drops to 43.50% and 41.67%, respectively, compared with 76.67% for Q/K. All variants achieve a clean AUC of 1.00, but Q/V and Q/K/V are less robust under transformations. With spectral projection disabled, Q/K substantially improves fidelity over K-only injection while maintaining nearly the same average TPR. Compared with Q-only injection, it yields slightly lower fidelity but improves average TPR by 6.00 percentage points. These comparisons motivate the use of Q/K injection to transfer clean-reference attention affinities while retaining watermarked value representations.

Table 4: Component ablation of TrajRing.

Variant	Q/K	Projection	Image Fidelity				Watermark Detection		
			PSNR ↑	SSIM ↑	DINOv2 ↑	LPIPS ↓	Clean AUC ↑	Avg. AUC ↑	Avg. TPR (%) ↑
Tree-Ring	–	–	12.64	0.52	0.73	0.50	1.00	0.99	**78.83**
Q/K only	✓	–	21.46	0.87	0.97	0.09	1.00	0.99	76.67
Projection only	–	✓	25.03	0.94	0.98	0.04	1.00	0.99	75.17
TrajRing	✓	✓	**27.55**	**0.96**	**0.99**	**0.02**	1.00	0.99	78.33

Table 5: Ablation of self-attention Q/K/V injection.

Injection	Image Fidelity				Watermark Detection		
	PSNR ↑	SSIM ↑	DINOv2 ↑	LPIPS ↓	Clean AUC ↑	Avg. AUC ↑	Avg. TPR (%) ↑
Q	22.53	0.89	0.98	0.08	1.00	0.98	70.67
K	13.34	0.54	0.44	0.53	1.00	**0.99**	**76.83**
V	25.47	0.91	0.98	0.05	1.00	0.94	61.67
Q/K	21.46	0.87	0.97	0.09	1.00	**0.99**	76.67
Q/V	30.02	**0.96**	**1.00**	**0.02**	1.00	0.89	43.50
K/V	22.68	0.89	0.98	0.07	1.00	0.94	65.33
Q/K/V	**30.50**	**0.96**	**1.00**	**0.02**	1.00	0.88	41.67

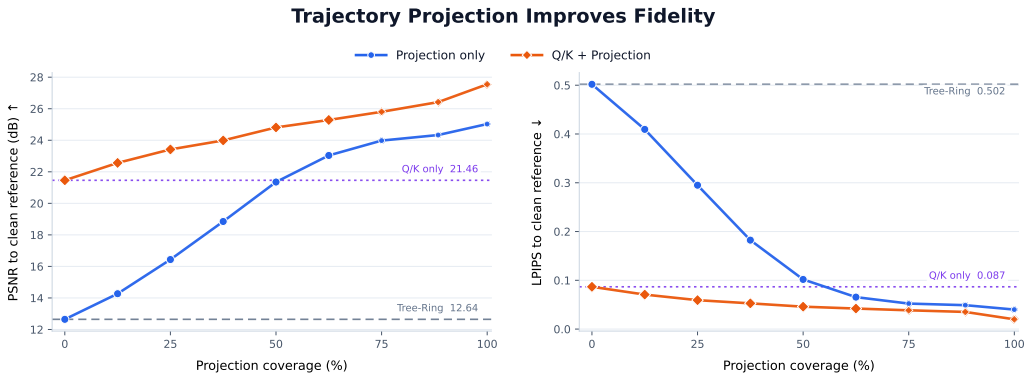


Figure 4: Effect of complementary projection coverage on paired fidelity.

5.5.2 Effect of Projection Coverage

We vary the fraction of complementary Fourier coefficients replaced with timestep-aligned clean references at each denoising step, keeping the watermark initialization and protected region fixed. We evaluate spectral projection both with and without Q/K injection to examine how projection coverage affects paired fidelity.

Figure 4 shows that increasing projection coverage consistently raises PSNR and lowers LPIPS, both with and without Q/K injection. The same watermark initialization therefore leads to different levels of final content deviation depending on the correction applied during denoising. Q/K injection provides additional fidelity gains across the evaluated coverage range, further supporting its complementary role. Because spectral projection acts only on complementary frequencies, these results show that paired fidelity can be improved without directly replacing coefficients within the protected watermark region.

6 Conclusion

We identified watermark-induced trajectory drift: localized watermark perturbations induce growing deviations in complementary frequencies during denoising. Based on this finding, we proposed TrajRing, which combines trajectory-aware spectral projection with clean-reference query-key injection to suppress drift without model training or per-image optimization. Experiments on DiffusionDB and MS-COCO, together with evaluations across four diffusion backbones, demonstrate consistent improvements in fidelity to paired clean generations while maintaining robust watermark detection. Under the evaluated configuration, TrajRing takes 4.77 s per image for both sampling passes, achieving approximately 24× and 98× speedups over PT-Mark and ZoDiac, respectively. Component ablations further support the complementary roles of spectral projection and attention guidance.

AI USE STATEMENT

We used ChatGPT solely for translation and language polishing of the manuscript. We reviewed and revised all AI-assisted text and take full responsibility for the final manuscript, including its scientific content and conclusions.

ETHICS STATEMENT

Our experiments use only prompts from publicly available datasets and model-generated images. This work aims to support the responsible use of generative models by enabling provenance verification while reducing the impact of watermarking on generated content.

REPRODUCIBILITY STATEMENT

We will publicly release the implementation and evaluation code to facilitate reproduction of our results. Section 4. describes the proposed method and watermark verification procedure. Section 5.1 presents the datasets, baselines, evaluation metrics, and generation settings, while Appendix A provides additional implementation details, controlled experimental protocols, and extended analyses.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024.

Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*, 2024a.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, pp. 338–354. Springer, 2024b.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint arXiv:2405.11135*, 2024.

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*, pp. 22466–22477, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Zheling Meng, Bo Peng, and Jing Dong. Latent watermark: Inject and detect watermarks in latent diffusion space. *IEEE Transactions on Multimedia*, 27:3399–3410, 2025.

Rui Min, Sen Li, Hongyang Chen, and Minhao Cheng. A watermark-conditioned diffusion model for ip protection. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2024.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 31620–31631, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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*, pp. 2117–2126, 2020.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930. IEEE, 2023.

Yaopeng Wang, Qingliang Wang, Zhibo Wang, Huiyu Xu, Jiacheng Du, Qiu Wang, Jia-Li Yin, and Kui Ren. Lora-key: User-centric lora watermarking for text-to-image diffusion models. *arXiv preprint arXiv:2605.29569*, 2026a.

Yaopeng Wang, Huiyu Xu, Zhibo Wang, Jiacheng Du, Zhichao Li, Yiming Li, Qiu Wang, and Kui Ren. Pt-mark: Invisible watermarking for text-to-image diffusion models via semantic-aware pivotal tuning. *IEEE Transactions on Dependable and Secure Computing*, 2026b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 893–911, 2023.

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8213–8224, 2025.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063, 2023.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12162–12171. IEEE, 2024.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019.

Lijun Zhang, Xiao Liu, Antoni Martin, Cindy Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 657–672, 2018.

A Additional Experimental Details and Results

A.1 Experimental Protocol and Implementation Details

Controlled studies follow the MS-COCO setting in Section 5.1. Within each comparison, prompts, initial Gaussian latents, generation seeds, watermark keys, and sampler settings are fixed except for the factor being varied. Fidelity is measured against the paired clean generation.

We use Stable Diffusion v2.1-base in FP16 at 512×512 resolution, with 50 second-order DPM-Solver multistep updates and a guidance scale of 7.5. A radius-10 watermark is embedded in latent channel 3 (zero-based). Clean Q/K representations are injected into nine self-attention layers in the U-Net upsampling blocks, while V remains from the watermarked branch.

At each step, z_t^w is first projected using the timestep-aligned clean reference z_t^c. The projected latent $\tilde{z}_t$ then undergoes a denoising update with clean Q/K injection to produce z_{t-1}^w. Solver iterations are indexed by $k = 0, \ldots, 49$, distinct from diffusion timestep t. The final update produces z_0^w, which is decoded by the VAE without further projection. Complementary frequencies refer to coefficients outside the watermark mask M_r defined in the main paper.

Conventional attacks. We apply JPEG compression (quality 25), random resized cropping (area scale 0.75; fixed aspect ratio 0.75), Gaussian blur (radius 4), noise with nominal standard deviation 0.1 relative to the $[0, 1]$ pixel range, brightness jitter using `ColorJitter(brightness=6)`, and a fixed 75° rotation. For brightness jitter, each invocation independently samples a factor $\beta \sim \mathcal{U}(0, 7)$. For noise, the evaluated implementation converts 255ϵ, with $\epsilon \sim \mathcal{N}(0, 0.1^2)$, to `uint8` before adding it to the 8-bit RGB image. Each attack precedes latent inversion and watermark detection. TPR is reported at 1% FPR. Table 1 additionally includes the clean condition, whereas Table 6 averages its six reconstruction-based attacks.

Reconstruction-based attacks. Inputs are converted to RGB and resized to 512×512. VAE reconstruction uses the Stable Diffusion v2.1-base VAE. Learned compression uses pretrained Cheng2020-anchor, MBT2018, and BMSHJ2018-hyperprior models from CompressAI at quality level 1. SDEdit uses Stable Diffusion v2.1-base with strength 0.2 and a 50-step sampling schedule. Following ZoDiac, Regenerate perturbs the VAE latent at diffusion timestep 60 and denoises from the corresponding position in a 50-step PNDM schedule. Both diffusion-based attacks use an empty prompt, guidance scale 7.5, and seed 1024.

A.2 Robustness to Reconstruction-Based Attacks

Table 6 evaluates robustness under six reconstruction-based attacks, including VAE reconstruction, learned compression, SDEdit, and diffusion-based regeneration. TrajRing achieves an AUC of 1.00 under five attacks and 0.99 under SDEdit, yielding an average AUC of 1.00. At the reported precision, these results match Tree-Ring under all evaluated attacks. Together with the fidelity improve-

Table 6: Detection AUC under reconstruction and regeneration attacks. Avg. averages the six listed attack conditions. Best values, including ties at the reported two-decimal precision, are bold.

Method	Reconstruction / Regeneration Attacks (AUC ↑)						
	VAE Rec.	Cheng2020	MBT2018	SDEdit	BMSHJ	Regenerate	Avg.
DwtDct	0.51	0.50	0.49	0.49	0.48	0.50	0.50
DwtDctSvd	0.70	0.46	0.51	0.49	0.47	0.67	0.55
RivaGAN	0.98	0.57	0.57	0.80	0.54	0.90	0.73
StegaStamp	**1.00**	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**
Tree-Ring	**1.00**	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**
ROBIN	**1.00**	0.99	0.99	**1.00**	0.99	**1.00**	0.99
ZoDiac	**1.00**	0.98	0.96	0.94	0.96	0.91	0.96
PT-Mark	**1.00**	0.99	0.98	0.99	0.99	**1.00**	0.99
TrajRing	**1.00**	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**

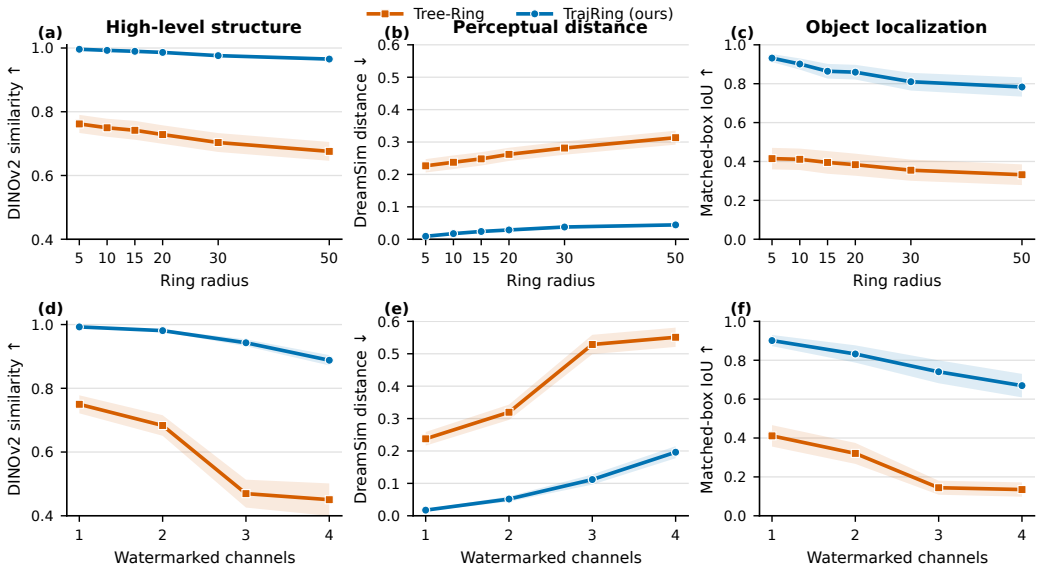


Figure 5: Semantic and structural preservation under varying watermark support. The top row varies the ring radius in one latent channel, while the bottom row varies the number of watermarked channels at radius 10. All metrics compare the watermarked output with its paired clean generation. Lines show means, and shaded regions indicate 95% confidence intervals.

ments in Table 1, this shows that TrajRing improves agreement with the clean generation while preserving robust watermark verification under reconstruction-based attacks.

A.3 Additional Semantic and Structural Preservation

We complement the pixel-level, perceptual, and representation-based metrics in the main paper with DreamSim distance, caption-level BERTScore F1, detected-class Jaccard similarity, and matched-object box IoU. Images without a clean-reference DETR detection above the fixed confidence threshold are excluded from the DETR-based object metrics.

Table 7 evaluates the default radius-10 watermark embedded in one latent channel. Compared with Tree-Ring, TrajRing achieves higher caption consistency, object-class agreement, and localization agreement with the paired clean generation. These results provide additional evidence of content preservation beyond pixel-level similarity.

Figure 5 further varies the ring radius and the number of watermarked latent channels. As the watermark support expands, Tree-Ring shows progressively lower semantic and structural agreement with the paired clean generation. TrajRing maintains higher agreement throughout the evaluated range, although its agreement with the clean reference also decreases as the watermark support grows, particularly when more latent channels are watermarked.

Table 7: Additional paired semantic and structural metrics under the default watermark setting. Caption denotes BERTScore F1, and class overlap denotes Jaccard similarity. Best values are bold.

Method	DreamSim ↓	Caption ↑	Class overlap ↑	Box IoU ↑
Tree-Ring	0.2376	0.7791	0.7519	0.4112
TrajRing	**0.0174**	**0.9254**	**0.9179**	**0.9013**

A.4 Extended Component and Design Ablations

A.4.1 Value Injection with Spectral Projection

The main-paper attention ablation disables spectral projection to isolate different attention components. Here, we compare Q/K and Q/K/V injection with projection enabled (Table 8) to examine the role of the watermarked value representations.

Table 8: Q/K versus Q/K/V injection with spectral projection enabled. The Q/K configuration corresponds to complete TrajRing. Avg. AUC and Avg. TPR average the six conventional transformations; TPR is measured at 1% FPR. Best values are bold.

Injection	Projection	PSNR ↑	LPIPS ↓	Clean AUC ↑	Avg. AUC ↑	Avg. TPR (%) ↑
Q/K	Yes	27.55	0.02	**1.00**	**0.99**	**78.33**
Q/K/V	Yes	**33.76**	**0.01**	0.99	0.89	49.83

Table 9: Scope of clean-reference Q/K injection with spectral projection disabled. Avg. TPR is measured at 1% FPR, averaged over the six conventional transformations, and reported in percent.

Scope	Layers	PSNR ↑	DINOv2 ↑	Avg. TPR (%) ↑
No Q/K	0	12.64	0.73	**78.83**
Mid only	1	12.62	0.75	78.33
Down only	6	16.48	0.89	77.67
All up-block	9	21.46	0.97	76.67
All self-attn.	16	**22.55**	**0.98**	72.67

Replacing V with its clean-reference counterpart increases PSNR from 27.55 to 33.76 dB and reduces LPIPS from 0.02 to 0.01, but decreases Avg. AUC from 0.99 to 0.89 and Avg. TPR from 78.33% to 49.83%. This trade-off supports retaining watermarked V while using clean-reference Q/K for spatial guidance.

A.4.2 Attention-Layer Scope

We disable spectral projection and vary the scope of clean-reference Q/K injection (Table 9). Injection into the middle block alone has little effect, whereas extending Q/K guidance to the upsampling blocks substantially improves paired fidelity. Applying Q/K injection to all 16 self-attention layers further increases PSNR from 21.46 to 22.55 dB and DINOv2 similarity from 0.97 to 0.98, but reduces Avg. TPR from 76.67% to 72.67%. We therefore use the nine self-attention layers in the upsampling blocks as the default scope, providing substantial fidelity recovery while avoiding the larger robustness reduction observed with all-layer injection.

A.4.3 Projection Schedule

With Q/K injection fixed, we vary the solver iterations at which spectral projection is enabled (Table 10). An interval a–b denotes projection before the denoising updates at iterations $k = a, \ldots, b$; the full 0–49 schedule corresponds to complete TrajRing in Table 4. Both delaying the start of projection and terminating it earlier reduce fidelity relative to the full schedule. Detection varies comparatively little across the tested windows: the 0–49 and 5–49 schedules both achieve an Avg. TPR of 78.33%. The full schedule additionally yields the best PSNR and LPIPS, supporting projection throughout denoising for content preservation.

Table 10: Projection-window ablation with Q/K injection fixed. Avg. TPR is reported at 1% FPR.

Projection steps	PSNR ↑	LPIPS ↓	Avg. TPR (%) ↑
0–25	25.14	0.04	74.67
0–35	25.72	0.04	73.83
0–49	**27.55**	**0.02**	**78.33**
5–49	26.41	0.03	**78.33**
10–49	26.05	0.03	77.83
15–49	25.75	0.04	77.67

A.5 Additional Trajectory Analysis

A.5.1 Controlled Projection-Strength Intervention

The main-paper coverage experiment varies the fraction of complementary Fourier coefficients constrained during denoising. Here, we instead fix the projection region and vary the projection strength $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, with Q/K injection disabled. At each denoising step, α controls how strongly the complementary coefficients of the current watermarked latent are moved toward their timestep-aligned clean references before the corresponding denoising update.

Table 11: Projection-strength intervention on 100 paired samples with Q/K injection disabled. A constant strength α is applied before each denoising update. Residuals are unnormalized measurements at the final latent state.

α	$R_0^{\perp} \downarrow$	$R_0^{\text{full}} \downarrow$	PSNR ↑	DINOv2 ↑	Clean AUC ↑
0	5127.94	5540.11	12.64	0.73	1.00
0.25	138.25	1341.47	23.69	0.97	1.00
0.50	66.99	1205.52	24.54	0.97	1.00
0.75	32.13	1162.11	24.86	**0.98**	1.00
1.00	**10.93**	**1139.06**	**25.03**	**0.98**	1.00

At solver iteration k, let $\widehat{\mathbf{z}}_k^w$ and $\widehat{\mathbf{z}}_k^c$ denote the Fourier representations of the current watermarked latent and its timestep-aligned clean reference, respectively. The relaxed projection is defined as

$$\widehat{\mathbf{z}}_k^{\alpha} = M_r \odot \widehat{\mathbf{z}}_k^w + (1 - M_r) \odot [(1-\alpha)\widehat{\mathbf{z}}_k^w + \alpha \widehat{\mathbf{z}}_k^c], \tag{4}$$

where M_r denotes the watermark-region mask defined in the main paper. The projected latent is then obtained by inverse Fourier transform and passed to the denoising model, following the same update order as the complete TrajRing framework.

Thus, $\alpha = 0$ leaves the complementary coefficients unchanged and recovers Tree-Ring, whereas $\alpha = 1$ fully aligns them with the paired clean reference, corresponding to the projection-only variant. Intermediate values implement under-relaxed projection, removing an α fraction of the current complementary-frequency discrepancy at each step. The same α is applied throughout all 50 solver iterations, and each setting is generated independently rather than obtained by post-hoc interpolation. Prompts, initial latents, watermark keys, watermark region, and sampler settings are fixed across all strength values. The study uses 100 paired samples.

For each strength, let

$$\delta_0^{(\alpha)} = z_0^{w,(\alpha)} - z_0^c \tag{5}$$

denote the final latent residual. Let $P_{\perp} = 1 - M_r$ select Fourier coefficients outside the watermark region. We measure the complementary residual as

$$R_0^{\perp}(\alpha) = \left\| P_{\perp} \odot \mathcal{F}\left(\delta_0^{(\alpha)}\right) \right\|_2, \tag{6}$$

and the full-spectrum residual as

$$R_0^{\text{full}}(\alpha) = \left\| \mathcal{F}\left(\delta_0^{(\alpha)}\right) \right\|_2. \tag{7}$$

Both quantities are averaged over the 100 paired samples and are computed using the unnormalized Fourier transform. The complementary residual isolates deviations outside the watermark region, whereas the full residual measures the overall spectral departure from the paired clean latent.

Table 11 shows a monotonic reduction in both residual measures as α increases, accompanied by improved paired fidelity. The largest change occurs between $\alpha = 0$ and 0.25, indicating that even partial correction of the complementary frequencies removes a substantial fraction of the accumulated deviation. Increasing α further continues to reduce the residuals, while the fidelity gains become more gradual. These results complement the coverage experiment in the main paper: with the watermark initialization and projection region fixed, both the extent and strength of complementary-frequency constraints affect the final deviation from the paired clean generation. Clean AUC remains 1.00 across all tested strengths.

A.5.2 Spatial Evolution of Latent Deviations

To complement the frequency-domain measurements above, we further examine how deviations from the clean trajectory are distributed spatially throughout denoising.

Figure 6 visualizes spatial residuals between the watermarked latents and their timestep-aligned clean references at representative denoising stages. These maps complement the frequency-domain analysis by revealing where trajectory deviations emerge and how their spatial structure evolves.

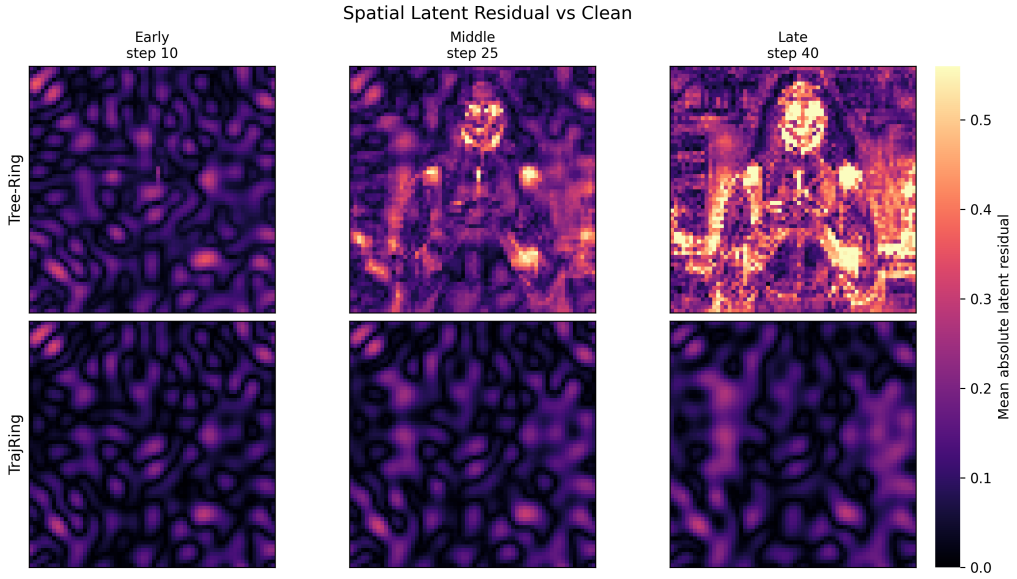


Figure 6: Spatial evolution of latent residuals relative to the paired clean trajectory. Residuals are averaged across latent channels and visualized using a shared color scale.

As denoising proceeds, Tree-Ring develops increasingly structured residuals around object regions and scene boundaries, indicating that the initially localized watermark perturbation is progressively reflected in spatially organized latent differences. In contrast, TrajRing maintains substantially weaker residual responses and remains more closely aligned with the clean trajectory across the displayed steps. This spatial evolution is consistent with watermark-induced trajectory drift and provides complementary evidence that TrajRing suppresses the structured deviations that emerge during denoising.