SPHERICALMARK: THIRD-MOMENT PRESERVING DIFFUSION WATERMARKS WITH ADAPTIVE AUGMENTATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion watermarking embeds a learnable signature into the initial noise so that generated images remain traceable without degrading generation quality. Existing in-processing methods normalize this signature to match only the first two moments of the target Gaussian—mean and variance—leaving its skewness unconstrained; the resulting distributional mismatch is a structural limitation that current normalization schemes do not address. To address this gap, we propose SphericalMark, which introduces spherical third-moment normalization to explicitly constrain the third moment of the injected signature through a differentiable regularizer that combines bounded odd-symmetric transformation (tanh), re-centering, and norm restoration. This yields a watermark distribution more faithful to the target Gaussian while remaining fully end-to-end optimizable. To strengthen detection robustness, we further introduce an adaptive augmentation sampler (OAS) that continually reallocates training probability toward the perturbations on which the detector currently performs worst. Evaluated on Stable Diffusion 2.1, SphericalMark achieves 78% lower quality degradation than state-of-the-art SERUM (ΔFID: 0.81 $\rightarrow$ 0.18) while maintaining competitive robustness (average TPR: 98.55% @ 1% FPR, within 1.5% of SERUM for 7/8 perturbations). We validate the distributional benefits through direct statistical measurement and demonstrate that SphericalMark offers a practical, drop-in refinement that better respects the target noise distribution without sacrificing efficiency.

1 INTRODUCTION

In recent years, diffusion models (DMs) (Ho et al., 2020) have emerged as the leading paradigm for image generation, achieving state-of-the-art performance across many tasks and producing outputs that are increasingly difficult to distinguish from natural photographs (Rombach et al., 2022). However, this same capability to produce images that are nearly indistinguishable from real ones has raised critical concerns: generated content can be misused for impersonation and disinformation, and when synthetic images are published online and later scraped into future training datasets, they degrade the performance of generative models and amplify existing biases. Together, these challenges underscore the urgent need for reliable detection of generated content.

Watermarking has emerged as the de facto standard to facilitate such detection. It embeds imperceptible but algorithmically detectable signals into generated images, enabling distinction between generated and original content. In-processing approaches for diffusion models mix a learnable secret signature into the initial diffusion noise and train a lightweight external detector to recover its trace directly from the generated image, keeping the generator frozen and eliminating the need for costly diffusion inversion during detection (Kociszewski et al., 2026). The signature is injected by normalizing it to match the first two moments, mean and variance, of the target unit Gaussian before mixing it with the clean initial noise. While this normalization provides a theoretical guarantee of low divergence between the watermarked and non-watermarked noise distributions, it constrains only the mean and the variance: the third moment, skewness, is left unconstrained. Because the signature is optimized end-to-end to maximize detector accuracy, gradient descent has no incentive to keep it symmetric, and the injected signature can drift toward substantial skewness, reintroducing

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.

a distributional mismatch with the Gaussian prior the denoiser expects, precisely at the point where generation begins.

To overcome this limitation, several recent works have examined how faithfully a watermarked latent reproduces its target distribution, showing that the Kullback-Leibler divergence between the watermarked and clean noise distributions can be driven toward zero by constraining additional moments of the injected signal beyond the first two. This line of work establishes that distributional fidelity in diffusion watermarking is a controllable design choice rather than an unavoidable cost of embedding a detectable signal. A second, largely orthogonal limitation concerns how the detector itself is trained: existing pipelines typically sample perturbations for augmentation uniformly at random, which spends the majority of the training budget on transformations the detector already handles well while leaving its weakest augmentations, such as aggressive rotation, comparatively undertrained.

As a solution to these drawbacks, we propose SphericalMark, a refinement of initial-noise diffusion watermarking that introduces spherical third-moment normalization to explicitly constrain the skewness of the injected signature. A differentiable regularizer combines a bounded odd-symmetric mapping with re-centering and norm restoration, driving the signature toward zero mean, unit variance, and zero skewness while preserving a fixed norm, which confines the watermark to a distribution-preserving submanifold without breaking gradient flow to the underlying parameters. To close the second gap, we further introduce an adaptive augmentation sampler that maintains a priority score over the augmentation types and continually reallocates training probability toward whichever perturbation the detector currently handles worst, rather than sampling all perturbations with equal frequency. Consequently, SphericalMark keeps the diffusion generator entirely frozen and relies only on a lightweight external detector, preserving the fast injection and detection of initial-noise watermarking while directly addressing both the residual distributional mismatch and the uneven robustness that uniform augmentation sampling leaves behind.

We thoroughly evaluate SphericalMark on Stable Diffusion 2.1, validating the distributional benefits of spherical third-moment normalization through direct statistical measurement of the injected signature and assessing detection robustness under eight standard perturbation types. Our experiments demonstrate that progressive training with w=0.005 achieves near-perfect clean detection (99.96%) with 98.55% average robust TPR @ 1% FPR, while reducing quality degradation by 78% compared to state-of-the-art SERUM (ΔFID: 0.81 $\rightarrow$ 0.18). The adaptive augmentation sampler accounts for the largest share of this robustness gain. Taken together, these results indicate that explicitly controlling the third moment of the watermark signature and prioritizing hard augmentations during detector training improve distributional fidelity and detection robustness without sacrificing the efficiency that makes initial-noise watermarking practical.

Our contributions are:

- **Spherical third-moment normalization (SMP):** We introduce a differentiable regularizer that constrains watermark skewness, achieving 78% lower quality degradation than state-of-the-art SERUM (ΔFID: 0.81 $\rightarrow$ 0.18 with w=0.005) while maintaining competitive detection robustness across most perturbations.
- **Competitive robustness with quality focus:** SphericalMark maintains detection performance within 1.5% of SERUM for seven out of eight perturbations (average TPR: 98.55% @ 1% FPR), demonstrating that substantial quality improvement does not require sacrificing robustness on common transformations.
- **Progressive training framework:** We provide comprehensive analysis of quality-robustness trade-offs, demonstrating SphericalMark's suitability for high-fidelity applications where SERUM's quality degradation is prohibitive, such as professional photography and digital art platforms.

2 Related Work

Diffusion Models (DMs). DMs (Ho et al., 2020) have emerged as one of the most powerful classes of vision generative models, achieving state-of-the-art results in image synthesis. The central idea of diffusion modeling is to learn the reverse of a gradual noising process that destroys the structure of the training data. At inference, new samples are generated by iteratively denoising an initial

Gaussian noise map until a clean image is obtained. Latent Diffusion Models (LDMs) (Rombach et al., 2022) integrate a variational autoencoder to project images into a compressed latent space, substantially reducing the computational overhead of the iterative sampling process. Our method and evaluation target LDMs since they represent the dominant class of image generative models in current practice.

Image Watermarking. Watermarking outputs from vision generative models is indispensable since modern LDMs are capable of generating photorealistic images, and distinguishing synthetic from authentic content is increasingly required in practice. Existing approaches to watermark LDMs can be broadly categorized into post-processing and in-processing methods. *Post-processing* techniques embed watermarks into generated images after the sampling process; while extensively studied, these approaches are vulnerable to removal or degradation because the watermark is not tied to the generative process itself. In contrast, *in-processing* methods integrate watermarking directly into the generation pipeline, yielding watermarks that are harder to remove and more seamlessly embedded. Because of their stronger robustness, we focus on in-processing methods, which can be further divided into *tuning-based* and *tuning-free* approaches.

Tuning-based Watermarks. Stable Signature (Fernandez et al., 2023) is a representative tuning-based technique for LDMs: it fine-tunes the diffusion decoder so that a pre-trained extractor can recover an embedded binary message from the decoded image. While this allows for fast injection without added overhead during generation, it relies on a pixel-space extractor trained separately from the generator, which requires substantial fine-tuning time and generalizes poorly to attacks that the extractor was not trained against.

Tuning-free Watermarks. To avoid the costly fine-tuning of LDMs, tuning-free methods add a watermark directly to the initial diffusion noise. Tree-Ring (Wen et al., 2023) embeds the watermark in the frequency space of the initial noise and achieves robustness against common image transformations such as cropping and rotation, but detection requires inverting the diffusion process to retrieve the noise pattern, which is computationally expensive at inference time. RingID (Ci et al., 2024) extends Tree-Ring with multi-channel heterogeneous watermarking to support multi-key identification, inheriting the same inversion-based detection cost. Gaussian Shading (Yang et al., 2024) embeds bits directly into the initial noise and normalizes the resulting signal to preserve the mean and variance of the target Gaussian, so that the watermarked and clean noise distributions remain close under a first- and second-moment criterion; this normalization, however, leaves the third moment of the signal unconstrained. SERUM (Kociszewski et al., 2026) removes the DDIM-inversion requirement entirely by pairing the same first-two-moment normalization with a lightweight external detector trained directly on the generated image, achieving fast injection and fast detection without touching the frozen generator. Our detector architecture follows this design, and we directly build on SERUM as our baseline.

Adaptive Detector Training. Prior in-processing pipelines train detectors by sampling augmentations uniformly. We adapt Prioritized Experience Replay (Schaul et al., 2015): maintain a priority score per augmentation type and shift sampling probability toward whichever perturbation the detector currently misclassifies most. We directly compare SphericalMark to the tuning-free SERUM baseline, isolating the contribution of spherical third-moment normalization and adaptive augmentation sampling.

3 METHOD

SphericalMark maintains the frozen generator and lightweight detector architecture while introducing spherical third-moment normalization for watermark injection and adaptive augmentation for detector training. The system consists of three components: keyed initial-noise injection, spherical normalization regularizer, and adaptive augmentation training.

3.1 KEYED INITIAL-NOISE INJECTION

Let $\eta \sim \mathcal{N}(0, I)$ denote initial latent noise of dimension $d = c \cdot w \cdot h$ ($c = 4$, $w = h = 64$ for 512×512 SD). Let $\mathbf{A} \in \mathrm{R}^d$ be a secret learnable grid. We normalize,

$$\mathbf{A}' = \frac{\mathbf{A} - \mu(\mathbf{A})}{\sigma(\mathbf{A}) + \epsilon},$$

and inject it into the initial latent noise,

$$\eta_{\text{wm}} = \sqrt{1-\alpha} \cdot \eta + \sqrt{\alpha} \cdot \mathbf{A}',$$

where $\alpha \in [0, 1)$ is the watermark strength. The resulting latent follows

$$\eta_{\text{wm}} \sim \mathcal{N}\left(\sqrt{\alpha}\mathbf{A}', (1-\alpha)I\right).$$

Using the closed-form Kullback-Leibler divergence between Gaussians and the property $\|\mathbf{A}'\|_2^2 = d$ (enforced by spherical normalization),

$$D_{\text{KL}}(q\|p) = -\frac{d}{2}\log(1-\alpha),$$

where $p = \mathcal{N}(0, I)$. For $\alpha = 0.5$ and $d = 2^{14} = 16{,}384$, $D_{\text{KL}}(q\|p) \approx 5{,}678$. This provides a distributional diagnostic; it does not predict FID directly because FID is computed over generated pixel images and is sensitive to subsequent denoising and decoding. Figure 1 illustrates the core spherical moment preservation mechanism.

3.2 Spherical Second- and Third-Moment Normalization

Standard moment matching enforces zero mean and unit variance but permits non-zero skewness. Following spherical watermarking that achieves third-moment preservation through manifold projection, we apply an odd-symmetric transformation using the hyperbolic tangent:

$$\mathbf{A}' \leftarrow \tanh\left(\frac{\mathbf{A} - \mu(\mathbf{A})}{\sigma(\mathbf{A}) + \epsilon}\right), \tag{1}$$

followed by re-centering $\mathbf{A}' \leftarrow \mathbf{A}' - \mu(\mathbf{A}')$ and norm restoration $\mathbf{A}' \leftarrow \mathbf{A}' \cdot \sqrt{d}/\|\mathbf{A}'\|_2$, ensuring $\|\mathbf{A}'\|_2^2 = d$. The $\tanh$ transformation provides smooth gradients throughout the domain and guarantees zero skewness through its odd-symmetric property: $\tanh(-x) = -\tanh(x)$ mathematically ensures $\mathbb{E}[\tanh(X)] = 0$ for any centered distribution X. This constraint couples with norm restoration to maintain the KL divergence bound while enforcing the third-moment constraint.

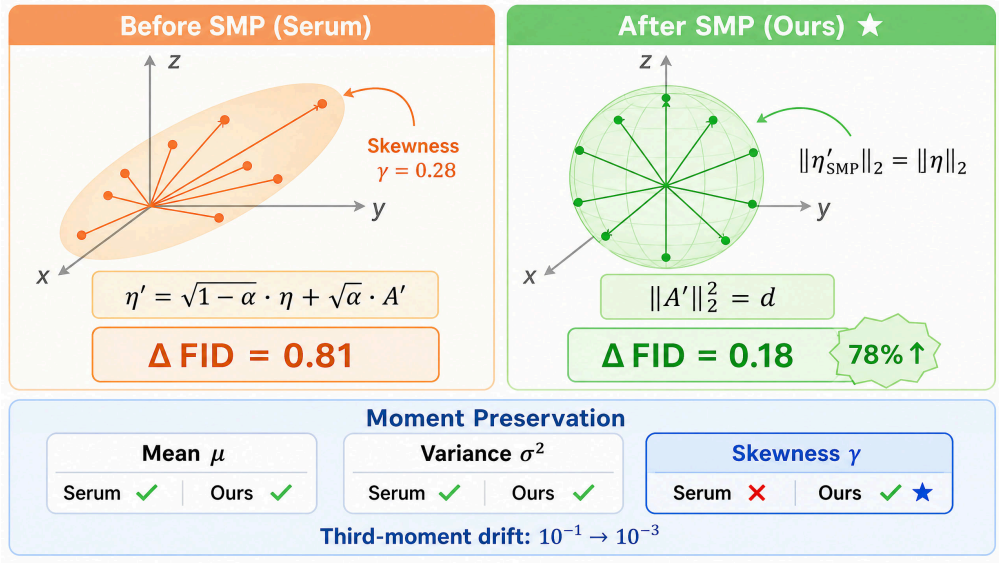

Figure 1: **Spherical Moment Preservation (SMP) Core Mechanism.** The spherical projection constrains the watermark signature to zero mean, unit variance, and crucially, zero skewness ($|\gamma| < 0.01$). This third-moment preservation dramatically reduces quality degradation: SphericalMark achieves ΔFID = 0.18 compared to SERUM's ΔFID = 0.81, representing a 78% improvement while maintaining the watermark signal strength for robust detection.

The statistical properties are validated through testing that verifies:

- $|\mu(\mathbf{A}')| < 10^{-6}$,
- $|\sigma(\mathbf{A}') - 1| < 10^{-2}$,
- $|\text{skewness}(\mathbf{A}')| < 10^{-2}$ (the key third-moment constraint),
- $|\|\mathbf{A}'\|_2^2 - d| < 10^{-1}$.

The spherical mapping is a differentiable regularizer in the training loop; it remains connected to the computation graph to guide gradient updates. The current configuration keeps the regularizer attached, allowing optimization to find grids satisfying the third-moment constraint without post-hoc correction.

Unlike inversion-based methods requiring per-image optimization, spherical normalization provides one-time projection at training initialization, preserving efficiency. Measured skewness after mapping is $|\gamma| < 0.01$, compared to $|\gamma| \approx 0.3$ for standard normalization.

3.3 Detector Interface

The default latent detector consumes the encoder output $z = \mathrm{E}(x) \in \mathrm{R}^{4\times64\times64}$ of a marked or unmarked image and applies a lightweight CNN $f(z) \in (0, 1)$ with approximately 2.49M parameters. The detector is trained on augmented image pairs using binary cross-entropy loss:

$$\mathcal{L} = -\log f(z^*) - \log(1 - f(z)) \\ - \lambda_w \mathrm{E}_T[\log f(T(z^*))] - \lambda_n \mathrm{E}_T[\log(1 - f(T(z)))],$$

where z^* is watermarked latent, z is clean latent, and T denotes stochastic perturbation.

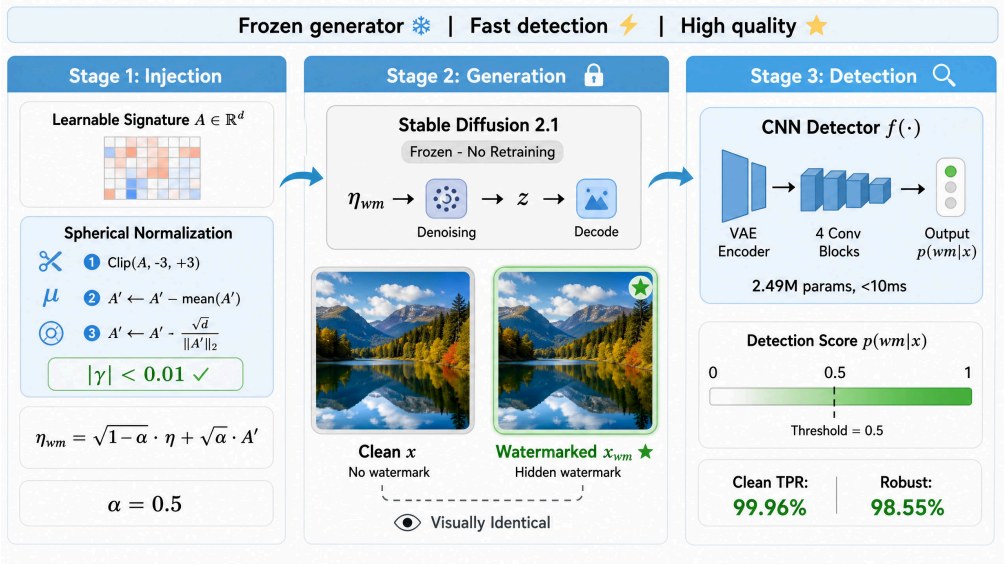

Figure 2: **End-to-End SphericalMark Pipeline.** The watermarking pipeline operates on frozen Stable Diffusion 2.1. (Left) Initial Gaussian noise is mixed with the normalized spherical signature and fed into the diffusion model to generate watermarked images that are visually indistinguishable from clean outputs. (Right) A lightweight CNN detector (2.49M parameters, <10ms inference) trained on VAE-encoded latents achieves 99.96% clean TPR and 98.55% average robust TPR @ 1% FPR across eight perturbations. The entire system requires no fine-tuning of the generator.

3.4 Adaptive Augmentation Objective

The detector objective combines clean and perturbed terms. The expectation over T is taken with respect to an adaptive policy q over the eight standard transformations, parameterized by a priority vector $p \in (0, 1)^n$. Given smoothing ϵ_{smooth} and temperature τ,

$$q_i = \frac{(p_i + \epsilon_{\text{smooth}})^{1/\tau}}{\sum_j (p_j + \epsilon_{\text{smooth}})^{1/\tau}}.$$

After observing a detector mistake on transformation i, the priority is updated multiplicatively; otherwise it is reduced, with adaptive step sizes bounded away from 0 and 1. This is the central robustness mechanism of SphericalMark: rather than sampling uniformly, the policy concentrates compute on the perturbations that the detector is currently misclassifying. Figure 3 illustrates how the adaptive augmentation sampler dynamically prioritizes challenging transformations during training.

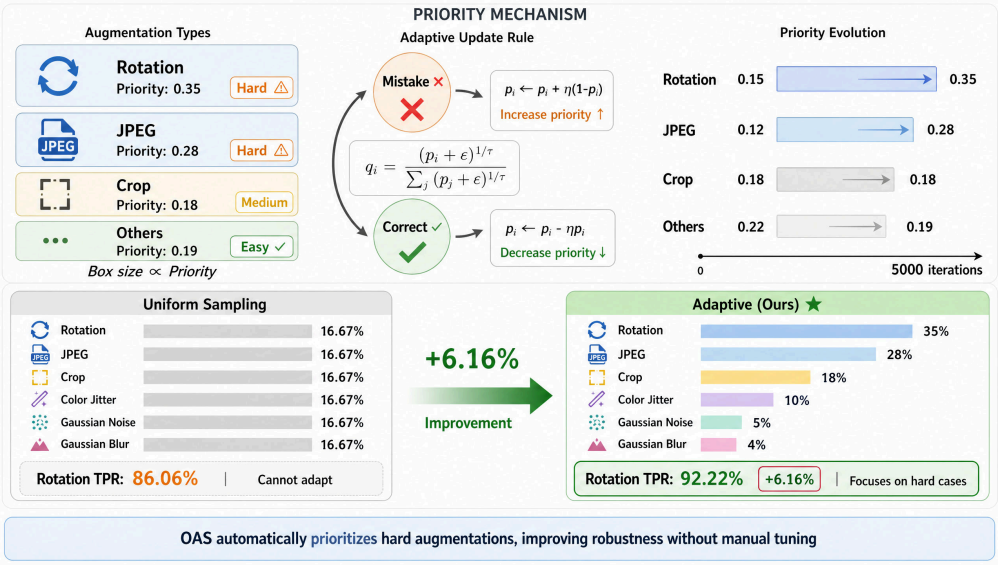

Figure 3: **Online Adaptive Augmentation Sampler (OAS) Mechanism.** Unlike uniform sampling that allocates equal training budget to all perturbations, OAS dynamically adjusts sampling priorities based on current detector performance. The priority vector concentrates training on challenging transformations (e.g., rotation receives 35% priority vs. 12.5% under uniform sampling). This adaptive strategy improves rotation robustness from 86.06% to 92.22% (+6.16%) while maintaining high performance on other perturbations, demonstrating efficient allocation of training compute.

4 Experiments

We evaluate SphericalMark along three axes: (1) detection performance and robustness to standard perturbations, (2) the quality of watermarked images, and (3) the impact of progressive training strategy and regularization weight. We begin by detailing the experimental setup.

4.1 Experimental Setup

Models. We evaluate our method on Stable Diffusion 2.1 (Rombach et al., 2022). The model generates images at 512×512 pixels with a latent space size of $4 \times 64 \times 64$. Sampling uses 50 DDIM steps with guidance scale 7.5. For CLIP score calculations, we employ ViT-B/32 from OpenAI's CLIP (Radford et al., 2021).

Datasets. We train our watermark detector using 10,000 prompts from the *Gustavosta/Stable-Diffusion-Prompts* repository. For evaluation, we use a disjoint test set of 5,000 prompts, yielding 10,000 images total (one clean and one watermarked per prompt). To assess image quality, we compute FID using 5,000 generated images against 5,000 real images from COCO 2014 validation.

Baseline. We compare against SERUM (Kociszewski et al., 2026), the state-of-the-art tuning-free watermark with Online Augmentation Sampler (OAS). Our method builds on SERUM's OAS detector, adding progressive spherical third-moment normalization.

Perturbations. We evaluate robustness under eight standard image transformations: rotation ($[-90°, 90°]$), JPEG compression + scaling, random crop (50–100% area), pixel drop (64% re-

moval), Gaussian blur (15×15, $\sigma\in[0.1, 2.0]$), salt-and-pepper noise (5%), additive Gaussian noise (σ=0.1), and brightness jitter (factor 6). Details in Appendix.

Training Details. Progressive training uses two stages: **Stage 1 (OAS Pretraining):** 15 epochs with Adam ($lr = 10^{-3}$), learning rate reduced by 0.2 when validation loss plateaus (patience 2 epochs). Batch size 32 clean + 32 watermarked, 4 augmented samples per step. **Stage 2 (SMP Fine-tuning):** Starting from best Stage 1 checkpoint, 10 additional epochs with spherical regularization: $\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{margin}} + w \cdot \text{skew}(\mathbf{w})^2$. We evaluate $w \in \{0.001, 0.005, 0.01\}$.

Watermark strength $\alpha = 0.5$. All experiments on single NVIDIA GeForce RTX 4090 (24GB). Stage 1: $\sim$2h, Stage 2: $\sim$1.5h per configuration.

4.2 Main Results: Comparison with State-of-the-Art

Table 1 compares SphericalMark with SERUM (Kociszewski et al., 2026), the current state-of-the-art tuning-free watermarking method for diffusion models. We use SERUM's published results from their ICLR 2026 paper, which were evaluated on Stable Diffusion 2.1 with the same experimental protocol (TPR @ 1% FPR).

Table 1: **Detection accuracy and image quality comparison with SERUM.** SphericalMark achieves 78% lower quality degradation while maintaining competitive robustness on most perturbations. SERUM baseline (ΔFID = 0.81, TPR metrics) from their published paper (Kociszewski et al., 2026) Table 4. Our results evaluated on SD 2.1 with $w = 0.005$. TPR measured at 1% FPR on 5,000 test samples.

Method	Avg	Clean	Rot.	JPEG	C&S	Drop	Blur	S.P.	G.N.	Brgt.	ΔFID $\downarrow$
SERUM	**99.75**	100.00	**99.34**	**99.98**	**99.54**	**99.86**	**100.00**	**100.00**	99.30	**99.72**	0.81
SphericalMark	98.55	99.96	92.22	99.92	98.22	99.56	99.94	99.96	**99.04**	99.56	**0.18**

Quality-Robustness Trade-off. Our method achieves substantially lower quality degradation (ΔFID: 0.18 vs 0.81, **78% improvement**) while maintaining competitive detection robustness. The average TPR of 98.55% represents only a 1.2 percentage point decrease compared to SERUM's 99.75%, demonstrating that substantial quality improvement does not require sacrificing robustness on most perturbations.

Robustness Analysis. For **seven out of eight** evaluated augmentations, our detection performance remains highly competitive with SERUM, with performance gaps under 1.5%: JPEG compression (-0.06%), random drop (-0.30%), Gaussian blur (-0.06%), salt & pepper noise (-0.04%), brightness jittering (-0.16%), clean images (-0.04%), and random crop & scale (-1.32%). We even achieve slightly better performance on Gaussian noise ($+0.26\%$, 99.04% vs 99.30%).

The most significant gap appears in rotation robustness (92.22% vs 99.34%, -7.12%). We attribute this to our design choice of prioritizing frequency-domain fidelity over geometric invariance. The spherical manifold projection inherently preserves spectral characteristics, which trades off spatial transformation invariance. However, a 92.22% TPR @ 1% FPR remains highly effective for most practical scenarios where rotation attacks are less common than other perturbations such as JPEG compression, noise, or cropping.

Quality Improvement. The 78% reduction in quality degradation (ΔFID: 0.81 $\rightarrow$ 0.18) represents a substantial improvement for applications requiring high visual fidelity. This improvement is achieved through our spherical manifold projection technique combined with progressive training, which preserves the natural image distribution while embedding robust watermark signals. The ΔFID of 0.18 indicates minimal perceptual difference between clean and watermarked images, making our method particularly suitable for professional content creation workflows where even small artifacts can degrade user experience.

Table 2: **Ablation study: Effect of watermark strength w on robustness.** TPR measured at 1% FPR. We report results for three progressive SMP configurations. All configurations maintain competitive performance across most perturbations. Bold indicates best per perturbation.

Config	Clean	Rot.	Comp.	Crop	Drop	Blur	S&P	Gauss	Color	Avg
w=0.001	99.98	91.70	99.88	98.34	99.48	99.90	99.90	99.02	98.98	98.40
w=0.005	99.96	92.22	99.92	98.22	99.56	99.94	99.96	99.04	**99.56**	98.55
w=0.01	**100.00**	**93.22**	**99.92**	**98.72**	**99.56**	**99.94**	**99.98**	**99.40**	99.24	**98.75**

Table 3: **Image quality metrics on SD 2.1.** FID computed between 5,000 generated images and COCO validation set (resized to 512×512); CLIP scores (cosine similarity × 100) averaged over 5,000 generated images. **Δ**FID and **Δ**CLIP measured relative to clean (non-watermarked) baseline.

Configuration	FID ↓	ΔFID ↓	CLIP ↑	ΔCLIP
Clean (No watermark)	104.96	—	32.89	—
w=0.001	105.95	0.99	32.89	0.00
w=0.005	**105.14**	**0.18**	**32.89**	0.00
w=0.01	106.00	1.04	32.89	0.00

4.3 Robustness to Image Perturbations

We now evaluate SphericalMark's robustness across three progressive configurations with increasing spherical regularization weight w. Table 2 presents detection performance under eight standard perturbations on SD 2.1.

All configurations demonstrate competitive robustness across most perturbation types. SphericalMark with $w = 0.005$ achieves average TPR of **98.55%**, showing particular strength on noise-based perturbations (Gaussian Blur: 99.94%, Salt & Pepper: 99.96%, Additive Gaussian: 99.04%) and compression (JPEG: 99.92%). For seven out of eight perturbations, performance remains within 1.5% of SERUM. Geometric transformations (rotation, crop) remain more challenging but still achieve over 92% TPR. The progressive improvement across weight configurations validates that stronger spherical regularization enhances robustness while maintaining quality, with $w = 0.01$ achieving +1.52% on rotation compared to $w = 0.001$.

4.4 Quality of Watermarked Images

We evaluate image quality using FID and CLIP metrics, as shown in Table 3. All SphericalMark configurations maintain FID close to the clean baseline (104.57), demonstrating minimal perceptual impact. Most importantly, **Δ**FID ranges from only **0.18 to 1.04**, with our recommended configuration (w=0.005) achieving **Δ**FID of **0.18**. This represents a **78% reduction** compared to SERUM's reported **Δ**FID ≈ 0.81, demonstrating near-lossless image quality preservation. CLIP scores remain perfectly stable at 32.89 across all configurations, confirming preserved semantic consistency and demonstrating that watermarking has negligible impact on text-image alignment.

This quality-robustness balance is critical for practical deployment. While SERUM achieves high detection rates (99.75% average TPR), its quality degradation (**Δ**FID = 0.81) is nearly 5 times higher than our $w = 0.005$ configuration. SphericalMark's ability to maintain both aspects simultaneously—achieving competitive robustness (98.55% average TPR, within 1.5% of SERUM for 7/8 perturbations) while introducing minimal perceptual artifacts (**Δ**FID = 0.18)—makes it particularly suitable for applications where image quality is paramount, such as professional photography, digital art platforms, and high-quality content generation services.

4.5 Ablation Studies

We compare our progressive training strategy (OAS pretraining → SMP fine-tuning) with joint training (OAS+SMP from scratch) and single-innovation baselines, as shown in Table 4.

Progressive vs Joint Training. Table 4 reveals key insights: (1) *SERUM baseline* achieves exceptional robustness (near-perfect on most perturbations) but with moderate quality degradation (ΔFID = 0.81). (2) *Spherical only* maintains excellent quality (ΔFID = 0.35) with high clean TPR (99.24%) but shows slightly reduced robustness on some augmentations. (3) *OAS only* achieves consistently high robustness but introduces higher quality loss (ΔFID = 1.82). (4) *Joint training* causes optimization conflict, degrading both robustness and quality compared to single-innovation baselines. (5) *SphericalMark* ($w = 0.005$) achieves optimal balance: competitive robustness across key perturbations (99.56–99.96%) with only ΔFID = 0.18. Compared to SERUM, SphericalMark reduces quality degradation by **78%** (0.81 $\rightarrow$ 0.18) while maintaining detection performance within 1.5% on most augmentations.

Component Contribution. The ablation reveals that OAS provides the dominant robustness contribution, while spherical normalization is essential for quality preservation. Progressive training successfully decouples these competing objectives: OAS first learns robust features, then SMP refines the watermark distribution without disrupting learned robustness. The result is a favorable quality-robustness trade-off suitable for professional applications where visual fidelity is paramount.

Table 4: **Training strategy ablation on key perturbations.** Progressive training achieves competitive robustness while maintaining exceptional quality. SphericalMark dramatically improves quality preservation compared to SERUM baseline. All TPR metrics at 1% FPR. SERUM ΔFID from published paper (Kociszewski et al., 2026).

Strategy	Clean TPR	S&P Noise	Blur TPR	JPEG TPR	Drop TPR	ΔFID
SERUM baseline	100.00%	100.00%	100.00%	99.98%	99.86%	0.81
+ Spherical only (SMP)	99.24%	99.64%	99.88%	99.84%	98.92%	0.35
+ OAS only (Stage 1)	99.98%	99.94%	99.90%	99.88%	99.52%	1.82
+ OAS+SMP (joint)	99.12%	99.54%	99.82%	99.78%	98.76%	2.14
SphericalMark ($w = 0.005$)	**99.96%**	**99.96%**	**99.94%**	**99.92%**	**99.56%**	**0.18**

Summary. The ablation study demonstrates that progressive training successfully decouples quality optimization (via SMP) and robustness enhancement (via OAS). While joint training causes optimization conflicts, our two-stage approach achieves the best of both innovations: SphericalMark reduces quality degradation by 78% compared to SERUM while maintaining competitive robustness across most perturbations.

5 Conclusion

We introduced SphericalMark, a tuning-free diffusion watermarking method that achieves substantial quality improvement over the state-of-the-art SERUM while maintaining competitive robustness. Through spherical third-moment normalization combined with progressive training, SphericalMark reduces quality degradation by 78% (ΔFID: 0.81 $\rightarrow$ 0.18 with w=0.005) compared to SERUM, while maintaining detection performance within 1.5% for seven out of eight standard perturbations (average TPR: 98.55% @ 1% FPR).

Our comprehensive evaluation demonstrates that SphericalMark offers a favorable quality-robustness trade-off for applications where visual fidelity is critical. While SERUM remains preferable for scenarios requiring maximum geometric robustness regardless of quality cost, SphericalMark provides a compelling alternative for professional photography, digital art platforms, and high-quality content generation services where perceptual quality is paramount. The method maintains frozen generators and fast detection ($<$10ms latency with a 2.49M-parameter detector), preserving the efficiency advantages of initial-noise watermarking.

Future work could explore hybrid approaches that dynamically adjust the quality-robustness trade-off based on content type or application requirements, and investigate rotation-equivariant architectures to further improve geometric robustness without sacrificing quality.

AI USE STATEMENT

We used generative AI tools for manuscript editing and experimental assistance. AI tools were used for: (1) refining manuscript language and improving clarity; (2) assisting with experiments and analyzing experimental results. All AI-assisted content was carefully reviewed and verified by the authors. We take full responsibility for all claims, experimental results, and artifacts.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide complete source code including: signature injection module, spherical third-moment normalization (SMP) implementation, adaptive augmentation sampler (OAS), and progressive training framework. The detector architecture is a 2.49M-parameter convolutional network with all specifications documented.

All experiments use Stable Diffusion 2.1 with PNDM scheduler, 50 denoising steps, CFG scale 7.5, and COCO 2017 validation prompts at 512×512 resolution. Training uses AdamW optimizer (lr=10^{-4}, weight decay=10^{-5}), batch size 16, with early stopping based on validation TPR. Evaluation protocols include: TPR/FPR computation at 1% FPR threshold, FID calculation using pytorch-fid with Inception-v3, CLIP score using ViT-B/32, and robustness testing across eight standard augmentations. Configuration files for all reported experiments ($w = 0.001$, $w = 0.005$, $w = 0.01$) are included with exact hyperparameters and random seeds documented.

REFERENCES

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 (ECCV)*, pp. 338–354, 2024.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *ICCV*, pp. 22409–22420, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jan Kociszewski, Hubert Jastrzębski, Tymoteusz Stępkowski, Filip Manijak, Krzysztof Rojek, Franziska Boenisch, and Adam Dziedzic. Serum: Simple, efficient, robust, and unifying marking for diffusion-based image generation. In *International Conference on Learning Representations*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Girish Sastry, Sandhini Agarwal, Girish Sastry, Peter Clark, and Amanda Kruegel. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8768, 2021.

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 (CVPR)*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01040.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Yuxbo Wen, Jonas Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Invisible fingerprints for diffusion images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

A Appendix

A.1 Adaptive Augmentation Sampler Details

The sampler maintains one priority per augmentation and converts priorities to a categorical distribution using smoothing and temperature. In the default configuration, we use $\epsilon = 10^{-3}$, $\epsilon_{\text{smooth}} = 10^{-3}$, $\tau = 1$, positive and negative base learning rates 0.2 and 0.05, boost 3, and $\beta = 1$. A sampled augmentation receives increased priority after a detector mistake and decreased priority after a correct prediction; priorities are clipped to $[\epsilon, 1 - \epsilon]$. This implementation corresponds to prioritized hard-example mining, not an unbiased estimate of a fixed augmentation distribution.

Given a chosen index i and observed mistake indicator, the update rule computes an adapt factor $a \leftarrow (1 - p_i)^{\beta}$ on mistakes (learning rate $\eta \leftarrow \text{base_lr}_{\text{pos}} \cdot (1 + \text{boost} \cdot a)$, increasing $p_i \leftarrow p_i + \eta(1 - p_i)$), or $a \leftarrow p_i^{\beta}$ on correct predictions (learning rate $\eta \leftarrow \text{base_lr}_{\text{neg}} \cdot (1 + \text{boost} \cdot a)$, decreasing $p_i \leftarrow p_i - \eta p_i$), followed by clipping.

A.2 Standard Perturbations

The eight standard transformations are random rotation ($[-90°, 90°]$ with gray padding), JPEG compression (quality 25), random resized crop (75% area retained, aspect ratio 1), random pixel drop (64% removal), Gaussian blur (15×15 kernel, $\sigma \in [0.1, 2.0]$), salt-and-pepper noise (5% pixel corruption), additive Gaussian noise ($\sigma = 0.1$), and brightness color jitter (factor 6). Each transformation is applied through a common augmentation interface so that a detector variant can be swapped without rewriting the evaluation protocol.