Novelty Assessment Report

Paper: Spherical Watermark: Encryption-Free, Lossless Watermarking for Diffusion Models
PDF URL: https://openreview.net/pdf?id=2eAGrunxVz
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Diffusion models have revolutionized image synthesis but raise concerns around content provenance and authenticity. Digital watermarking offers a means of tracing generated media, yet traditional schemes often introduce distributional shifts and degrade visual quality. Recent lossless methods embed watermark bits directly into the latent Gaussian prior without modifying model weights, but still require per-image key storage or heavy cryptographic overhead. In this paper, we introduce Spherical Watermark, an encryption-free and lossless watermarking framework that integrates seamlessly with diffusion architectures. First, our binary embedding module mixes repeated watermark bits with random padding to form a high-entropy code. Second, the spherical mapping module projects this code onto the unit sphere, applies an orthogonal rotation, and scales by a chi-square-distributed radius to recover exact multivariate Gaussian noise. We theoretically prove that the watermarked noise distribution preserves the target prior up to third-order moments, and empirically demonstrate that it is statistically indistinguishable from a standard multivariate normal distribution. Adopting Stable Diffusion, extensive experiments confirm that Spherical Watermark consistently preserves high visual fidelity while simultaneously improving traceability, computational efficiency, and robustness under attacks, thereby outperforming both lossy and lossless approaches.

Core Task Landscape

This paper addresses: **Lossless Watermarking for Diffusion Models**
A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:
- **Watermark Embedding Mechanism**
- **Robustness and Security**
- **Domain-Specific Extensions**
- **Application-Oriented Frameworks**

Complete Taxonomy Tree

- Lossless Watermarking for Diffusion Models Survey Taxonomy
- Watermark Embedding Mechanism
  - Latent Noise Manipulation
  - Lossless Noise Distribution Preservation ★ (2 papers)
    - [0] Spherical Watermark: Encryption-Free, Lossless Watermarking for Diffusion Models (Anon et al., 2026) View paper
    - [1] Gaussian shading: Provable performance-lossless image watermarking for diffusion models (Zijin Yang, 2024) View paper
  - Noise Pattern Injection (3 papers)
    - [4] Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust (Wen, 2023) View paper
    - [18] Shallow Diffuse: Robust and Invisible Watermarking through Low-Dimensional Subspaces in Diffusion Models (Li Wenda, 2024) View paper
    - [39] T2SMark: Balancing Robustness and Diversity in Noise-as-Watermark for Diffusion Models (Yang JinDong, 2025) View paper
  - Latent Code Seeding and Mapping (3 papers)
    - [20] Robust watermarking for diffusion model generated images (Ziqi Liu, 2025) View paper
    - [38] DGADM-GIS: Deterministic Guided Additive Diffusion Model for Generative Image Steganography (Chengsheng Yuan, 2025) View paper
    - [42] SDS-TG: Secure Diffusion Steganography in Text-Guided Generative Images (Haozhong Yang, 2025) View paper
  - Model-Integrated Watermarking
  - Decoder Fine-Tuning (3 papers)
    - [44] TraceMark-LDM: Authenticatable Watermarking for Latent Diffusion Models via Binary-Guided Rearrangement (Luo, 2025) View paper
    - [45] The Stable Signature: Rooting Watermarks in Latent Diffusion Models (Pierre Fernandez, 2023) View paper
    - [48] SWA-LDM: Toward Stealthy Watermarks for Latent Diffusion Models (Yang, 2025) View paper
  - Full Model Training with Watermark Task (3 papers)
    - [2] Image watermarking of generative diffusion models (Vice Jordan, 2025) View paper
    - [13] SleeperMark: Towards Robust Watermark against Fine-Tuning Text-to-image Diffusion Models (Zilan Wang, 2024) View paper
    - [19] Embedding Watermarks in Diffusion Process for Model Intellectual Property Protection (Yang, 2024) View paper
  - Sampling Process Guidance (4 papers)
    - [23] Dynamic watermarks in images generated by diffusion models (Yunzhuo Chen, 2025) View paper

Narrative

Core task: lossless watermarking for diffusion models. The field addresses how to embed imperceptible ownership or provenance signals into images, videos, and other media generated by diffusion processes without degrading output quality. The taxonomy organizes research into four main branches. Watermark Embedding Mechanism explores how watermarks are injected—whether by manipulating latent noise distributions (e.g., Gaussian Shading[1], Tree Ring Watermarks[4]), conditioning the diffusion process itself (e.g., Watermark Conditioned Diffusion[16]), or modifying model weights and training procedures (e.g., Stable Signature[45]). Robustness and Security examines defenses against removal attacks and adversarial perturbations, with works like Attack Resilient Watermarking[3] and Groot[5] focusing on maintaining detectability under distortions. Domain-Specific Extensions adapt watermarking to specialized modalities such as video (Videomark[6], VideoShield[32]), 3D meshes (Robust Reversible 3D[43]), audio (Invisible Audio Watermarking[34]), and text-to-shape generation (T2SMark[39]). Application-Oriented Frameworks target practical deployment scenarios including federated learning (Federated Diffusion Watermarking[7]), black-box settings (Black Box Watermarking[9]), and copyright protection (DiffusionShield Copyright[22]).

A particularly active line of work centers on latent noise manipulation, where methods preserve the statistical properties of the noise distribution to achieve losslessness while embedding detectable patterns. Spherical Watermark[0] sits within this cluster, emphasizing lossless noise distribution preservation through spherical coordinate transformations. This contrasts with approaches like Gaussian Shading[1], which also manipulates noise but may prioritize robustness trade-offs differently, and with training-based methods such as Stable Signature[45] that bake watermarks into model parameters rather than per-sample noise. Meanwhile, robustness-focused branches explore adversarial resilience (Attack Resilient Watermarking[3]) and adaptive attacks (Groot[5]), highlighting tensions between imperceptibility, capacity, and security. Open questions persist around balancing these competing objectives, scaling to diverse modalities, and ensuring watermark persistence across post-processing pipelines without sacrificing the core promise of lossless generation quality.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Gaussian shading: Provable performance-lossless image watermarking for diffusion models

Authors: Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, et al. (6 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â⃞ achieve a high-capacity watermark of 256 bits, surpassing â⃞ of performancelossless watermarking for diffusion models, and â⃞ Reversible image watermarking based on integer-to-integer â⃞

Relationship Analysis

Both papers belong to the Lossless Noise Distribution Preservation category, focusing on embedding watermarks while maintaining the original Gaussian noise distribution in diffusion models. They share the goal of provably lossless watermarking by ensuring watermarked noise is statistically indistinguishable from standard Gaussian noise. The key difference is that the original paper (Spherical Watermark) uses a spherical mapping approach with orthogonal rotation and chi-square scaling, while the candidate paper (Gaussian Shading) employs stream cipher encryption and distribution-preserving sampling via quantile functions to achieve the same lossless property.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Spherical Watermark framework for lossless, encryption-free watermarking

Description: The authors propose a novel watermarking framework that embeds watermarks into diffusion models without requiring encryption or per-image key storage. The framework uses binary embedding and spherical mapping modules to transform watermark bits into Gaussian noise that is statistically indistinguishable from the original prior, enabling lossless watermarking with improved traceability and computational efficiency.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Videomark: A distortion-free robust watermarking framework for video diffusion models

URL: View paper

Brief Assessment

Videomark[6] focuses on video diffusion models with frame-wise watermarking and temporal matching, while the original paper addresses image diffusion models with spherical mapping and binary embedding modules. The technical approaches differ fundamentally in their application domains and architectural designs.

2. Shallow Diffuse: Robust and Invisible Watermarking through Low-Dimensional Subspaces in Diffusion Models

URL: View paper

Brief Assessment

Shallow Diffuse[18] focuses on decoupling watermarking from diffusion sampling using low-dimensional subspaces and null space embedding. The ORIGINAL paper's Spherical Watermark uses binary embedding and spherical mapping modules with orthogonal rotations and chi-square scaling to achieve lossless watermarking. These are fundamentally different technical approaches to watermarking diffusion models.

3. DiffusionShield: A Watermark for Data Copyright Protection against Generative Diffusion Models

URL: View paper

Brief Assessment

DiffusionShield Data Copyright[55] focuses on copyright protection by embedding watermarks that are learned and reproduced by generative models to detect infringement. The original paper proposes a lossless, encryption-free framework using spherical mapping to preserve the Gaussian prior distribution. These represent fundamentally different technical approaches and objectives.

4. Black-box forgery attacks on semantic watermarks for diffusion models

URL: View paper

Brief Assessment

Black Box Forgery[52] focuses on attacking existing semantic watermarks (Tree-Rings, Gaussian Shading) rather than proposing a new watermarking framework. The candidate demonstrates forgery attacks on watermarks but does not present an alternative lossless, encryption-free watermarking method that would challenge the novelty of the Spherical Watermark framework.

5. Intellectual Property Protection of Diffusion Models via the Watermark Diffusion Process

URL: View paper

Brief Assessment

IP Protection Watermark[54] focuses on training a watermark diffusion process (WDP) alongside the standard diffusion process, without imprinting watermarks during task generation. In contrast, the original paper embeds watermarks directly into the noise input via spherical mapping without modifying model weights or training processes.

6. Towards Effective User Attribution for Latent Diffusion Models via Watermark-Informed Blending

URL: View paper

Brief Assessment

Watermark Informed Blending[15] focuses on user attribution through ready-to-use configuration and pixel-level operations, not on lossless, encryption-free frameworks using spherical mapping and binary embedding modules as described in the original paper.

7. Gaussian shading: Provable performance-lossless image watermarking for diffusion models
URL: View paper

Prior Art Analysis

Gaussian Shading[1] demonstrates that a lossless watermarking framework for diffusion models without encryption was already proposed and implemented. The candidate paper presents a method that embeds watermarks into diffusion models by mapping watermark bits to Gaussian noise that is statistically indistinguishable from the original prior, achieving performance-lossless watermarking. This directly refutes the novelty claim of the original paper's Spherical Watermark framework, as Gaussian Shading[1] already established the core concepts of lossless watermarking without model modification, distribution-preserving sampling, and theoretical guarantees of indistinguishability.

Evidence

Evidence 1 - **Rationale**: Both papers claim to be the first to propose lossless watermarking frameworks that map watermarks to Gaussian noise indistinguishable from the original distribution. Gaussian Shading[1] explicitly states this achievement, demonstrating prior work exists. - **Original**: we introduce spherical watermark, an encryption-free and lossless watermarking framework that integrates seamlessly with diffusion architectures. first, our binary embedding module mixes repeated watermark bits with random padding to form a highentropy code. second, the spherical mapping module proj... - **Candidate**: we propose gaussian shading, a diffusion model watermarking technique that is both performance-lossless and trainingfree, while serving the dual purpose of copyright protection and tracing of offending content. our watermark embedding is free of model parameter modifications and thus is plug-and-pla...

Evidence 2 - **Rationale**: Both papers describe multi-module frameworks for lossless watermarking. Gaussian Shading[1] presents watermark diffusion, randomization, and distribution-preserving sampling modules, which are conceptually similar to the original paper's binary embedding and spherical mapping modules. - **Original**: in this paper, we introduce spherical watermark, a simple yet effective lossless scheme that is encryption-free and robust against common attacks. our method integrates seamlessly with pretrained diffusion models via three modules: binary embedding, spherical mapping, and diffusion integration. - **Candidate**: building upon this insight, we propose a watermarking method named gaussian shading, designed to ensure no deterioration in model performance. the embedding process encompasses three primary elements: watermark diffuse, randomization, and distribution-preserving sampling. watermark diffusion spreads...

Evidence 3 - **Rationale**: While the original paper claims encryption-free design as novel, Gaussian Shading[1] uses stream cipher encryption (ChaCha20) but still achieves lossless watermarking. This shows that the encryption-free aspect is not the fundamental novelty, as lossless watermarking was already achieved with encryption in prior work. - **Original**: compared to existing lossless watermarking schemes, our encryption-free approach omits key storage overhead, enabling an excellent trade-off between undetectability and watermark robustness. - **Candidate**: watermark randomization. if we know the distribution of the diffused watermark sd, we can directly utilize distribution-preserving sampling to obtain the corresponding latent representations zs t . however, in practical scenarios, its distribution is always unknown. hence, we introduce a stream key ...

8. Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models
URL: View paper

Brief Assessment

Watermark Embedded Adversarial[53] focuses on embedding visible watermarks into adversarial examples to prevent diffusion models from imitating copyrighted images, not on lossless watermarking of the diffusion model's latent prior itself. The technical approaches are fundamentally different: the candidate uses adversarial perturbations with GANs for copyright protection, while the original proposes spherical mapping to preserve Gaussian noise distribution.

9. Embedding Watermarks in Diffusion Process for Model Intellectual Property Protection
URL: View paper

Brief Assessment

Embedding Watermarks Diffusion[19] embeds watermarks into the intermediate diffusion process during training, not into the Gaussian prior. The candidate modifies the forward diffusion sequence and training objectives, whereas the original transforms watermark bits into statistically indistinguishable Gaussian noise inputs without model modification.

Contribution 2: Spherical mapping module with theoretical guarantees
Description: The authors develop a spherical mapping technique that projects binary codes onto the unit sphere, applies orthogonal rotation, and scales by a chi-square-distributed radius. They theoretically prove that the resulting watermarked noise preserves the target Gaussian prior up to third-order moments and forms a spherical 3-design, ensuring statistical indistinguishability from standard Gaussian noise.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A Comprehensive Review of Three-Dimensional Watermarking Algorithms
URL: View paper

Brief Assessment

3D Watermarking Review[51] focuses on 3D object watermarking (meshes, Gaussian splats) with geometric transformations like scaling into unit spheres/cubes. The ORIGINAL paper addresses 2D image generation via diffusion models, using spherical mapping to transform binary codes into Gaussian noise that preserves statistical properties for latent-space watermarking. These are fundamentally different application domains and technical approaches.

Contribution 3: Encryption-free design eliminating key storage overhead
Description: The authors design a watermarking system that does not require per-image cryptographic keys or nonces, unlike prior lossless methods. This eliminates substantial storage and management overhead while maintaining strong traceability and robustness against attacks, offering improved computational efficiency compared to cryptography-based approaches.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Keyless, Robust and Cost-Effective Privacy Protection Algorithm for Videos Using Steganography
URL: View paper

Brief Assessment

Keyless Privacy Protection[61] focuses on video steganography for piracy protection using randomized pixel-based watermarks, not on diffusion model watermarking or latent-space Gaussian noise embedding. The technical domains and methodologies differ fundamentally.

2. Enhanced least significant qubit watermarking scheme for quantum images: G. Luo et al.
URL: View paper

Brief Assessment

The candidate paper (Least Significant Qubit[58]) focuses on quantum image watermarking using LSB techniques, which is fundamentally different from the original paper's diffusion model watermarking framework. The candidate's 'keyless' approach applies to quantum images, not to diffusion-based generative models.

3. Watermarking and authentication of quantum images based on restricted geometric transformations
URL: View paper

Brief Assessment

Quantum Image Authentication[56] focuses on quantum image watermarking using geometric transformations, not diffusion model watermarking. The technical domains and implementation contexts are fundamentally different.

4. Keyless Public Watermarking for Intellectual Property Authentication
URL: View paper

Brief Assessment

The candidate paper context is not available (marked as 'n/a'), making it impossible to perform a meaningful comparison or extract evidence regarding encryption-free watermarking methods.

5. Tri-layer stego for enhanced security-A keyless random approach
URL: View paper

Brief Assessment

Tri Layer Stego[59] focuses on multi-stage steganographic embedding using space-filling curves for image hiding, not on eliminating cryptographic key storage in watermarking systems for diffusion models.

6. NEURAL WATERMARKING: LACK OF A SECRET KEY IS STILL LACK OF SECURITY
URL: View paper

Brief Assessment

Neural Watermarking Security[62] focuses on attacking keyless watermarking schemes (specifically WAM) to demonstrate security vulnerabilities, rather than proposing an encryption-free design as a positive contribution. The paper critiques the absence of secret keys rather than offering it as a feature.

7. A Contrastive Semantic Watermarking Framework for Large Language Models
URL: View paper

Brief Assessment

Contrastive Semantic Watermarking[57] focuses on text watermarking for LLMs using semantic token selection and contrastive learning, not image watermarking for diffusion models. The candidate addresses keyless watermarking in NLP contexts, whereas the original contribution concerns encryption-free image watermarking that eliminates per-image cryptographic keys in visual generation systems.

Appendix: Text Similarity Detection

Textual similarity detection checked 18 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Gaussian shading: Provable performance-lossless image watermarking for diffusion models

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Spherical Watermark: Encryption-Free, Lossless Watermarking for Diffusion Models View paper
- [1] Gaussian shading: Provable performance-lossless image watermarking for diffusion models View paper
- [2] Image watermarking of generative diffusion models View paper
- [3] Attack-resilient image watermarking using stable diffusion View paper
- [4] Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust View paper
- [5] Groot: Generating robust watermark for diffusion-model-based audio synthesis View paper
- [6] Videomark: A distortion-free robust watermarking framework for video diffusion models View paper
- [7] Robust Watermarking for Federated Diffusion Models with Unlearning-Enhanced Redundancy View paper
- [8] WaterFlow: Learning Fast & Robust Watermarks using Stable Diffusion View paper
- [9] Performance-lossless black-box model watermarking View paper
- [10] Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model View paper
- [11] Robust watermarking for diffusion models based on STDM and latent space fine-tuning View paper
- [12] TAG-WM: Tamper-Aware Generative Image Watermarking via Diffusion Inversion Sensitivity View paper
- [13] SleeperMark: Towards Robust Watermark against Fine-Tuning Text-to-image Diffusion Models View paper
- [14] Robin: Robust and invisible watermarks for diffusion models with adversarial optimization View paper
- [15] Towards Effective User Attribution for Latent Diffusion Models via Watermark-Informed Blending View paper
- [16] A watermark-conditioned diffusion model for ip protection View paper
- [17] Plug-and-Hide: Provable and Adjustable Diffusion Generative Steganography View paper
- [18] Shallow Diffuse: Robust and Invisible Watermarking through Low-Dimensional Subspaces in Diffusion Models View paper
- [19] Embedding Watermarks in Diffusion Process for Model Intellectual Property Protection View paper
- [20] Robust watermarking for diffusion model generated images View paper

- [21] Unified High-binding Watermark for Unconditional Image Generation Models View paper
- [22] DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models View paper
- [23] Dynamic watermarks in images generated by diffusion models View paper
- [24] Safe-SD: Safe and Traceable Stable Diffusion with Text Prompt Trigger for Invisible Generative Watermarking View paper
- [25] WaterDiff: Perceptual Image Watermarks Via Diffusion Model View paper
- [26] Hey That's Mine Imperceptible Watermarks are Preserved in Diffusion Generated Outputs View paper
- [27] DERO: Diffusion-Model-Erasure Robust Watermarking View paper
- [28] Attention to Neural Plagiarism: Diffusion Models Can Plagiarize Your Copyrighted Images! View paper
- [29] Guidance Watermarking for Diffusion Models View paper
- [30] CLUE-MARK: Watermarking Diffusion Models using CLWE View paper
- [31] MarkPlugger: Generalizable Watermark Framework for Latent Diffusion Models without Retraining View paper
- [32] VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking View paper
- [33] HMARK: Radioactive Multi-Bit Semantic-Latent Watermarking for Diffusion Models View paper
- [34] Invisible Watermarking for Audio Generation Diffusion Models View paper
- [35] Watermarking via Gaussian noise modulation in diffusion models View paper
- [36] A chaotic hierarchical encryption/watermark embedding scheme for multi-medical images based on rowâ□column confusion and closed-loop bi-directional diffusion View paper
- [37] CoSDA: Enhancing the Robustness of Inversion-based Generative Image Watermarking Framework View paper
- [38] DGADM-GIS: Deterministic Guided Additive Diffusion Model for Generative Image Steganography View paper
- [39] T2SMark: Balancing Robustness and Diversity in Noise-as-Watermark for Diffusion Models View paper
- [40] Watermarking Discrete Diffusion Language Models View paper
- [41] BitMark: Watermarking Bitwise Autoregressive Image Generative Models View paper
- [42] SDS-TG: Secure Diffusion Steganography in Text-Guided Generative Images View paper
- [43] Robust Reversible Watermarking forÂ 3D Models Based onÂ Auto Diffusion Function View paper
- [44] TraceMark-LDM: Authenticatable Watermarking for Latent Diffusion Models via Binary-Guided Rearrangement View paper
- [45] The Stable Signature: Rooting Watermarks in Latent Diffusion Models View paper
- [46] PlugMark: A Plug-in Zero-Watermarking Framework for Diffusion Models View paper
- [47] DTR: Dynamic Tree-Ring Watermarking Framework for Diffusion-Based Video Generation View paper
- [48] SWA-LDM: Toward Stealthy Watermarks for Latent Diffusion Models View paper
- [49] SuperMark: Robust and Training-free Image Watermarking via Diffusion-based Super-Resolution View paper
- [50] TabWak: A Watermark for Tabular Diffusion Models View paper
- [51] A Comprehensive Review of Three-Dimensional Watermarking Algorithms View paper
- [52] Black-box forgery attacks on semantic watermarks for diffusion models View paper
- [53] Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models View paper
- [54] Intellectual Property Protection of Diffusion Models via the Watermark Diffusion Process View paper
- [55] DiffusionShield: A Watermark for Data Copyright Protection against Generative Diffusion Models View paper
- [56] Watermarking and authentication of quantum images based on restricted geometric transformations View paper
- [57] A Contrastive Semantic Watermarking Framework for Large Language Models View paper
- [58] Enhanced least significant qubit watermarking scheme for quantum images: G. Luo et al. View paper
- [59] Tri-layer stego for enhanced security-A keyless random approach View paper
- [60] Keyless Public Watermarking for Intellectual Property Authentication View paper
- [61] Keyless, Robust and Cost-Effective Privacy Protection Algorithm for Videos Using Steganography View paper
- [62] NEURAL WATERMARKING: LACK OF A SECRET KEY IS STILL LACK OF SECURITY View paper