

Robust Yet Imperceptible Diffusion Watermarking For AI-Generated Image Verification

Shuren Qi, Wenying Wen, Mengxin Shen, Yushu Zhang, and Fenglei Fan

Abstract—Diffusion models have enabled highly realistic image generation, but their misuse also raises growing concerns about misleading visual content. Digital watermarking has emerged as a promising proactive solution for verifying AI-generated images in cooperative deployment settings. An effective watermarking method for diffusion models should ideally satisfy two key properties: imperceptibility and robustness. However, existing methods still face a clear trade-off between them: pixel-level approaches usually preserve image fidelity well but are vulnerable to regeneration-based removal, whereas content-level approaches are often more robust but may compromise reproducibility, pixel-level consistency, or deployment efficiency due to retraining.

In this paper, we revisit diffusion watermarking from the perspective of embedding space design. We argue that a major difficulty lies in finding an embedding space that simultaneously captures components that are both visually insensitive and stable under common distortions. To address this issue, we propose a plug-and-play watermarking framework that embeds watermark information into the circular harmonic expansion of the initial noise in diffusion models. This embedding space provides two favorable properties for watermarking: amplitude-based invariance, which supports reliable detection under geometric transformations, and reconstructability with frequency selectivity, which helps preserve the original generation behavior while limiting perceptual distortion.

Extensive experiments demonstrate that the proposed method achieves a strong robustness-imperceptibility trade-off without retraining the diffusion model. In particular, it shows strong robustness against natural distortions, geometric transformations, and regeneration-based expert attacks, while substantially improving imperceptibility over representative content-level baselines. Moreover, the method remains lightweight and broadly compatible with existing diffusion-model pipelines, making it practical for active AI-generated image verification.

Index Terms—Diffusion model, Image generation, Digital watermarking, Imperceptibility, Robustness

This work was supported in part by the Young Talent Support Project (Hong Kong and Macau) of Guangzhou Association for Science and Technology under Grant QT-2025-047, the Startup Fund of the City University of Hong Kong, the National Natural Science Foundation of China under Grant 62522112, the Jiangxi Key Research and Development Program under Grant 20261BCE310050, and the Ganpo Talent Program of Jiangxi Province under Grant gpyc20240012. (Corresponding author: Wenying Wen)

Shuren Qi and Fenglei Fan are with the Department of Data Science, City University of Hong Kong, Hong Kong (e-mail: shurenqi@cityu.edu.hk; fenglfan@cityu.edu.hk).

Wenying Wen and Yushu Zhang are with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China, and also with the Jiangxi Provincial Key Laboratory of Multimedia Intelligent Processing, Nanchang, China (e-mail: wenyingwen@sina.cn; zhangyushu@jxufe.edu.cn).

Mengxin Shen is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: mengxin_shen@163.com).

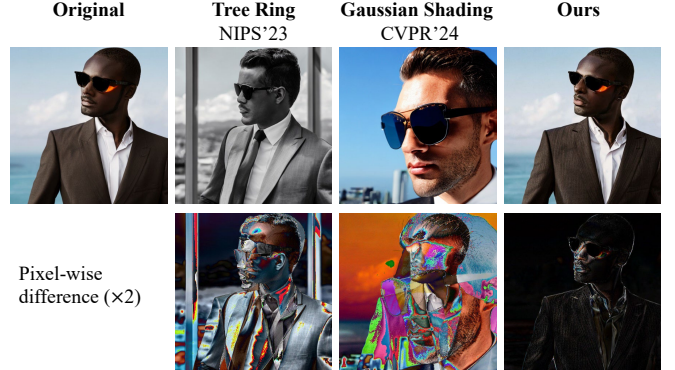


Fig. 1. Comparison between an image generated by the original diffusion model and the corresponding watermarked images produced by Tree Ring (NIPS'23), Gaussian Shading (CVPR'24), and the proposed method. Compared with the most relevant content-level baselines, our method preserves substantially stronger pixel-level consistency with the original output.

I. INTRODUCTION

DIFFUSION models [1]–[7] have become a dominant paradigm for image generation due to their ability to synthesize highly realistic and diverse visual content. Commercial systems such as Stable Diffusion [8], DALL-E 2 [9], and Imagen [10] have made high-quality image generation accessible to non-expert users. While this accessibility enables many beneficial applications, it also raises growing concerns about the large-scale creation and dissemination of misleading visual content. Notable examples include the fake images of Trump's arrest¹ and the Balenciaga Pope², both of which spread widely online. These developments have motivated increasing interest in practical mechanisms for identifying and labeling generated content.

Watermarking provides an active alternative by embedding identifiable information into the generated content as a prior for subsequent verification [11]. For diffusion models, an effective watermarking method should ideally satisfy two key properties: *imperceptibility* and *robustness*. Imperceptibility concerns the fidelity of a watermarked output to its corresponding unwatermarked output, while robustness requires the watermark to remain detectable after natural distortions or deliberate removal attempts. We operationalize imperceptibility as *pairwise output consistency*: the corresponding unwatermarked and watermarked outputs are produced with the same diffusion model, prompt, seed, sampler, and other generation

¹<https://www.bbc.com/news/world-us-canada-65069316>

²<https://www.forbes.com/sites/danidiplacido/2023/03/27/why-did-balenciaga-pope-go-viral/>

settings, and are evaluated as a matched pair. This protocol holds the generation conditions fixed when assessing changes associated with watermark embedding. Semantic consistency measures whether a generated image remains compatible with its prompt. Pairwise output consistency measures the change associated with watermark embedding when the model and generation settings are held fixed, thereby directly reflecting reproducibility and preservation of the provider’s original output in the API setting.

Existing diffusion watermarking methods often face a trade-off between robustness and strict imperceptibility, particularly in plug-and-play settings without retraining. Section II reviews this trade-off and its relation to the choice of embedding space.

Motivated by this observation, we propose a plug-and-play watermarking framework for diffusion models that achieves a strong trade-off between robustness and imperceptibility. Our key idea is to embed the watermark into the circular harmonic expansion of the initial noise. In this representation, the coefficients encode geometric information at different frequencies, making it possible to select components that are both relatively robust and less visually intrusive. Moreover, the embedding space offers two desirable properties for watermarking: reconstructability, which helps preserve the original generation behavior, and amplitude-based invariance, which improves detection stability under geometric transformations such as rotation and flipping. As a result, the proposed method remains lightweight, requires no retraining, and is broadly compatible with existing diffusion-model pipelines.

In this work, we focus on a practical active-verification setting, in which a protected diffusion system embeds a watermark during generation and a verifier later determines whether a suspicious image carries a valid watermark. Rather than pursuing user-level attribution or identity tracing, we focus on a softer and more practical verification objective. This choice avoids embedding user-specific information and thereby helps avoid introducing unnecessary external concerns related to privacy, freedom of expression, or governance. At the same time, by remaining plug-and-play and broadly compatible with existing pipelines, the proposed framework is particularly attractive for low-cost deployment in real-world generation systems.

Our main contributions are summarized as follows.

- *Perspective:* We revisit diffusion watermarking from the perspective of embedding space design and highlight its importance in balancing robustness and strict imperceptibility, rather than merely preserving high-level semantics.
- *Methodology:* We propose a plug-and-play watermarking method that embeds watermark information into the circular harmonic expansion of the initial noise in diffusion models. This embedding space enables the selection of components that are relatively robust to common distortions while remaining less disruptive to the generated output, and it does so without retraining and with broad compatibility with existing diffusion pipelines.
- *Results:* Extensive experiments show that the proposed method achieves consistently strong robustness against signal degradations, geometric transformations, and regeneration-based expert attacks, while substantially

improving imperceptibility over representative content-level baselines [12], [13], including an absolute SSIM improvement of at least 0.15.

II. RELATED WORK

This section reviews representative studies on generated-image detection for diffusion models. According to whether a detection method requires intervention before image distribution, existing approaches can be broadly divided into *passive detection*, which operates directly on released images, and *image watermarking*, which embeds verifiable information before release as a prior for subsequent verification.

A. Passive Detection for Diffusion Models

Passive detection methods for diffusion-generated images [14], [19]–[24] usually learn a discriminative feature space from large-scale collections of real and generated images. Their goal is to identify artifacts, statistical regularities, or model-specific traces that distinguish synthetic content from authentic images, without requiring any modification to the generation pipeline. Beyond image-level real/fake classification, passive image forensics also addresses localized manipulation traces, as exemplified by the dense feature interaction network of Yao et al. [25] for image inpainting localization.

Representative methods differ mainly in the type of features they exploit. Ojha et al. [20] use the feature space of large pretrained vision-language models and apply nearest-neighbor retrieval and linear probing to detect fake images across multiple generative models. Ricker et al. [14] leverage the autoencoder of latent diffusion models and use reconstruction error as a signal for distinguishing generated images from real ones. Sha et al. [19] further study both image-only and image-text joint models for detection and attribution, and analyze how textual information affects the realism of generated content. Other studies explore forensic cues derived from inversion behavior, spectral statistics, or cross-model generalization [21]–[24].

Although these methods have shown promising clean-set performance, passive detection still faces practical limitations. Since no explicit source-side prior is embedded into the image at generation time, the detector must rely on subtle statistical traces that may be weakened by common post-processing operations, image editing, compression, resizing, or distribution shifts. Maintaining reliable high-confidence detection under real-world distortions therefore remains challenging, motivating active forensic mechanisms that introduce stronger and more controllable signals during generation.

B. Image Watermarking for Diffusion Models

Image watermarking provides an active mechanism by embedding verifiable information into generated content before release. Recent studies have investigated watermarking specifically for Stable Diffusion models, including the method of Yuan et al. [26] and the flexible, lightweight framework of Luo et al. [27]. As summarized in Table I, diffusion watermarking methods can be roughly categorized as *pixel-level*

TABLE I
COMPARISON OF REPRESENTATIVE RELATED METHODS IN TERMS OF KEY PROPERTIES.

Method	Passive detection	Pixel-level watermarking			Content-level watermarking			
	AEROBLADE [14] CVPR'24	DwtDct [15] DW&S'07	RivaGAN [16] arXiv'19	Stable Signature [17] ICCV'23	Tree Ring [12] NeurIPS'23	Gaussian Shading [13] CVPR'24	AquaLoRA [18] ICML'24	Ours
Detection confidence		✓	✓	✓	✓	✓	✓	✓
Invariance and robustness					✓	✓	✓	✓
Imperceptibility	✓	✓	✓	✓			✓	✓
Plug-and-play	✓	✓	✓		✓	✓		✓

watermarking and *content-level watermarking*, depending on whether the watermark is injected into relatively microscopic image details or into components more tightly coupled with the generation process.

1) *Pixel-level watermarking*: Pixel-level watermarking usually operates after denoising, for example by embedding the watermark directly into the generated image [15], [16], [28]–[32] or by modifying downstream image-decoding modules [17], [33]–[35]. Because the embedding is applied to low-level visual components, such methods often preserve image fidelity well and can offer strong imperceptibility in the conventional pixel-level sense.

Representative post-hoc methods include traditional signal-processing methods such as DwtDct [15] and deep-learning-based approaches such as RivaGAN [16]; both can be integrated without changing the underlying diffusion model. Other methods adapt model components to carry watermarks more natively. Stable Signature [17] embeds a fixed watermark into the decoder of latent diffusion models through lightweight fine-tuning. Xiong et al. [33] introduce a message matrix and fuse it with intermediate representations to support flexible message embedding, while Ci et al. [35] incorporate user-specified watermarks through a lightweight adapter and hybrid tuning strategy.

Despite their fidelity advantages, pixel-level schemes are generally less robust to strong removal attempts, especially regeneration-based attacks. Since the watermark is often embedded after the main denoising trajectory or into components that affect only local visual details, it can be largely suppressed when the image is regenerated or re-rendered by another generative model [36].

2) *Content-level watermarking*: Content-level watermarking typically operates before or during denoising, for example by embedding watermark information into the initial noise [12], [13], [37], [38] or by modifying the denoiser through additional training [18], [39]. Coupling the watermark more closely with the generation process generally improves robustness to regeneration and other strong transformations.

Tree Ring [12] embeds a structured watermark into the Fourier space of the initial noise to improve detectability after transformations. Gaussian Shading [13] increases robustness by replicating watermark information and sampling a new initial noise for generation. Other methods explore source tracing or identity-aware watermark construction in the latent noise space [37], [38]. Training-based approaches such as AquaLoRA [18] incorporate watermarking capability into the U-Net and improve the robustness–imperceptibility trade-off compared with earlier content-level methods.

Content-level watermarking still faces several challenges. Robustness is often strengthened through fixed structural patterns, redundancy, or stronger perturbations, which may reduce pixel-level consistency with the corresponding original output even when image semantics remain largely preserved. Methods that retrain model components introduce additional computational cost and weaken plug-and-play deployability. Comparatively less attention has been paid to whether the embedding space itself is intrinsically aligned with both robustness and strict imperceptibility.

3) *Summary*: Overall, pixel-level methods usually preserve fidelity well but remain vulnerable to regeneration-based removal, whereas content-level methods are often more robust but may sacrifice reproducibility, strict imperceptibility, or deployment efficiency. These observations motivate an embedding-space view of diffusion watermarking, in which the transform domain is selected to support both stable detection and faithful reconstruction.

III. MATHEMATICAL PREPARATION

This section reviews the properties of circular harmonic expansion used by the proposed framework, including orthogonality, reconstructability, and the transformation of moment magnitudes under rotation and flipping. Throughout this section, V_{nm} denotes a circular harmonic basis function and $P_{nm}(f) = \langle f, V_{nm} \rangle$ denotes the corresponding moment. The PCET basis used in our implementation is specified in Eq. (11).

A. Orthogonality

Orthogonality means that two distinct basis functions, V_{nm} and $V_{n'm'}$, are mutually orthogonal, so their corresponding moments do not contain redundant information [40]. Over the unit disk, the orthogonality condition is

$$\iint_{x^2+y^2 \leq 1} V_{nm}(x, y) V_{n'm'}^*(x, y) dx dy = \delta_{nn'} \delta_{mm'}, \quad (1)$$

where δ denotes the Kronecker delta function, i.e., $\delta_{ij} = [i = j]$. The basis function V_{nm} can be decomposed into a radial basis function R_n and an angular basis function A_m as

$$V_{nm}(r, \theta) = R_n(r) A_m(\theta), \quad (2)$$

where $A_m(\theta) = e^{jm\theta}$. Substituting this decomposition into the orthogonality condition yields

$$\begin{aligned} & \int_0^{2\pi} \int_0^1 R_n(r) A_m(\theta) R_{n'}^*(r) A_{m'}^*(\theta) r dr d\theta \\ &= \int_0^{2\pi} A_m(\theta) A_{m'}^*(\theta) d\theta \int_0^1 R_n(r) R_{n'}^*(r) r dr. \end{aligned} \quad (3)$$

Since the angular basis functions satisfy

$$\int_0^{2\pi} A_m(\theta) A_{m'}^*(\theta) d\theta = \int_0^{2\pi} e^{im\theta} e^{-im'\theta} d\theta = 2\pi \delta_{mm'}, \quad (4)$$

the radial basis functions should satisfy

$$\int_0^1 R_n(r) R_{n'}^*(r) r dr = \frac{1}{2\pi} \delta_{nn'}. \quad (5)$$

The radial basis function adopted in our implementation, defined later in Eq. (12), satisfies this condition.

B. Reconstructability

Since the basis functions of circular harmonic expansion form an orthonormal basis on the unit disk, the expansion provides the optimal $\mathcal{L}^2$ approximation of a function over any selected index set [40]. Specifically, let $f \in \mathcal{L}^2(x^2 + y^2 \leq 1)$ and $\mathcal{S} \subseteq \mathbb{Z}^2$. For any set of complex coefficients $\{\alpha_{nm} : (n, m) \in \mathcal{S}\}$,

$$\begin{aligned} & \left\| f(x, y) - \sum_{(n, m) \in \mathcal{S}} P_{nm}(f) V_{nm}(x, y) \right\| \\ & \leq \left\| f(x, y) - \sum_{(n, m) \in \mathcal{S}} \alpha_{nm} V_{nm}(x, y) \right\|. \end{aligned} \quad (6)$$

Thus, among all linear combinations supported on the same index set $\mathcal{S}$, the coefficients given by the circular harmonic moments $P_{nm}(f)$ minimize the reconstruction error. Equality holds if and only if $\alpha_{nm} = P_{nm}(f)$.

C. Rotation Invariance

Let f^{rot} denote a rotated version of f with rotation angle β , i.e., $f^{\text{rot}}(r, \theta) = f(r, \theta + \beta)$. The corresponding moments satisfy

$$\begin{aligned} P_{nm}(f^{\text{rot}}) &= \int_0^{2\pi} \int_0^1 f(r, \theta + \beta) R_n^*(r) e^{-jm\theta} r dr d\theta \\ &= \int_0^{2\pi} \int_0^1 f(r, \theta) R_n^*(r) e^{-jm(\theta - \beta)} r dr d\theta \\ &= e^{jm\beta} P_{nm}(f). \end{aligned} \quad (7)$$

Taking the magnitude gives

$$|P_{nm}(f^{\text{rot}})| = |e^{jm\beta} P_{nm}(f)| = |P_{nm}(f)|. \quad (8)$$

Therefore, the magnitude of the circular harmonic moment is invariant to rotation.

D. Flipping Invariance

Let f^{fp} denote the vertically flipped version of f , i.e., $f^{\text{fp}}(r, \theta) = f(r, -\theta)$. The corresponding moments satisfy

$$\begin{aligned} P_{nm}(f^{\text{fp}}) &= \int_0^{2\pi} \int_0^1 f(r, -\theta) R_n^*(r) e^{-jm\theta} r dr d\theta \\ &= \int_0^{2\pi} \int_0^1 f(r, \theta) R_n^*(r) e^{jm\theta} r dr d\theta \\ &= P_{n, -m}(f). \end{aligned} \quad (9)$$

Taking the magnitude gives

$$|P_{nm}(f^{\text{fp}})| = |P_{n, -m}(f)|. \quad (10)$$

Thus, flipping reflects the moment magnitudes across $m = 0$. A reflection about another axis introduces a unit-magnitude phase factor in addition to the same mapping $m \leftrightarrow -m$, so the magnitude relation is unchanged. In the proposed method, the embedding support and watermark amplitudes are constructed symmetrically across $m = 0$; consequently, this permutation preserves the embedded magnitude pattern and the detection score.

IV. METHOD

In this section, we first describe the practical setting and design objective, as illustrated in Fig. 2. We then present the proposed watermarking framework, whose overall pipeline is shown in Fig. 3.

A. Problem Setting

We consider an API-based deployment with four roles, as illustrated in Fig. 2. Alice is the model provider who operates a proprietary diffusion model and embeds a preset watermark during image generation. Bob is a benign user who expects high-quality outputs with minimal interference from watermarking, whereas Mallory may generate misleading content and manipulate the released images to evade detection. Justin is a trusted verifier, such as a platform moderator, regulator, or judicial authority, who has authorized access to the protected model and detector configuration and assesses suspicious content using the prescribed verification procedure.

The design objective is to support reliable verification while preserving the usability of the underlying diffusion model. Accordingly, the watermarking mechanism should be plug-and-play for low-cost deployment, maintain high pairwise output consistency under matched generation settings, and remain detectable after common distortions and deliberate removal attempts. Here, imperceptibility follows the pairwise criterion defined in Section I, including pixel-level consistency and reproducibility under the same generation conditions.

Verification does not require generation-specific side information, including the original prompt and seed, a paired unwatermarked image, or the user's identity. A positive decision indicates that the recovered watermark evidence is consistent with the protected system's preset configuration.

B. Watermark Embedding

The framework is built on circular harmonic expansion, which provides the structural properties used by both watermark embedding and detection. Orthogonality and reconstructability support selective coefficient modification followed by inverse reconstruction of the initial noise; the explicit two-dimensional moment indices provide a coordinate system for specifying the embedding support; and the magnitude transformation laws support stable comparison under rotation and flipping. The analytical forward and inverse mappings

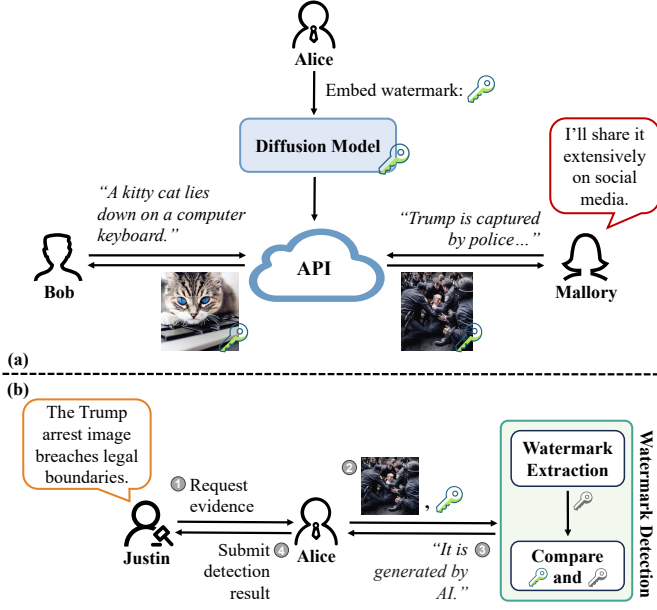


Fig. 2. Deployment setting considered in this work. (a) Bob and Mallory access Alice's image generation API. (b) For suspicious content, Justin assesses whether the submitted image exhibits watermark evidence consistent with Alice's protected generation system.

keep the procedure training-free and leave the diffusion-model parameters unchanged. These properties are reviewed in Section III.

1) *Mapping the initial noise to the circular harmonic space:* For the concrete implementation in this work, we instantiate the circular harmonic expansion with the Polar Complex Exponential Transform (PCET) [41]. PCET provides the explicit radial basis used in the subsequent moment computation and reconstruction. The framework-level properties described above follow from circular harmonic expansion. Let f denote the continuous form of the initial noise x_T . The PCET moment $P_{nm}(f)$ of order n and repetition m in polar coordinates (r, θ) is defined as

$$\begin{aligned} P_{nm}(f) &= \int_0^{2\pi} \int_0^1 f(r, \theta) V_{nm}^*(r, \theta) r dr d\theta \\ &= \int_0^{2\pi} \int_0^1 f(r, \theta) R_n^*(r) e^{-jm\theta} r dr d\theta, \end{aligned} \quad (11)$$

where $*$ denotes complex conjugation, $j = \sqrt{-1}$, and $(n, m) \in \mathbb{Z}^2$. The basis function is given by $V_{nm}(r, \theta) = R_n(r) e^{jm\theta}$, where R_n satisfies the orthogonality condition reviewed in Section III-A. For PCET, the radial basis function is

$$R_n(r) = \frac{1}{\sqrt{\pi}} \exp(j2n\pi r^2). \quad (12)$$

Intuitively, Eq. (11) maps the initial noise f into the space spanned by the basis set $\{V_{nm}\}$, and the resulting moments $\{P_{nm}(f)\}$ provide a complete representation of f in that space [42].

The magnitude transformation laws of circular harmonic expansion under rotation and flipping are detailed in Sections III-C and III-D. Performance under JPEG compression,

blur, additive noise, brightness changes, resizing, and the other tested distortions is evaluated empirically in Section V-E.

Let the maximum order be $K \in \mathbb{Z}$, and define

$$\Omega = \{(n, m) : |n| \leq K, |m| \leq K, n, m \in \mathbb{Z}\}. \quad (13)$$

Then the corresponding moment set is

$$\mathcal{P} = \{P_{nm}(f) : (n, m) \in \Omega\}, \quad (14)$$

with cardinality $|\mathcal{P}| = (2K + 1)^2$.

2) *Embedding the watermark into selected moments:* We define the embedding support in the $n > 0$ half-plane, covering the first and second quadrants of the index plane. Specifically, let the embedding region be

$$D \subset \{(n, m) \in \Omega : n > 0\}. \quad (15)$$

Here, D denotes a finite, prespecified set of coefficient indices within the allowed range. To match the flipping transformation, it is symmetric across $m = 0$: if $(n, m) \in D$, then $(n, -m) \in D$. Let the binary watermark be

$$\mathcal{W} = \{w_{nm} : w_{nm} \in \{0, 1\}, (n, m) \in D\}, \quad (16)$$

where $|\mathcal{W}| = |D|$.

The watermark amplitudes follow the same symmetry, with $w_{n,m} = w_{n,-m}$. The embedding and verification stages use the same pair $(D, \mathcal{W})$. The restriction $n > 0$ retains one representative from each centrosymmetric pair, and the corresponding coefficients in $n < 0$ are assigned by the conjugate-symmetry update below.

We use an embedding function E to modify the selected moments $\{P_{nm}(f) : (n, m) \in D\}$ according to the watermark bits:

$$Q_{nm} = E(P_{nm}, w_{nm}) = \begin{cases} \alpha_1 e^{j\varphi_{nm}}, & w_{nm} = 0, \\ \alpha_2 e^{j\varphi_{nm}}, & w_{nm} = 1, \end{cases} \quad (17)$$

where Q_{nm} denotes the watermarked version of $P_{nm}(f)$, $(\alpha_1, \alpha_2) \in \mathbb{R}_{\geq 0}^2$ with $\alpha_1 \neq \alpha_2$, and $\varphi_{nm} = \arg(P_{nm}(f))$. Equation (17) assigns α_1 or α_2 to the magnitude of each selected moment and retains its original phase.

The resulting watermarked moment set is

$$\begin{aligned} \mathcal{T} &= \{T_{nm} : (n, m) \in \Omega\} \\ &= \{P_{nm}(f) : (n, m) \in \Omega \setminus D\} \cup \{Q_{nm} : (n, m) \in D\}. \end{aligned} \quad (18)$$

Since PCET moments at centrosymmetric indices are complex conjugates, we enforce the same symmetry in the modified moment set. Therefore, for indices mirrored to the third and fourth quadrants, the moments are updated as

$$T_{nm} = T_{-n, -m}^*, \quad n < 0. \quad (19)$$

Together, the embedding and conjugate-symmetry updates change the coefficients in D and their centrosymmetric counterparts. Coefficients at all other indices retain their original values. With an orthogonal circular harmonic basis, the change in the reconstructed initial noise is determined by the coefficient differences on this selected support. Thus, D specifies the scope of the modification, and (α_1, α_2) specify the target magnitudes. The pairwise metrics in Table V and

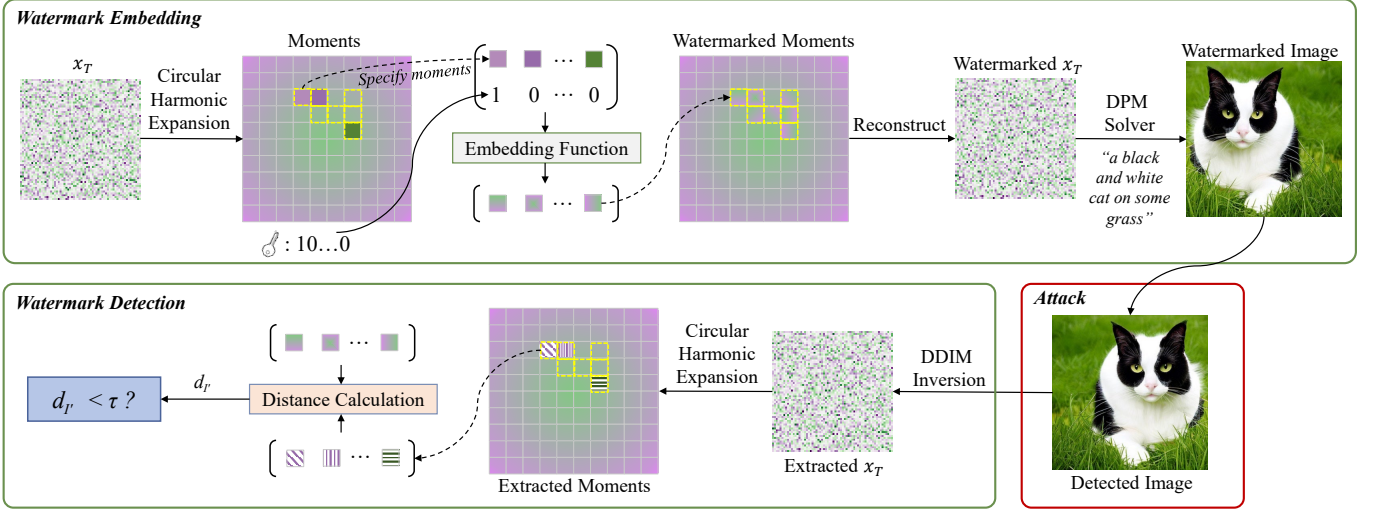


Fig. 3. Overview of the proposed framework. The pipeline consists of two stages: watermark embedding and watermark detection. The embedding function modifies the moments within the embedding region (yellow dotted box) according to the preset watermark, as defined in Eq. (17). During detection, the extracted moments are compared with the target amplitudes in the same region to compute the decision distance, as defined in Eq. (22).

the qualitative differences in Fig. 6 and Appendix provide empirical evidence of the resulting output consistency under the evaluated settings.

3) *Inverse mapping and reconstruction of the initial noise:* The watermarked initial noise is reconstructed by mapping the modified moment set $\mathcal{T}$ back to the spatial domain:

$$f_w(r, \theta) = \sum_{n=-K}^K \sum_{m=-K}^K T_{nm} V_{nm}(r, \theta). \quad (20)$$

The reconstructed initial noise f_w is then fed into the diffusion model to generate the watermarked image I_w .

4) *Why reconstructability matters:* The reconstructability of circular harmonic expansion follows from the orthogonality of the basis functions (see Section III-B). This property is important for watermark embedding because it allows us to modify selected coefficients while still reconstructing an initial noise that remains close to the original one. In practice, this helps preserve *imperceptibility*. In general, using a larger value of K reduces truncation error and improves reconstruction fidelity, although it also increases computational cost.

C. Watermark Detection

1) *Recovering the initial noise by DDIM inversion:* Given a query image I' , we first apply DDIM inversion [43] to estimate the corresponding initial noise. DDIM inversion reverses the denoising trajectory and provides an approximation of the initial noise x_T that generated the image x_0 . Although the inversion is not exact, the residual estimation error does not prevent stable detection in practice. Let the recovered noise be denoted by f' .

2) *Comparing the watermark in the circular harmonic space:* Given the maximum order K , the recovered noise f' is mapped into the same circular harmonic space using Eq. (11). This yields the moment set

$$\mathcal{P}' = \{P'_{nm}(f') : (n, m) \in \Omega\}. \quad (21)$$

To determine whether the query image I' is watermarked, we compute a distance score $d_{I'}$. Given the embedding region D and the preset watermark $\mathcal{W}$, the score is defined as

$$d_{I'} = \frac{1}{|D|} \sum_{(n,m) \in D} d_{nm}, \quad (22)$$

where d_{nm} denotes the absolute difference between the target magnitude determined by the watermark bit and the extracted magnitude:

$$d_{nm} = \begin{cases} |\alpha_1 - |P'_{nm}||, & w_{nm} = 0, \\ |\alpha_2 - |P'_{nm}||, & w_{nm} = 1. \end{cases} \quad (23)$$

Because D contains one representative from each centrosymmetric conjugate pair, each pair contributes once to the average in Eq. (22). The score is nonnegative and equals zero when the extracted magnitudes exactly match the target pattern; closer matches yield smaller values. Its absolute scale depends on the protected model and detector configuration. For a given deployment configuration, I' is classified as watermarked when $d_{I'} < \tau$, where τ denotes the decision threshold.

3) *Why magnitude-based comparison improves robustness:* As shown in Sections III-C and III-D, rotation multiplies each circular harmonic moment by a unit-magnitude phase factor, whereas flipping maps its magnitude from index (n, m) to $(n, -m)$. Therefore, rotation preserves each selected magnitude, and flipping permutes the selected magnitudes across $m = 0$. Because $(D, \mathcal{W})$ is symmetric across $m = 0$, both transformations preserve the target magnitude pattern used in Eqs. (22) and (23). Comparing magnitudes rather than complex-valued moments consequently improves robustness to rotation and flipping while remaining consistent with the embedding rule in Eq. (17).

V. EXPERIMENTS



Fig. 4. Examples of image regeneration attacks used in Section V-C. (a) No attack; (b) VAE-based Bmshj'18 with $q = 3$; (c) VAE-based Cheng'20 with $q = 3$; (d) diffusion-model-based regeneration using SDv2.1 with $step = 60$.

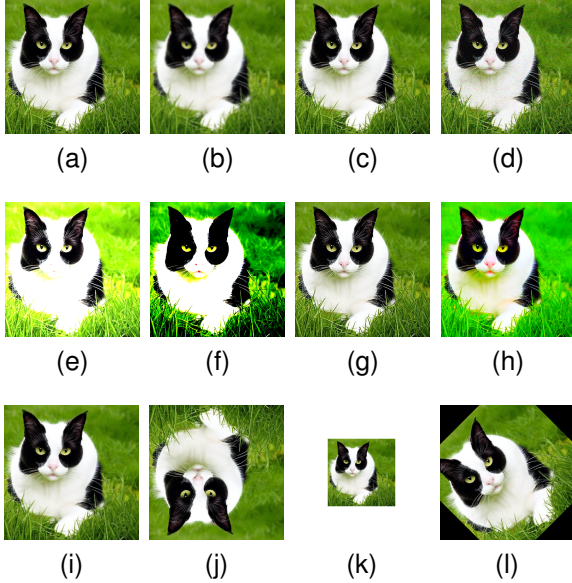


Fig. 5. Examples of natural distortions used in Section V-E. (a) No distortion; (b) Gaussian blur with $r = 2$; (c) JPEG compression with $qf = 70$; (d) Gaussian noise with $\mu = 0$ and $\sigma = 0.1$; (e) brightness adjustment with $f = 2$; (f) contrast adjustment with $f = 4$; (g) sharpness adjustment with $f = 4$; (h) saturation adjustment with $f = 4$; (i) median blur with $k = 3$; (j) vertical flip; (k) resizing with $f = 0.5$; (l) rotation with $degree = 45$.

This section evaluates the proposed method from three complementary perspectives: *imperceptibility*, *robustness*, and practical deployability. For imperceptibility, corresponding unwatermarked and watermarked outputs are generated with the same diffusion model, prompt, seed, sampler, and other generation settings to assess pairwise output consistency. We then evaluate robustness under common distortions and regeneration-based attacks, while keeping the method in a plug-and-play setting without retraining.

A. Experimental Setup

1) *Default experimental configuration*: Unless otherwise specified, all experiments use Stable Diffusion v2.1 (SDv2.1) [8]. Images are generated with the DPM-Solver sampler [44] using 50 sampling steps, a classifier-free guidance (CFG) scale of 7.5, and a resolution of 512×512 . Detection uses 50-step DDIM inversion with a CFG scale of 1 and an empty prompt, reflecting the unavailability of the original generation prompt at detection time.

For the proposed method, we use $K = 250$, $\alpha_1 = 0.001$, and $\alpha_2 = 0.04$. To reduce bias from manually selecting

individual coefficient positions, we construct D by mapping a fixed handwritten digit “8” template to the PCET index plane and reflecting the selected positions across $m = 0$. The paired positions share the same assigned amplitude, yielding a reproducible set of 176 selected coefficient positions near the origin in the non-DC, low-order region. Embedding and verification use the resulting index set and amplitude pattern.

We use captions from the validation split of MS-COCO [45] as prompts. The benchmarking experiment uses 5,000 generated images for metric computation; for the remaining experiments, we randomly sample 1,000 images from this set.

2) *Watermark baselines*: We compare our method with four pixel-level watermarking methods, namely DwtDct [15], DwtDctSvd [15], RivaGAN [16], and Stable Signature [17], as well as two training-free content-level watermarking methods, Tree Ring [12] and Gaussian Shading [13]. These baselines cover representative design choices with different trade-offs, as summarized in Table I. Unless otherwise noted, all baseline methods are evaluated using their default configurations.

3) *Metrics*: We evaluate pairwise output consistency primarily with Structural Similarity Index (SSIM) [46] and Learned Perceptual Image Patch Similarity (LPIPS) [47] on matched image pairs, together with Watermarked-to-Original Fréchet Inception Distance (WO-FID) [48] on the corresponding output sets. SSIM and LPIPS quantify structural and perceptual differences at the image-pair level, while WO-FID measures the distribution shift between the watermarked and unwatermarked sets. We additionally report CLIP-Score [49] and Generated-to-Real Fréchet Inception Distance (GR-FID) [48] as auxiliary generation-quality metrics. CLIP-Score measures prompt-image alignment, and GR-FID measures the distribution gap between generated and real images.

For watermark detection and robustness evaluation, each evaluated condition yields an empirical receiver operating characteristic (ROC) curve constructed from the detection scores of its watermarked and unwatermarked samples. We compute the area under the curve (AUC) over the complete ROC curve. The true positive rate (TPR) is reported at the empirical operating point with the highest TPR among thresholds satisfying a false positive rate (FPR) below 1%; the ROC curve and operating point are determined separately for each evaluated condition. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

B. Benchmarking Experiments

Following the evaluation protocol adopted in Tree Ring [12] and Gaussian Shading [13], we compare our method with the baselines from two perspectives: 1) image generation quality, measured by CLIP-Score and GR-FID; and 2) watermark detection performance in the clean setting, measured by AUC and TPR. The results are reported in Table II.

For CLIP-Score, all methods produce results close to that of the original model, indicating similar prompt-image alignment under this metric. For GR-FID, all methods except Stable Signature and Tree Ring achieve values comparable to or

TABLE II
BENCHMARK RESULTS. SDv2.1 DENOTES THE ORIGINAL UNWATERMARKED MODEL. AUC AND TPR ARE COMPUTED ON CLEAN WATERMARKED IMAGES WITHOUT ANY ATTACK.

Method	CLIP-Score $\uparrow$	Metrics		
		GR-FID $\downarrow$	AUC $\uparrow$	TPR $\uparrow$
SDv2.1	0.329	25.06	-	-
DwtDct	0.329	24.57	0.9951	0.889
DwtDctSvd	0.329	24.67	1.0000	1.000
RivaGAN	0.327	24.28	0.9999	0.998
Stable Signature	0.331	25.63	1.0000	1.000
Tree Ring	0.329	25.58	1.0000	0.999
Gaussian Shading	0.329	24.44	1.0000	1.000
Ours	0.329	24.06	1.0000	1.000

TABLE III
RESULTS UNDER IMAGE REGENERATION ATTACKS, REPORTING TPR FOR PIXEL-LEVEL AND CONTENT-LEVEL METHODS. “DM” DENOTES DIFFUSION-MODEL-BASED ATTACKS.

Method	VAE based		DM based	
	Bmshj’18	Cheng’20	SDv2.1	Average
<i>Pixel-level</i>				0.165
DwtDct	0.005	0.002	0.003	0.003
DwtDctSvd	0.103	0.124	0.230	0.152
RivaGAN	0.014	0.017	0.123	0.051
Stable Signature	0.541	0.813	0.003	0.452
<i>Content-level</i>				0.987
Tree Ring	0.976	0.993	0.943	0.971
Gaussian Shading	1.000	1.000	1.000	1.000
Ours	0.990	0.983	1.000	0.991

lower than that of the original model, and our method has the lowest value in this comparison. For clean-image detection, all methods except DwtDct achieve near-saturated AUC and TPR.

Accordingly, the clean-detection results in Table II serve as a sanity check that establishes normal detector operation without attacks. Together with the closely grouped generation-quality metrics, they provide a common clean reference for the subsequent evaluation. The following analyses focus on regeneration robustness in Tables III–IV and pairwise output consistency in Table V, where the differences among design choices become more informative.

C. Security Experiments

In this experiment, we evaluate robustness against regeneration-based removal attacks following Zhao et al. [36]. This attack family is particularly relevant in diffusion watermarking because it can substantially weaken many imperceptible watermarks while preserving visual quality, and therefore serves as a strong test of the robustness–imperceptibility trade-off. We consider three attack instances: two VAE-based variants, Bmshj’18 [50] and Cheng’20 [51], and one diffusion-model-based variant using SDv2.1. Example outputs are shown in Fig. 4.

Tables III and IV show a clear category-level separation under regeneration attacks: the pixel-level methods have an av-

TABLE IV
RESULTS UNDER IMAGE REGENERATION ATTACKS, REPORTING AUC FOR PIXEL-LEVEL AND CONTENT-LEVEL METHODS.

Method	VAE based		DM based	
	Bmshj’18	Cheng’20	SDv2.1	Average
<i>Pixel-level</i>				0.6969
DwtDct	0.4856	0.4898	0.5012	0.4922
DwtDctSvd	0.8088	0.8161	0.8193	0.8147
RivaGAN	0.5855	0.5563	0.8247	0.6555
Stable Signature	0.9462	0.9753	0.5537	0.8251
<i>Content-level</i>				0.9990
Tree Ring	0.9987	0.9998	0.9957	0.9981
Gaussian Shading	1.0000	1.0000	1.0000	1.0000
Ours	0.9987	0.9984	1.0000	0.9990

TABLE V
QUANTITATIVE PAIRWISE OUTPUT-CONSISTENCY RESULTS FOR IMPERCEPTIBILITY EVALUATION. IN EACH COLUMN, THE BEST RESULT IS HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST IS MARKED WITH AN ASTERISK (*).

Method	Metrics		
	SSIM $\uparrow$	LPIPS $\downarrow$	WO-FID $\downarrow$
Tree Ring	0.47	0.50	43.81
Gaussian Shading	0.20	0.74	48.32
Ours ($\alpha_2 = 0.02$)	0.75	0.20	26.50
Ours ($\alpha_2 = 0.04$)	0.62*	0.31*	35.02*

erage TPR of 0.165 and an average AUC of 0.6969, while the content-level methods reach 0.987 and 0.9990, respectively. The evaluated regeneration attacks reconstruct the released image through a VAE or diffusion model, replacing many output pixels and weakening watermark signals carried by low-level image details. The content-level methods considered here embed their signals before or within denoising, coupling the watermark with the generation process; under the evaluated attacks, these signals remain substantially more detectable. The proposed method belongs to this category because it modifies the circular harmonic moments of the initial noise before denoising. With regeneration robustness established for all three content-level methods, the subsequent comparison examines how faithfully each method preserves the corresponding unwatermarked output through the pairwise output-consistency evaluation in Section V-D.

D. Imperceptibility Experiments

We evaluate imperceptibility using the pairwise output-consistency criterion defined in Section I. For each sample, the corresponding unwatermarked and watermarked outputs use the same diffusion model, prompt, seed, sampler, and other generation settings. The evaluation combines qualitative comparison of these matched outputs with SSIM and LPIPS at the image-pair level and WO-FID at the set level. This distinction is also reflected in the existing results: the closely clustered CLIP-Scores in Table II coexist with substantial differences in SSIM, LPIPS, and WO-FID in Table V.

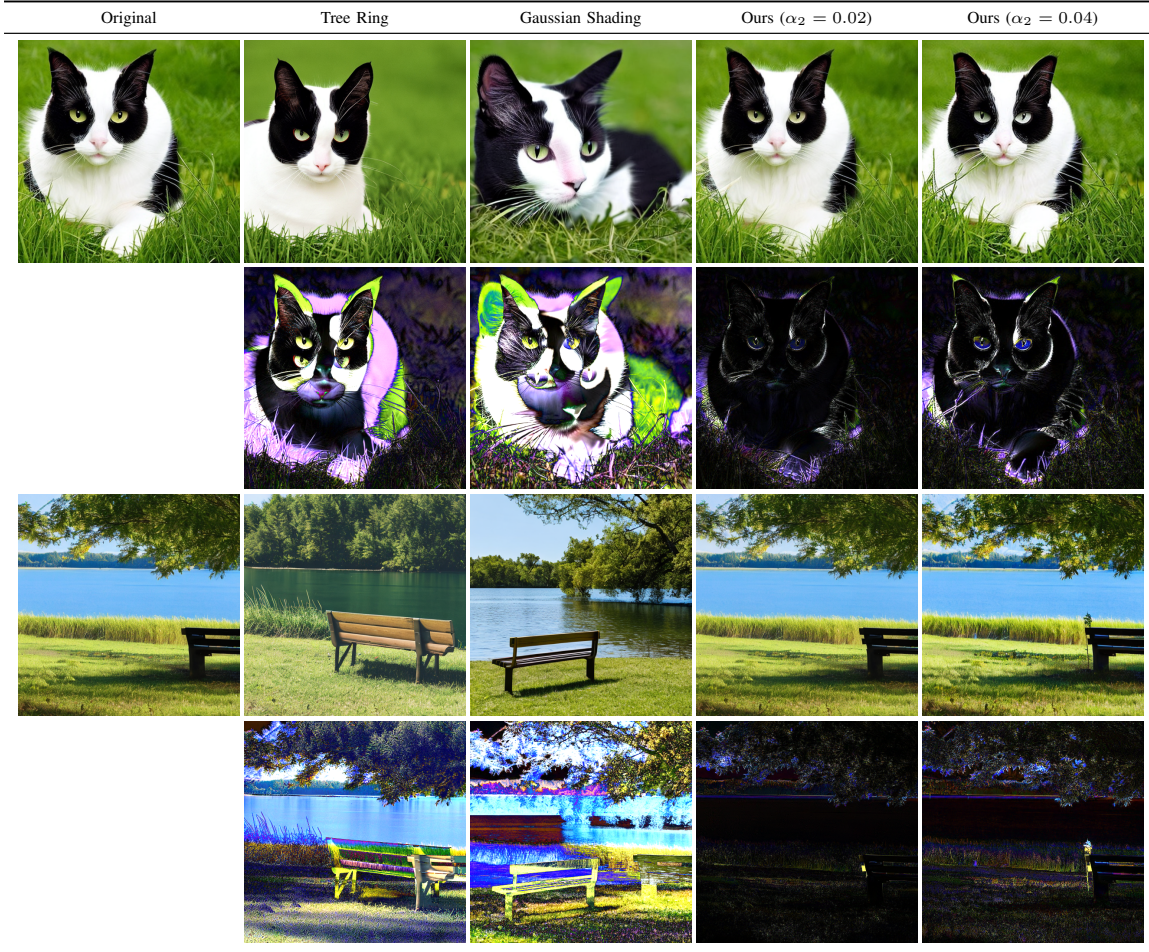


Fig. 6. Qualitative pairwise output-consistency comparison for imperceptibility evaluation. Each example shows the unwatermarked image, the watermarked images generated under the same settings, and their RGB difference maps, computed channelwise as $D_{\text{vis}} = \text{clip}(2|I_w - I_0|, 0, 255)$. The factor 2 is used only for visualization; SSIM, LPIPS, and WO-FID are computed from the unscaled images. The prompts are: 1) a black and white cat on some grass; and 2) a bench next to tall grass at the edge of a body of water. More qualitative examples are provided in Appendix.

In this experiment, we report qualitative comparisons in Fig. 6 and quantitative results in Table V. With α_1 fixed at 0.001, the two values of α_2 represent operating points of the same method. The $\alpha_2 = 0.02$ setting applies a milder magnitude modification and is included to show the high pairwise output consistency achievable at this operating point. The default $\alpha_2 = 0.04$ setting provides a larger nominal separation between the two target magnitude levels and is used in the main robustness evaluations.

For the qualitative comparison, we generate corresponding unwatermarked and watermarked outputs using the same diffusion model, prompt, seed, sampler, and other generation settings. Let I_0 and I_w denote a matched pair of unwatermarked and watermarked RGB images. Their displayed difference is computed independently in each channel, after conversion to a signed representation, as $D_{\text{vis}} = \text{clip}(2|I_w - I_0|, 0, 255)$. The factor 2 is used only to improve visibility, and the maps serve as qualitative evidence; SSIM, LPIPS, and WO-FID are computed directly from the unscaled images. In the examples shown, Gaussian Shading produces the largest visual deviations from the corresponding unwatermarked outputs, and Tree Ring preserves broad semantic content while showing

more changes in local structures. The outputs from our method remain closer to their corresponding unwatermarked images, supporting stronger pairwise output consistency under the criterion adopted in this work. Additional qualitative examples are provided in Appendix.

Table V quantifies the matched-output comparison using SSIM, LPIPS, and WO-FID. The $\alpha_2 = 0.02$ operating point achieves the best result in all three metrics, and the default $\alpha_2 = 0.04$ operating point ranks second. Relative to Tree Ring, the corresponding SSIM values represent absolute gains of 0.28 and 0.15, respectively. LPIPS and WO-FID exhibit the same ordering. Under the matched generation settings, these results indicate stronger pairwise output consistency than the evaluated content-level baselines.

E. Robustness Experiments

We next evaluate robustness under common post-processing operations. A practical watermarking method should remain detectable after realistic distortions and transformations while preserving the original generation behavior. The 11 evaluated distortions are organized into signal degradations (Gaussian blur, JPEG compression, Gaussian noise, brightness, contrast,

TABLE VI

ROBUSTNESS RESULTS, REPORTING TPR UNDER DIFFERENT NATURAL DISTORTIONS. THE TOP OF THE TABLE LISTS THE DISTORTION TYPES AND THEIR CORRESPONDING INTENSITY PARAMETERS, WHERE “GAU.” STANDS FOR GAUSSIAN, “MED.” FOR MEDIAN, “VER.” FOR VERTICAL, AND “HOR.” FOR HORIZONTAL. FOR GAU. NOISE, THE PARAMETER IS σ . THE ROW HEADERS REPRESENT METHODS, WITH “T.R.” REFERRING TO TREE RING AND “G.S.” TO GAUSSIAN SHADING. IN THE LAST COLUMN, THE BEST VALUE IS HIGHLIGHTED IN **BOLD**. VALUES BELOW 0.5 ARE MARKED WITH A DAGGER ($\dagger$) THROUGHOUT THE TABLE.

	Signal Degradations														Geometric Transformations						Avg.		
	Gau. Blur		JPEG		Gau. Noise		Brightness		Contrast		Sharpness		Saturation		Med. Blur		Flip		Resize			Rotation	
	2	4	70	50	0.1	0.2	2	6	4	6	4	6	4	6	3	7	Ver.	Hor.	0.5	0.3		45	75
T.R.	1.000	0.999	1.000	0.999	0.994	0.958	0.999	0.916	0.992	0.978	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.216 [†]	0.314 [†]	0.926
G.S.	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.989	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.029 [†]	0.030 [†]	1.000	1.000	0.008 [†]	0.010 [†]	0.821
Ours	1.000	0.969	1.000	1.000	0.981	0.919	1.000	0.883	0.993	0.980	1.000	1.000	0.999	0.996	1.000	1.000	0.994	0.999	1.000	1.000	0.944	0.900	0.980

TABLE VII

ROBUSTNESS RESULTS, REPORTING AUC UNDER DIFFERENT NATURAL DISTORTIONS.

	Signal Degradations														Geometric Transformations								Avg.
	Gau. Blur		JPEG		Gau. Noise		Brightness		Contrast		Sharpness		Saturation		Med. Blur		Flip		Resize		Rotation		
	2	4	70	50	0.1	0.2	2	6	4	6	4	6	4	6	3	7	Ver.	Hor.	0.5	0.3	45	75	
T.R.	1.0000	1.0000	1.0000	1.0000	0.9996	0.9974	0.9999	0.9944	0.9995	0.9986	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9267	0.9408	0.9935
G.S.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.5990	0.6014	1.0000	1.0000	0.4928	0.4889	0.9173
Ours	1.0000	0.9978	1.0000	1.0000	0.9973	0.9860	1.0000	0.9802	0.9997	0.9984	1.0000	1.0000	0.9999	0.9997	1.0000	1.0000	0.9995	1.0000	1.0000	1.0000	0.9949	0.9915	0.9975

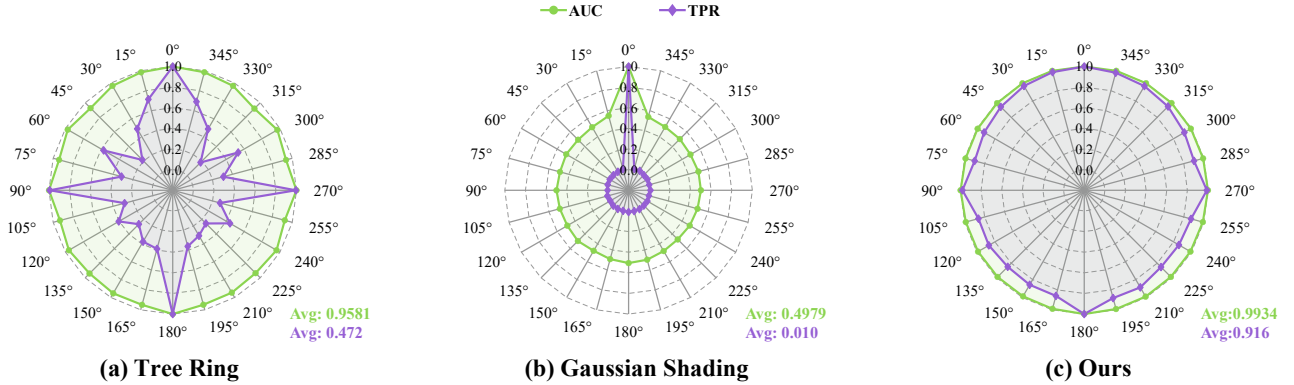


Fig. 7. AUC and TPR under rotation at different angles. The rotation angle ranges from 15 to 345 degrees in increments of 15 degrees. The average AUC and TPR for each method are shown in the lower-right corner of each radar plot in green and purple, respectively.

sharpness, saturation, and median blur) and geometric transformations (flipping, resizing, and rotation), following the two groups in Tables VI and VII. Example outputs are shown in Fig. 5.

Under the eight signal-degradation types, all three content-level methods generally maintain high AUC and TPR across the evaluated settings. These results characterize the empirical stability of each complete embedding-and-detection pipeline under the tested degradations.

The clearest differences appear under geometric transformations. All three methods remain stable under resizing. Under flipping, Gaussian Shading shows a pronounced decrease; Tree Ring and our method retain high detection performance. The detailed rotation results in Fig. 7 further separate the methods. Tree Ring maintains relatively high AUC, but its TPR varies periodically with the rotation angle, recovering near multiples of 90 degrees and often reaching local minima near odd multiples of 45 degrees. Gaussian Shading has low average AUC and TPR under rotation. Our method remains stable across the tested angular range; at the representative angles

of 45 and 75 degrees in Table VI, its TPR is 0.944 and 0.900, respectively.

The magnitude transformation laws of circular harmonic expansion provide the direct analytical basis for our method’s stability under rotation and flipping, as detailed in Sections III-C and III-D. The signal-degradation and resizing results report the empirical behavior of the evaluated pipelines under the tested settings. Overall, the proposed method maintains strong detection under both groups, with its most evident advantage occurring for the tested arbitrary-angle rotations. Together with Section V-D, these results indicate a favorable balance between robustness and pairwise output consistency among the evaluated content-level methods.

F. Ablation Experiments

In this experiment, we study several factors that may affect performance. Specifically, we vary the diffusion model version, CFG scale, generation steps, and inversion steps, and report both imperceptibility metrics (SSIM, LPIPS, and WO-FID) and detection metrics (AUC and TPR). The results are

TABLE VIII
ABLATION RESULTS. AUC AND TPR ARE AVERAGED OVER THE CLEAN
CONDITION AND ALL 22 NATURAL-DISTORTION CONDITIONS REPORTED
IN TABLE VI. DEFAULT SETTINGS ARE HIGHLIGHTED IN GRAY.

Factors		Metrics				
		SSIM↑	LPIPS↓	WO-FID↓	AUC↑	TPR↑
Models	v1.4	0.63	0.30	35.46	0.9983	0.984
	v2.0	0.63	0.31	34.61	0.9981	0.983
	v2.1	0.62	0.31	35.02	0.9976	0.981
CFG	6	0.63	0.31	35.22	0.9982	0.985
	7.5	0.62	0.31	35.02	0.9976	0.981
	14	0.59	0.33	34.51	0.9951	0.956
Generation Steps	25	0.61	0.32	35.18	0.9976	0.979
	50	0.62	0.31	35.02	0.9976	0.981
	75	0.63	0.31	34.82	0.9979	0.984
Inversion Steps	25	-	-	-	0.9975	0.983
	50	-	-	-	0.9976	0.981
	75	-	-	-	0.9976	0.981

summarized in Table VIII. Since our main claim is that the proposed embedding strategy provides a favorable robustness–imperceptibility trade-off in a plug-and-play setting, the ablation focuses on whether this behavior remains stable under practical changes to the generation and detection pipeline.

1) *Models*: We consider three commonly used versions of Stable Diffusion. The results show that changing the backbone version leads to only minor fluctuations across all five metrics, suggesting that the proposed method generalizes well across different Stable Diffusion releases.

2) *Classifier-Free Guidance (CFG) scale*: The CFG scale controls how strongly the generated image follows the input prompt. Although the default setting is 7.5, users often adjust it over a wider range in practice. We therefore evaluate two boundary cases, 6 and 14. When the CFG scale increases to 14, SSIM, LPIPS, AUC, and TPR become slightly worse than those under the default setting, although the changes remain small in magnitude. A possible reason is that a larger generation-time CFG scale introduces additional mismatch for DDIM inversion, since the true CFG value is unavailable at detection time [12]. Overall, the method remains relatively stable across this practical range of CFG settings.

3) *Generation and inversion steps*: The generation step controls how many denoising iterations are used to transform the initial noise into an image. In practice, users often vary this number to balance quality and efficiency, so we test 25, 50, and 75 steps. The results show that changing the generation step only slightly affects both imperceptibility and robustness, with a mild tendency toward improved performance at larger step counts.

During detection, the original generation step is typically unknown. We therefore also vary the inversion step and observe that its influence on AUC and TPR is minimal. This suggests that the proposed detector is not overly sensitive to moderate mismatch in inversion settings.

VI. CONCLUSION

In this work, we study diffusion model watermarking from the perspective of two practical requirements, namely imperceptibility and robustness. Existing methods often face a trade-off between these two objectives, partly because their embedding spaces are not well aligned with image components that are both stable under distortions and minimally disruptive to generation quality. In particular, we emphasize that imperceptibility in this setting should not be relaxed to merely preserving semantics, since such a loose criterion may still alter reproducibility and the original generation behavior.

To address this issue, we propose a watermarking framework that embeds information into the circular harmonic space of the initial noise. This design leverages several useful properties of that space, including reconstructability for faithful recovery, frequency-domain controllability for low-visibility embedding, and invariance-related characteristics for improved robustness under common distortions and geometric transformations. At the same time, the proposed method remains plug-and-play, requires no retraining, and is broadly compatible with existing diffusion-model pipelines.

Extensive experiments show that the proposed method achieves strong robustness against regeneration attacks and a wide range of natural distortions, while also providing favorable imperceptibility compared with representative baselines. These results suggest that embedding watermarks in an appropriate invariant space is a promising direction for diffusion model watermarking, especially when strict imperceptibility and practical deployability are both taken into account. In addition, we focus on practical active verification of AI-generated images in a cooperative deployment setting rather than user-level attribution or identity-oriented tracing, which keeps the problem scope useful and well aligned with the core technical objective.

There are several directions for future work. First, although the current method performs well in our setting, its detection pipeline relies on DDIM inversion, whose approximation error may affect detection accuracy; improving inversion fidelity and extending compatibility to a wider range of sampling pipelines would therefore be valuable. Second, the circular harmonic transform and its inverse involve nontrivial matrix computation, especially for larger K , so reducing computational overhead and improving efficiency in resource-constrained scenarios remain important practical goals. Third, since the method modifies the deployed generation pipeline, real-world adoption would benefit from further study of secure deployment mechanisms, including access control, model protection, and safeguards against unauthorized modification.

REFERENCES

- [1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, pp. 6840–6851, 2020.
- [2] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*, 2015, pp. 2256–2265.
- [3] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in neural information processing systems*, 2019.

- [4] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.
- [5] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=PIKWVd2yBkY>
- [6] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*, 2021, pp. 8162–8171.
- [7] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.
- [8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.
- [11] S. Dathathri, A. See, S. Ghaissas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova *et al.*, "Scalable watermarking for identifying large language model outputs," *Nature*, vol. 634, no. 8035, pp. 818–823, 2024.
- [12] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-rings watermarks: Invisible fingerprints for diffusion images," in *Advances in Neural Information Processing Systems*, 2023, pp. 58047–58063.
- [13] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. 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, pp. 12162–12171.
- [14] J. Ricker, D. Lukovnikov, and A. Fischer, "Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9130–9140.
- [15] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*. Morgan kaufmann, 2007.
- [16] K. A. Zhang, L. Xu, A. Cuesta-Infante, and K. Veeramachaneni, "Robust invisible video watermarking with attention," *arXiv preprint arXiv:1909.01285*, 2019.
- [17] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22466–22477.
- [18] W. Feng, W. Zhou, J. He, J. Zhang, T. Wei, G. Li, T. Zhang, W. Zhang, and N. Yu, "AquaLoRA: Toward white-box protection for customized stable diffusion models via watermark LoRA," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 13423–13444.
- [19] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3418–3432.
- [20] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24480–24489.
- [21] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [22] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model deepfakes," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP*, 2024, pp. 446–457.
- [23] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22445–22455.
- [24] S. Qi, Y. Zhang, C. Wang, Z. Xia, J. Weng, and X. Cao, "Hierarchical invariance for robust and interpretable vision tasks at larger scales," *arXiv preprint arXiv:2402.15430*, 2024.
- [25] Y. Yao, T. Han, S. Jia, and S. Lyu, "Dense feature interaction network for image inpainting localization," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 1636–1648, 2025.
- [26] Z. Yuan, L. Li, Z. Wang, and X. Zhang, "Watermarking for stable diffusion models," *IEEE Internet of Things Journal*, vol. 11, no. 21, pp. 35238–35249, 2024.
- [27] H. Luo, L. Li, and X. Zhang, "A flexible and lightweight watermarking framework for stable diffusion models," *IEEE Internet of Things Journal*, vol. 13, no. 7, pp. 13950–13964, 2026.
- [28] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou, and M. Douze, "Watermarking images in self-supervised latent spaces," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3054–3058.
- [29] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, 2018, p. 682–697.
- [30] Z. Jia, H. Fang, and W. Zhang, "Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 41–49.
- [31] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2117–2126.
- [32] X. Kang, J. Huang, Y. Q. Shi, and Y. Lin, "A dwt-dft composite watermarking scheme robust to both affine transform and jpeg compression," *IEEE transactions on circuits and systems for video technology*, pp. 776–786, 2003.
- [33] C. Xiong, C. Qin, G. Feng, and X. Zhang, "Flexible and secure watermarking for latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1668–1676.
- [34] C. Kim, K. Min, M. Patel, S. Cheng, and Y. Yang, "Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8974–8983.
- [35] H. Ci, Y. Song, P. Yang, J. Xie, and M. Z. Shou, "Wmadapter: Adding watermark control to latent diffusion models," *arXiv preprint arXiv:2406.08337*, 2024.
- [36] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li, "Invisible image watermarks are provably removable using generative ai," *arXiv preprint arXiv:2306.01953*, 2023.
- [37] H. Ci, P. Yang, Y. Song, and M. Z. Shou, "Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification," *arXiv preprint arXiv:2404.14055*, 2024.
- [38] L. Lei, K. Gai, J. Yu, and L. Zhu, "Diffustrace: A transparent and flexible watermarking scheme for latent diffusion model," *arXiv preprint arXiv:2405.02696*, 2024.
- [39] R. Min, S. Li, H. Chen, and M. Cheng, "A watermark-conditioned diffusion model for ip protection," *arXiv preprint arXiv:2403.10893*, 2024.
- [40] T. V. Hoang, "Image representations for pattern recognition," Ph.D. dissertation, Université Nancy II, 2011.
- [41] P.-T. Yap, X. Jiang, and A. C. Kot, "Two-dimensional polar harmonic transforms for invariant image representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1259–1270, 2009.
- [42] S. Qi, Y. Zhang, C. Wang, J. Zhou, and X. Cao, "A survey of orthogonal moments for image representation: Theory, implementation, and evaluation," *ACM Computing Surveys (CSUR)*, pp. 1–35, 2021.
- [43] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.
- [44] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.
- [45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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*. Springer, 2014, pp. 740–755.
- [46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

- [47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [50] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.
- [51] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.



Fenglei Fan received the Ph.D. degree from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2021. He is currently an Assistant Professor with the Department of Data Science, City University of Hong Kong, Hong Kong, China, where he directs the Frontier of Artificial Networks Group. He was previously a postdoctoral researcher with Cornell University, NY, USA, and a Research Assistant Professor with The Chinese University of Hong Kong, Hong Kong, China. His research interests include NeuroAI, model compression, and medical imaging.



Shuren Qi received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2024. He is currently a Postdoctoral Fellow with the Department of Data Science, City University of Hong Kong, Hong Kong, China. He was previously a Postdoctoral Fellow with The Chinese University of Hong Kong, Hong Kong, China. His research interests include invariants, representations, and geometric deep learning.



Wenying Wen received the Ph.D. degree from Chongqing University, Chongqing, China, in 2013. She is currently a Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. Her research interests include privacy, security, and trustworthy AI.



Mengxin Shen received the M.S. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2026. Her research interests include media forensics and security.



Yushu Zhang (Senior Member, IEEE) received the Ph.D. degree from Chongqing University, Chongqing, China, in 2014. He is currently a Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include privacy, security, and trustworthy AI. He serves on the Editorial Boards of *IEEE Transactions on Dependable and Secure Computing*, *Signal Processing*, and *Information Sciences*.

APPENDIX A

ADDITIONAL QUALITATIVE RESULTS FOR IMPERCEPTIBILITY

To further illustrate the imperceptibility of the proposed method, we provide additional qualitative comparisons beyond Fig. 6 in the main text. Figs. A1–A3 show nine more examples generated under the same prompt and seed, together with the corresponding pixel-wise differences between the original and watermarked outputs. These examples consistently show that our method preserves stronger pixel-level consistency with the original generation result than the compared content-level baselines.

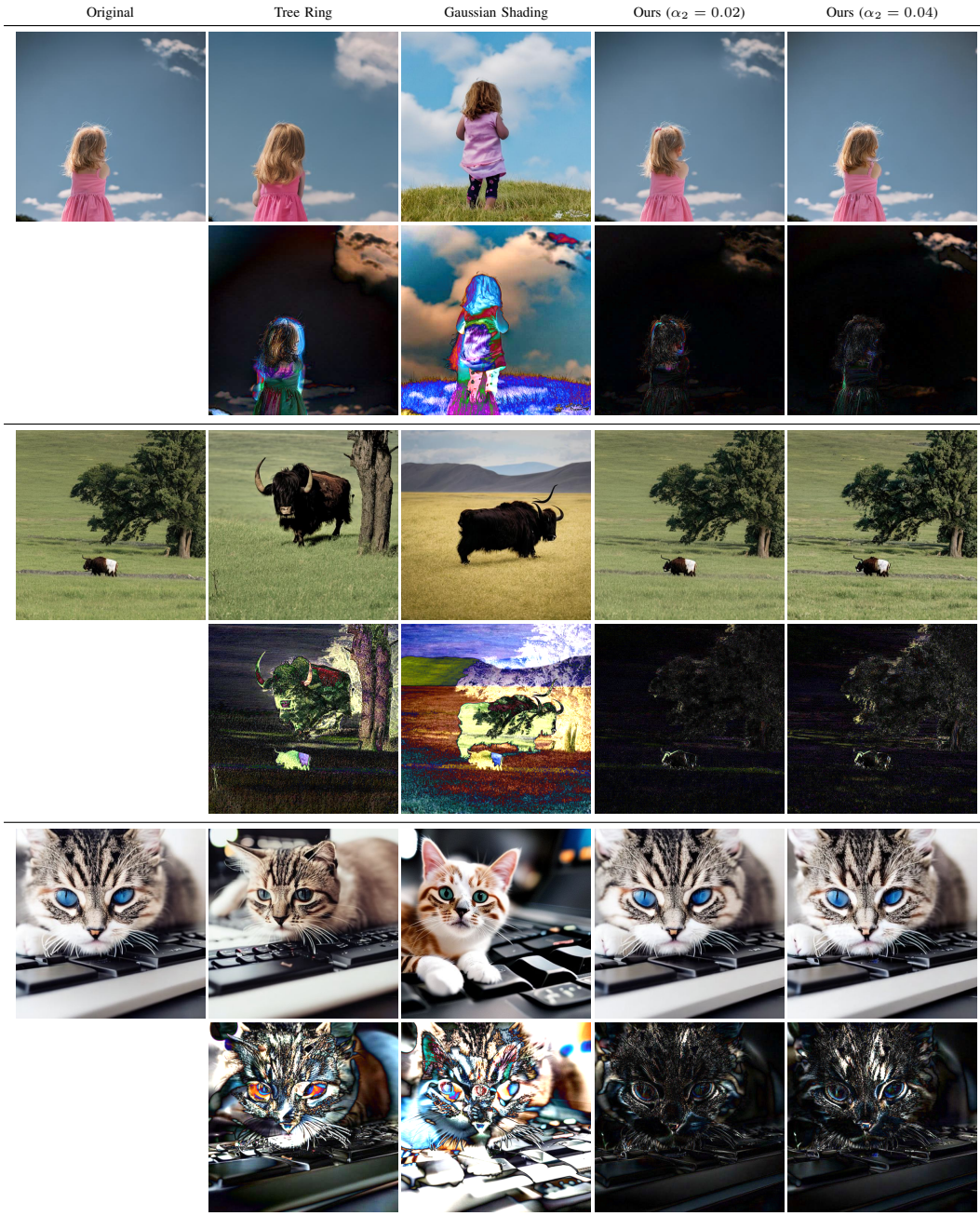


Fig. A1. Additional qualitative pairwise output-consistency results for the imperceptibility experiments. Each example shows the unwatermarked image, the corresponding watermarked images, and RGB difference maps computed with the same channelwise definition as Fig. 6. The prompts are: (a) A little girl looking down a hill with her skis. (b) A yak walking through a field toward a tree. (c) A kitty cat lying on a computer keyboard.

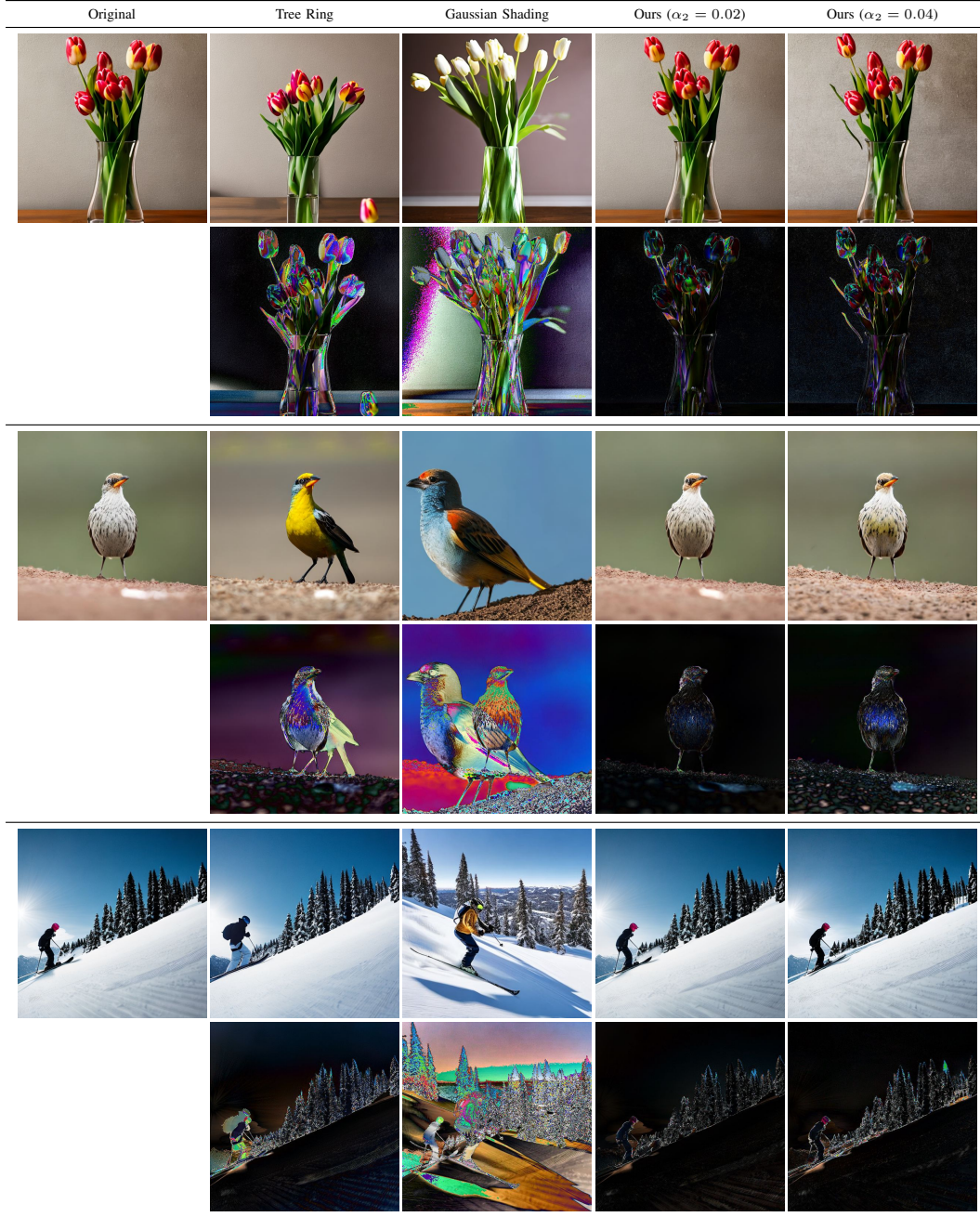


Fig. A2. Additional qualitative results for the imperceptibility experiments (continued). Each example shows the unwatermarked image, the corresponding watermarked images, and the associated pixel-wise difference visualizations. The prompts are: (d) A clear vase sitting on a table and holding long-stemmed tulips. (e) A bird standing on the ground and looking around. (f) A person skiing on a snowy hilltop.

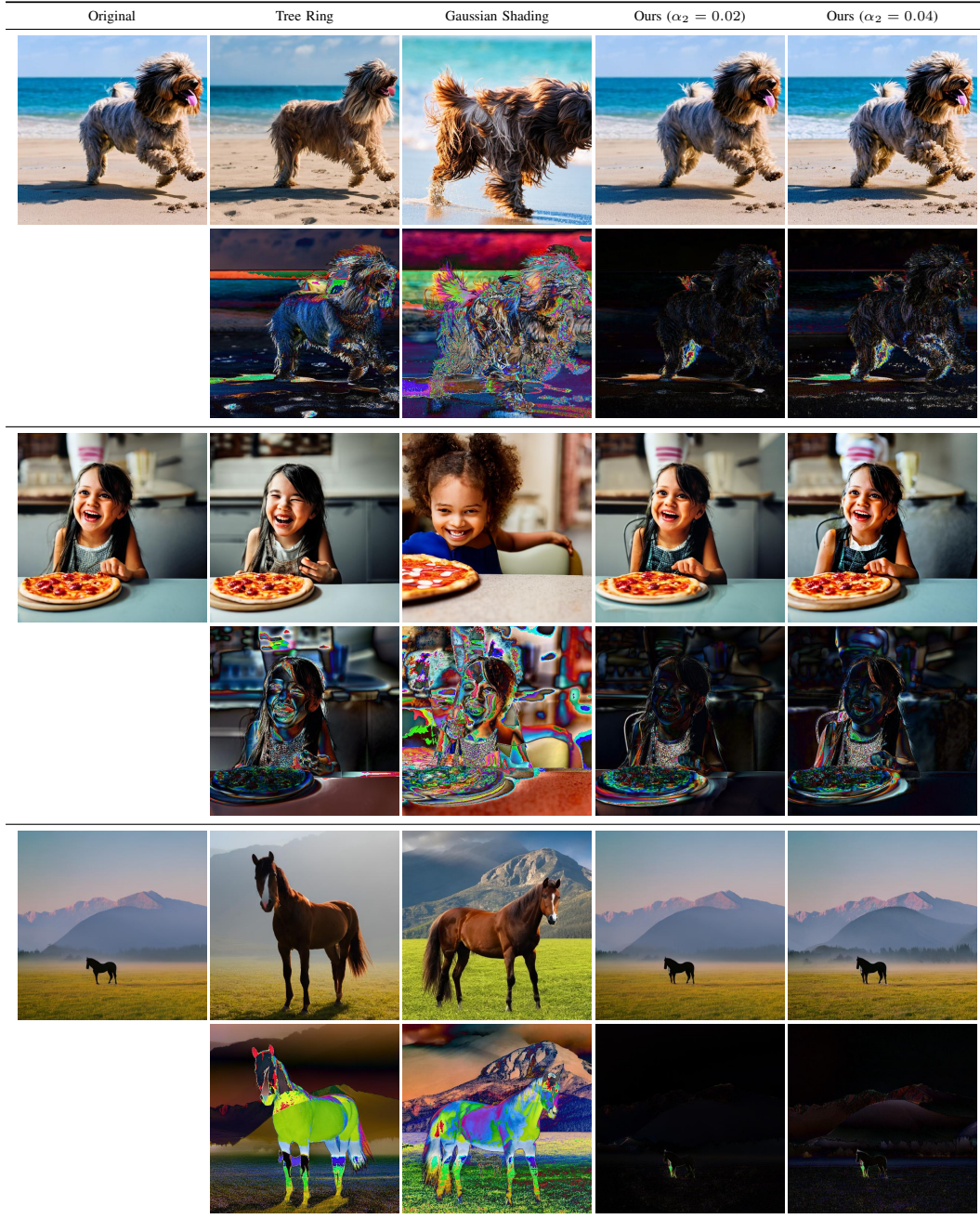


Fig. A3. Additional qualitative results for the imperceptibility experiments (continued). Each example shows the unwatermarked image, the corresponding watermarked images, and the associated pixel-wise difference visualizations. The prompts are: (g) A shaggy dog at the beach with a frisbee. (h) A little girl grinning at her plate of pizza. (i) A horse in a grassy field against a foggy mountain range.