

DeepFreqMark: End-To-End Learnable Frequency-Domain Watermarking with Spherical Attack Simulation for Latent Diffusion Models

Chen-Hsiu Huang*, Mario Köppen† and Ja-Ling Wu*

* National Taiwan University, Taipei, Taiwan

E-mail: {chenhsiu48,wjl}@cmlab.csie.ntu.edu.tw

† Kyushu Institute of Technology, Fukuoka, Japan

E-mail: mkoeppe@ieee.org

Abstract—The proliferation of AI-generated images produced by Latent Diffusion Models (LDMs) has raised critical concerns regarding copyright infringement and misinformation. Although existing frequency-domain watermarking methods embed handcrafted geometric patterns into the initial latent noise prior to generation, they suffer from limited capacity and rigid pattern designs. We propose DeepFreqMark, an end-to-end learnable frequency-domain watermarking framework that replaces manual pattern engineering with a neural message encoder and decoder. To circumvent the computational bottleneck caused by Denoising Diffusion Implicit Model (DDIM) inversion during training, we introduce a Spherical Linear Interpolation (Slerp)-based attack simulation. This approach operates directly on the noise latent while strictly preserving the Gaussian variance. Extensive experiments demonstrate that DeepFreqMark achieves significantly lower Bit Error Rates (BER) than baseline methods under real-world attacks and scales to 256 bits message capacity. Our source code is available at <https://github.com/chenhsiu48/DeepFreqMark>.

Index Terms—DeepFreqMark, Latent Diffusion Model, LDM Watermarking, Spherical Attack Simulation, Slerp.

I. INTRODUCTION

The rapid advancement of Latent Diffusion Models (LDMs) [1], such as Stable Diffusion [2], has democratized high-quality image synthesis. While beneficial to creative industries, the proliferation of AI-generated content raises critical concerns about copyright infringement and misinformation. Consequently, developing robust, invisible watermarking techniques to trace the provenance of generated images is paramount.

Hur et al. [3] categorize watermarking methods for diffusion models into two paradigms: “post-generation” and “in-generation.” Post-generation methods apply watermarks to the final generated image. However, they are vulnerable to bypass attacks, and optimizing them for high robustness often severely degrades visual quality. In contrast, in-generation methods embed watermarks directly into the generative process, offering enhanced resilience while preserving image quality.

Recent advances in LDM watermarking have shifted toward in-generation approaches. Methods such as Tree-Rings [4], METR [5], RingID [6], and HSTR/HSQR [7] embed watermarks in the initial Gaussian-noise latent prior to the reverse diffusion process. Because the Denoising Diffusion

Implicit Models (DDIMs) [8] inversion provides a deterministic mapping from the generated image back to the initial noise, these frequency-domain modifications appear as robust, invisible watermarks in the final pixel space.

Despite their success, existing frequency-domain watermarking techniques suffer from two critical limitations. First, they rely on handcrafted geometric patterns embedded exclusively in the Fast Fourier Transform (FFT) domain. This rigid design limits the maximum message capacity and prevents generalization to other domains, such as the Discrete Cosine Transform (DCT). Second, existing methods assume that mid-to-low-frequency modifications naturally survive image-level attacks. They do not optimize for these attacks during the embedding phase because of the computational cost of DDIM inversions.

We introduce DeepFreqMark to address these challenges. Our contributions are summarized as follows:

- **End-to-end Neural Embedding:** We propose a learnable framework that replaces rigid, handcrafted frequency patterns, extending semantic watermarking to both the DCT and FFT domains and scaling the payload capacity to 256 bits.
- **Slerp-based Attack Simulation:** We introduce a training mechanism that utilizes Spherical Linear Interpolation (Slerp) without DDIM inversion. By preserving the Gaussian variance of the noise latent, this simulation accurately approximates severe image-space distortions while avoiding the overhead of DDIM inversion.
- **Enhanced Robustness:** Extensive experiments demonstrate that DeepFreqMark achieves significantly lower Bit Error Rates (BER) than baseline methods across various real-world attacks, while preserving the visual fidelity and diversity of the generated images.

II. RELATED WORK

Most initial-noise LDM watermarking methods use the FFT domain for embedding. For instance, Wen et al. [4] proposed Tree-Rings, which embeds concentric ring patterns in the low-frequency regions of the FFT spectrum. However, its limited capacity restricts its utility for multi-user tracking. To address

this, Ci et al. [6] introduced RingID to enhance multi-key identification capabilities. Similarly, METR [5] increased the message capacity by assigning binary bits to individual rings, enabling the encoding of up to 2^{16} messages.

Despite these advancements, prior methods that used asymmetric frequency modifications discarded the imaginary parts during the inverse FFT, thereby compromising frequency-domain integrity. To mitigate this, HSTR/HSQR [7] introduced Hermitian symmetric patterns to enforce real-valued constraints.

III. METHODOLOGY

A. Neural Frequency Embedding

Our framework consists of a message encoder, $\mathcal{E}$, and a corresponding message decoder, $\mathcal{D}$. Given a binary message $m \in \{0,1\}^L$, where L denotes the message length in bits, the encoder $\mathcal{E}$ projects m into a continuous space of size $C \times D_{\text{wm}} \times D_{\text{wm}}$, where C is the channel dimension and $D_{\text{wm}} = 32$ is the spatial watermark dimension.

For the DCT Domain scheme in Fig. 1, the encoder outputs a single-channel, real-valued watermark Δ_{DCT} , which is injected directly into the DCT-transformed latent via:

$$Z'_T = Z_T \oplus \Delta_{\text{DCT}}, \quad (1)$$

where $Z_T = \text{DCT}(z_T)$ is the DCT-transformed initial noise latent, Z'_T denotes the watermarked noise latent in the frequency domain, and $\oplus$ signifies the element-wise addition applied to the upper-left, low-frequency region. This specific spectral band is selected to maximize robustness against common image-level attacks while minimizing distortion.

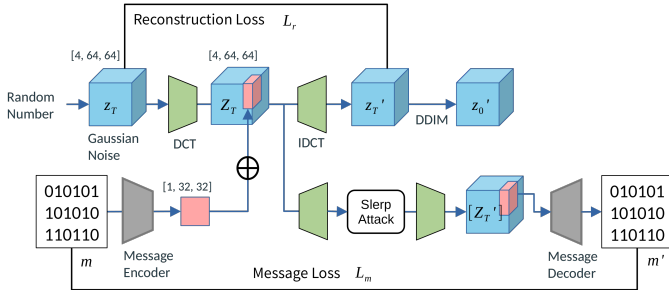


Fig. 1. DeepFreqMark embedding process in the DCT domain.

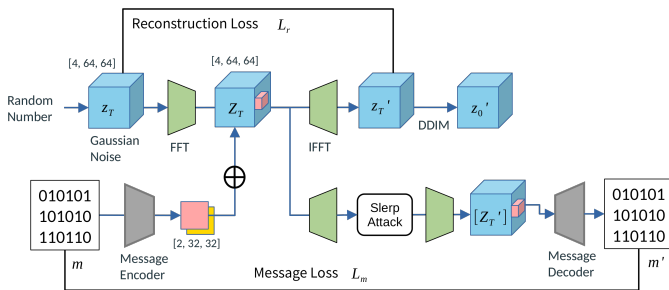


Fig. 2. DeepFreqMark embedding process in the FFT domain.

For the FFT Domain scheme in Fig. 2, the encoder outputs a two-channel, complex-valued watermark Δ_{FFT} . To satisfy the Hermitian Symmetry requirement for generating real-valued spatial latents, we designate a “free half-region” located strictly to the right of the vertical DC axis. We then mathematically reflect and complex-conjugate this region to construct its symmetric counterpart:

$$F[M - k, N - l] = \overline{F[k, l]}, \quad (2)$$

thereby guaranteeing that the inverse FFT yields a real-valued noise latent. During this process, the FFT-transformed noise latent is center-shifted, and the complex watermark is injected into the central frequency region. In both the DCT and FFT domains, the watermark is embedded in the last channel of the transformed noise latent Z_T .

B. Training Objective

To train the message encoder and decoder end-to-end, we optimize a joint loss function that balances watermark imperceptibility in the latent space with message-extraction accuracy.

Reconstruction Loss: To ensure that the watermarked latent z'_T remains statistically consistent with the initial Gaussian-noise latent $z_T \sim \mathcal{N}(0, \mathbf{I})$, we apply a Mean Squared Error (MSE) loss:

$$\mathcal{L}_r = \text{MSE}(z_T, z'_T). \quad (3)$$

Minimizing $\mathcal{L}_r$ guarantees that the frequency perturbation does not disrupt the variance or the generative stability of the downstream diffusion model.

Message Loss: To evaluate the accuracy of the extracted watermark, we employ the Binary Cross-Entropy (BCE) loss. The loss is computed between the original message m and the recovered message m' predicted by the decoder:

$$\mathcal{L}_m = \text{BCE}(m, m'). \quad (4)$$

The overall optimization objective is defined as a weighted sum of the reconstruction and message-extraction losses:

$$\mathcal{L} = \mathcal{L}_r + \alpha \mathcal{L}_m, \quad (5)$$

where α is a balancing hyperparameter that governs the trade-off between preserving latent fidelity and maximizing message-retrieval accuracy.

C. Watermark Encoder and Decoder Architecture

The message encoder and decoder networks are designed to learn an end-to-end continuous representation of a discrete binary message and its corresponding extraction logic within a localized $D_{\text{wm}} \times D_{\text{wm}}$ frequency patch. The detailed architectures of the encoder and decoder for both the DCT and FFT domains are illustrated in Figures 3 and 4, respectively.

Message Encoder ($\mathcal{E}$): Given an input binary message $m \in \{0,1\}^L$, the encoder first utilizes a fully connected linear layer to project the L -dimensional vector into a higher-dimensional

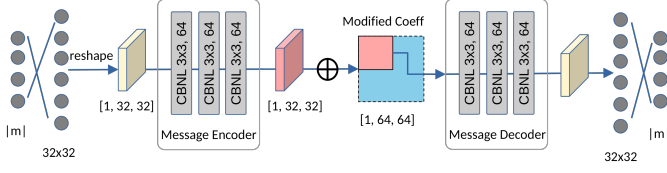


Fig. 3. Architectural overview of the message encoder and decoder for DCT domain embedding.

representation. This intermediate continuous vector is then reshaped into a two-dimensional feature map of size $1 \times 32 \times 32$.

The reshaped tensor is then processed through a sequence of convolutional blocks. Each standard block comprises a 3×3 2D convolutional layer with a stride of 1 and padding of 1, followed by 2D Batch Normalization (BN) to stabilize training and a LeakyReLU activation to introduce non-linearity. The final convolutional block omits the activation layer to allow an unconstrained continuous output range.

For the DCT framework, the final layer produces a real-valued watermark tensor $\Delta_{\text{DCT}} \in \mathbb{R}^{1 \times D_{\text{wm}} \times D_{\text{wm}}}$. Conversely, for the FFT framework, the output is a two-channel tensor $\Delta_{\text{FFT}} \in \mathbb{R}^{2 \times D_{\text{wm}} \times D_{\text{wm}}}$, representing the real and imaginary components of the complex watermark.

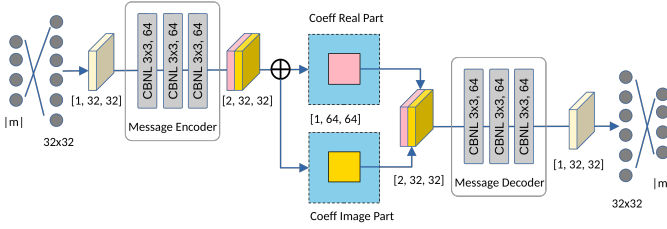


Fig. 4. Architectural overview of the message encoder and decoder for FFT domain embedding, utilizing a two-channel structure to separate real and imaginary components.

Message Decoder ($\mathcal{D}$): Let I_w denote the generated watermarked image and $\tilde{I}_w$ its attacked counterpart. The recovered noise latent, denoted as $[z_T]$, is obtained via DDIM inversion:

$$[z_T] = \text{DDIM}^{-1}(\tilde{I}_w). \quad (6)$$

During the message extraction phase, a localized frequency patch $\hat{\Delta}$ is isolated from the transformed attacked latent $[Z_T]$. For the DCT case, this is a single-channel patch $\hat{\Delta}_{\text{DCT}} \in \mathbb{R}^{1 \times D_{\text{wm}} \times D_{\text{wm}}}$ extracted from the upper-left, low-frequency region:

$$[Z_T] = \text{DCT}([z_T]), \quad (7)$$

$$\hat{\Delta}_{\text{DCT}} = \ominus([Z_T]), \quad (8)$$

where $\ominus(\cdot)$ denotes the frequency coefficient cropping operator. For the FFT case, the $\ominus$ operator isolates the designated centered-frequency region and separates the complex values into real and imaginary channels, forming a two-channel input $\hat{\Delta}_{\text{FFT}} \in \mathbb{R}^{2 \times D_{\text{wm}} \times D_{\text{wm}}}$.

The message decoder $\mathcal{D}$ mirrors the encoder's structure with a symmetric series of convolutional blocks to distill watermark features. The resulting feature map is flattened into a 1D vector and passed through a fully connected layer that maps the features back to the original message dimensionality L , thereby recovering the binary sequence:

$$m' = \mathcal{D}(\hat{\Delta}). \quad (9)$$

D. Spherical Attack Simulation (Slerp)

A core challenge in training robust watermarking frameworks for LDMs is the computational bottleneck of simulating real-world image distortions. Empirically, when standard image-level attacks (such as JPEG compression or Gaussian blur) are applied to a generated image, the subsequent DDIM inversion yields an attacked noise latent $[z_T]$ that exhibits a significant Mean Squared Error (MSE) relative to the original watermarked latent z_T , as reported in Table I. In this context, the cheng2020-anchor_3 and bmshj2018-factorized_3 attacks represent neural image re-compression [9] [10] at quality level 3. Concurrently, the diff_attacker_60 attack employs a diffusion-based purification model [11] designed to remove the watermark.

TABLE I
MSE OF DDIM INVERSION OF NOISE LATENTS UNDER VARIOUS IMAGE ATTACKS.

Attack Type	MSE Error	Attack Type	MSE Error
cheng2020-anchor_3	0.8077	brightness_0.5	0.1007
bmshj2018-factorized_3	0.8123	contrast_0.5	0.1230
diff_attacker_60	0.6286	Gaussian_noise	0.9091
jpeg_attacker_50	0.6010	Gaussian_blur	0.4024

However, computing the full DDIM inversion loop at every training iteration to optimize for robustness is computationally prohibitive. To bridge this gap efficiently, DeepFreqMark introduces a surrogate attack simulation that operates directly on the noise latent. According to the Gaussian Annulus Theorem, the probability mass of a high-dimensional Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ is heavily concentrated within a narrow spherical shell, effectively forming a high-dimensional “hollow soap bubble.” Consequently, simulating latent distortions via standard linear interpolation (Lerp) is mathematically flawed; it cuts through the interior of the hypersphere, artificially shrinking the variance and destroying the generative stability of the LDM. To resolve this, we employ Spherical Linear Interpolation (Slerp) to accurately traverse the surface of this hypersphere, simulating the trajectory of an attack via:

$$\text{Slerp}(z'_T, z_r, s) = \frac{\sin((1-s)\Omega)}{\sin(\Omega)} z'_T + \frac{\sin(s\Omega)}{\sin(\Omega)} z_r, \quad (10)$$

where z'_T represents the watermarked noise latent, and $z_r \sim \mathcal{N}(0, \mathbf{I})$ denotes a random Gaussian noise vector that serves as the target direction for the destructive attack energy. The scalar parameter Ω denotes the angular distance between z'_T

and z_r . To simulate varying degrees of distortion, the attack strength parameter s is uniformly sampled during training via $s \sim \mathcal{U}(0, S)$, where S defines the upper bound of the allowable attack strength.

Because Slerp operates strictly on the spherical manifold, it is mathematically guaranteed to preserve the unit variance and the underlying Gaussian profile of the noise. This surrogate allows us to explicitly optimize the encoder and decoder to withstand severe image-space distortions by proxying attacks directly on the spatial noise latent, thereby bypassing the computationally expensive overhead of DDIM inversion.

IV. EXPERIMENTS

A. Experimental Setup

We train separate DeepFreqMark models for both the DCT and FFT embedding domains, each with a message capacity of 32 bits. To train the encoder and decoder networks, we randomly sample initial noise latents $z_T \sim \mathcal{N}(0, \mathbf{I})$ in batches of 100. The training procedure spans 300 epochs, with each epoch consisting of 5,000 steps. For each latent, we generate a random binary message and apply the corresponding frequency-domain embedding. Throughout training, we apply a Slerp-based attack simulation with a maximum strength of $S = 0.6$.

To evaluate the robustness of the trained models, we test DeepFreqMark with Stable Diffusion v1.5 [1]. We randomly sample 100 text prompts from the DiffusionDB dataset [12]. For each prompt, we generate a watermarked image using the trained encoder, then apply a suite of common image attacks, as listed in Table I, to simulate real-world distortions. Finally, we perform DDIM inversion on the attacked images to recover the noise latent and extract the message to compute the Bit Error Rate (BER).

B. Robustness

Table II presents the BER of the extracted messages under distinct attacks for both the DCT and FFT embedding methods, comparing models trained without ($S = 0$) and with ($S = 0.6$) the Slerp-based attack simulation. The empirical results indicate that the Slerp simulation substantially improves robustness, yielding a significant reduction in average BER across all attacks.

TABLE II
COMPARISON OF BER ACROSS DCT AND FFT METHODS

	DCT, S=0	DCT, S=0.6	FFT, S=0	FFT, S=0.6
Un-attacked	0.1250%	0.0000%	0.0938%	0.0000%
cheng2020-anchor_3	2.2500%	0.9375%	3.9063%	0.8750%
bmsjh2018-factorized_3	1.8750%	0.7813%	3.6875%	0.5625%
diff_attacker_60	2.4688%	1.0000%	3.0938%	0.3750%
jpeg_attacker_50	0.6250%	0.3438%	1.0000%	0.0313%
brightness_0.5	0.1875%	0.0625%	0.0625%	0.0000%
contrast_0.5	0.1563%	0.1250%	0.1250%	0.0000%
Gaussian_noise	2.0313%	0.9375%	5.3438%	0.9688%
Gaussian_blur	0.2500%	0.0625%	0.4688%	0.0313%
Average	1.1076%	0.4722%	1.9757%	0.3160%

Table III compares DeepFreqMark with existing hand-crafted, FFT-based watermarking methods in terms of message reconstruction accuracy and payload capacity, using performance metrics reported directly in the original publications. Unlike these baseline methods, which rely on a two-stage “verification-then-identification” pipeline, DeepFreqMark directly decodes the precise message bits from the transformed noise latents. Crucially, our learnable framework achieves significantly higher bit-extraction accuracy and a greater embedding capacity.

TABLE III
COMPARISON OF ACCURACY AND CAPACITY ACROSS DIFFERENT WATERMARKING METHODS.

Method	Accuracy	Capacity
Ours (FFT, S=0.6)	99.68%	32 bits
Ours (FFT, S=0.6)	99.44%	256 bits
Tree-Rings [4]*	99.50%	0 bits
Tree-Rings [4]	7.70%	11 bits
RingID [6]	94.20%	11 bits
METR [5]	85.45%	16 bits
HSTR [7]*	88.90%	0 bits
HSQR [7]	98.50%	72 bits

* Verification accuracy.

C. Qualitative Results

The qualitative results in Fig. 5 further illustrate the visual impact of training with and without the Slerp attack simulation. The figure provides a side-by-side comparison of the original unwatermarked images and the watermarked outputs from the DCT and FFT models. As observed, models trained without attack simulation ($S = 0$) produce images that are slightly more coherent with the unwatermarked baseline, but their embedded signals are more vulnerable to attacks. Empirically, setting the attack strength to $S = 0.6$ yields images that maintain high semantic fidelity to the original prompt while improving watermark resilience.

Furthermore, Fig. 6 illustrates the visual diversity of images generated from the same prompt under different watermark messages. This diversity stems from the learnable nature of our neural message encoder, which optimizes spectral modifications to balance imperceptibility and robustness. Despite variations in the underlying messages, the generated images remain highly coherent with the target prompt. This confirms that the framework does not converge to a rigid, fixed embedding pattern, thereby enabling a vast and unique message space for provenance tracking.

D. Attack Strength

To investigate the impact of attack strength in the Slerp simulation, we evaluate the BER across a range of maximum attack strengths, $S \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$. As illustrated in Fig. 7, increasing the maximum attack strength S generally leads to a reduction in BER. Specifically, the configuration with $S = 0.9$ achieves the lowest BER.

However, this improvement in robustness comes at a cost. As shown in Fig. 8, the visual quality of the generated images

below 1.5% for both the DCT and FFT models for payloads up to 128 bits. In practical scenarios, such low error rates can be seamlessly corrected using standard error-correcting codes, such as Bose-Chaudhuri-Hocquenghem (BCH) codes.

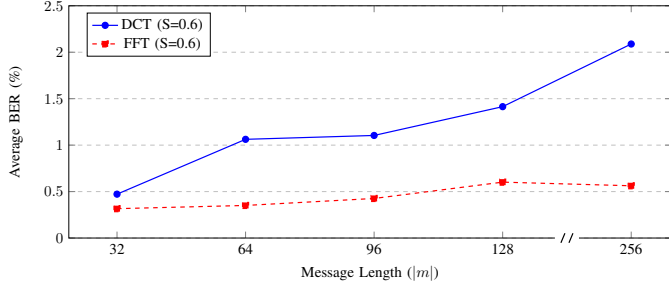


Fig. 9. Average BER vs. message length (at $S = 0.6$). The gap between $|m| = 128$ and $|m| = 256$ is shortened for better visualization.

F. Superiority of FFT over DCT

Comparing the DCT and FFT embedding methods in Figures 7 and 9, we observe that the FFT-based DeepFreqMark variants consistently achieve a lower BER across varying attack strengths and message lengths.

The superior robustness of the FFT-based approach stems from its complex-valued embedding space. Unlike real-valued DCT coefficients, FFT provides the encoder with magnitude and phase components, granting greater degrees of freedom. Furthermore, the mandatory Hermitian Symmetry required for real-valued latents introduces structural redundancy in the FFT spectrum, granting the decoder natural resilience against magnitude-distorting and quantization attacks.

V. CONCLUSION

We presented DeepFreqMark, a novel end-to-end neural frequency-domain watermarking framework designed to trace the provenance of images generated by LDMs. By replacing rigid, handcrafted geometric patterns with a learnable embedding strategy, our method successfully generalizes across both the DCT and FFT domains and scales the payload capacity to 256 bits. A core innovation of our approach is the Slerp-based spherical attack simulation, which preserves the Gaussian distribution of the noise latent while efficiently proxying real-world image distortions. Extensive experiments demonstrate the effectiveness and scalability of our framework. Consequently, DeepFreqMark provides a highly resilient, scalable, and visually coherent solution for generative content attribution. Our source code is available at <https://github.com/chenhsu48/DeepFreqMark>.

ACKNOWLEDGMENTS

The author gratefully acknowledges support from the National Science and Technology Council (NSTC) of Taiwan, which sponsored this joint research with Dr. Köppen during his visiting research stay at Kyushu Institute of Technology under Grant No. 114-2917-I-002-022.

REFERENCES

- [1] 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. 10 684–10 695.
- [2] D. Podell et al., “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [3] H. Hur, M. Kang, S. Seo, and J.-U. Hou, “Latent diffusion models for image watermarking: A review of recent trends and future directions,” *Electronics*, vol. 14, no. 1, p. 25, 2024.
- [4] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, “Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust,” *arXiv preprint arXiv:2305.20030*, 2023.
- [5] A. Varlamov, D. Diatlova, and E. Spirin, “Metri: Image watermarking with large number of unique messages,” *arXiv preprint arXiv:2408.08340*, 2024.
- [6] H. Ci, P. Yang, Y. Song, and M. Z. Shou, “Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification,” in *European conference on computer vision*, Springer, 2024, pp. 338–354.
- [7] S. J. Lee and N. I. Cho, “Semantic watermarking reinvented: Enhancing robustness and generation quality with fourier integrity,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 18 759–18 769.
- [8] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *International Conference on Learning Representations*, 2021.
- [9] Z. Cheng, H. Sun, Y. Tian, and A. Vetro, “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.
- [10] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” *arXiv preprint arXiv:1802.01436*, 2018.
- [11] X. Zhao et al., “Invisible image watermarks are provably removable using generative ai,” *Advances in neural information processing systems*, vol. 37, pp. 8643–8672, 2024.
- [12] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, “DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models,” *arXiv:2210.14896 [cs]*, 2022. [Online]. Available: <https://arxiv.org/abs/2210.14896>