Visual Quality Improved Watermarking based on Dual-Reference Loss for Deepfake Attribution

Qiushi Li
Media Integration and Communication Center (MICC), University of Florence
Florence, Italy
qiushi.li@unifi.it

Stefano Berretti
Media Integration and Communication Center (MICC), University of Florence
Florence, Italy
stefano.berretti@unifi.it

Roberto Caldelli
National Interuniversity Consortium for Telecommunications (CNIT), Florence, Italy
Universitas Mercatorum, Rome, Italy
roberto.caldelli@unifi.it

Abstract

The rapid advancement of image generation models like Stable Diffusion raises concerns about potential misuse, making robust watermarking techniques essential for the authentication and attribution of synthetic content, particularly in combating deepfakes. However, simultaneously ensuring high-quality image generation and accurate watermark extraction remains challenging. Through an analysis of existing methods, we identify a critical limitation: their loss functions often adopt a *single reference* (either the input image or the clean-generated image) for optimizing image fidelity, leading to suboptimal performance. In this paper, we conduct an in-depth study of the image-quality loss term in diffusion-based watermarking. By analyzing the distinct impacts of using the input image versus the clean-generated image as references during optimization, we reveal that jointly considering both references significantly improves robustness and visual quality. Extensive experiments demonstrate that our *dual-reference* approach achieves superior performance in both watermark extraction accuracy and generation fidelity compared to single-reference baselines. We advocate for this paradigm to advance reliable watermarking in generative models.

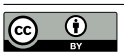
CCS Concepts

• **Security and privacy** → **Digital rights management**; **Privacy protections**.

Keywords

Intellectual Property Protection, Stable Diffusion, DNN Watermarking, Deepfake Attribution

ACM Reference Format:
Qiushi Li, Stefano Berretti, and Roberto Caldelli. 2025. Visual Quality Improved Watermarking based on Dual-Reference Loss for Deepfake Attribution. In *Proceedings of the 1st Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media (DFF '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746265.3759660



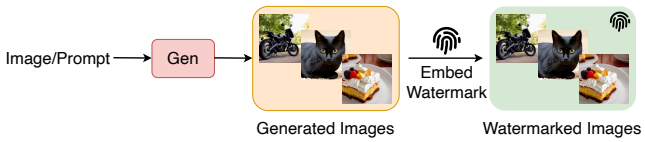
DFF '25, Dublin, Ireland

ACM ISBN 979-8-4007-2047-5/2025/10
https://doi.org/10.1145/3746265.3759660

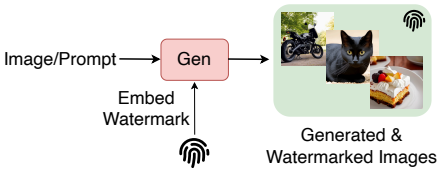
1 Introduction

Nowadays, AI-Generated Content (AIGC) has emerged as a highly popular and rapidly evolving technology and research field. With the advancement of generative techniques, especially Diffusion Models [6, 10, 11], their capabilities, such as producing photorealistic images and streamlining content creation, are progressively reshaping people's lifestyles. However, alongside these advancements, significant challenges arise from the potential misuse of generated images, which poses serious societal risks that demand continuous attention. In this context, proactive measures, particularly watermarking techniques tailored to generative models, have become a crucial solution for protecting the model intellectual property and ensuring the traceability of both models and their generated content (*attribution*).

Current watermarking-based deepfake attribution methods for Latent Diffusion Models (LDMs) can be broadly categorized into two approaches: *post-generation watermarking* methods [1, 12, 17], and *joint generation-embedding* methods [2–4, 9, 13, 14, 16]. Post-generation watermarking methods can embed and extract watermarks directly from images generated by LDMs (see Fig. 1a). While simple and practical, these methods require additional processing of the output images and remain vulnerable to information tampering and evasion attacks. Joint generation-embedding methods, instead, integrate watermarking during the image generation process, significantly reducing visible artifacts (see Fig. 1b). These approaches typically utilize the generated output images as feedback to appropriately fine-tune part or all of the generative model's weights, which ensures both successful watermark extraction and minimal impact on image quality. Currently, the joint generation-embedding approach has gained increasing attention in the field, as it inherently combines image generation and watermarking into a unified framework. However, existing joint generation-embedding approaches often employ limited-scope image quality supervision strategies by relying on *single-reference* loss functions, i.e., only one image is used as a reference during loss computation. These methods typically adopt either (1) the input-image as a reference (e.g., WOUAF [9]), where the loss function enforces content/distribution alignment between input and generated images during training, or (2) the clean-generated-image (i.e., the non-watermarked generated image) as a reference (e.g., Stable Signature [4] and AquaLoRA [3]), which directly minimizes the watermark-related distortion introduced to the generated images. While the former approach dominates most generative watermarking works for maintaining content consistency, the latter becomes crucial when watermark encoding



(a) Pipeline of post-generation watermarking.



(b) Pipeline of joint generation-embedding.

Figure 1: The two main generative watermarking pipelines for generated image attribution.

affects critical components (noise vectors, U-Net, or text-encoder) that may otherwise cause content deviation between input and output.

In this work, we mainly focus on a three-party scenario involving:

(1) *The model owner*, who develops a high-quality generative model and uploads it to a third-party platform.
(2) *The third-party platform*, which hosts the model, embeds watermarks into generated images, and decodes them for traceability, before distributing the model to end users.
(3) *The end user*, who accesses the model for specific applications.

To address the limitations of existing loss designs in such scenarios, we conduct an analysis of image-quality supervision signals and their impact on model performance. Our key contributions include:

- We identify the limitations of single-reference loss functions in existing methods, leading to suboptimal performance.
- We propose a *dual-reference loss function* that jointly considers both the input image and the clean-generated image as references during optimization.
- We demonstrate the effectiveness of our dual-reference approach through extensive experiments, achieving superior performance in both watermark extraction accuracy and generation fidelity compared to single-reference baselines.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related works. Section 3 presents our proposed approach in detail. Comprehensive experimental results are reported and analyzed in Section 4. Finally, the paper is concluded in Section 5, where future research directions are also outlined.

2 Related Work

The advancement of LDMs has significantly enhanced the quality and efficiency of content creation and editing. Text-to-image (T2I) generation models, in particular, have demonstrated remarkable progress in both output fidelity and content diversity, enabling broader real-world applications. However, critical challenges remain in preventing model misuse, ensuring copyright protection, and attributing generated content to combat disinformation. Watermarking techniques for LDMs have thus emerged as a pivotal solution, attracting growing research attention.

Existing LDM-based watermarking methods mainly fall into two categories: post-embedding watermarking, and joint generation-embedding methods, as shown in Fig. 1.

The former solution represents traditional image-oriented watermarking, which directly embeds and extracts watermarks from generated outputs, such as classical frequency-domain methods DCT/DWT-based techniques [1], and deep learning methods like HiDDeN [17] and StegaStamp [12]. Because they directly operate on images, they are model-agnostic watermarking methods. While simple to implement, such methods require additional processing and remain vulnerable to post-processing attacks.

Recent research has focused on integrating watermark embedding into the image generation process, injecting watermarks into model parameters, while fine-tuning the generative model. Similar to the two-stage watermarking pipeline proposed by Yu et al. [14], Ditria and Drummond [2], and Zhao et al. [16], the watermark information is indirectly embedded into generated images by training LDMs with watermarked training datasets. The drawback of this approach is that it requires modifying training data and retraining the generative model for each new user deployment. Fernandez et al. [4] proposed a new watermark embedding method for LDMs called Stable Signature. This method uses a pretrained HiDDeN as the watermark extractor and employs the fixed watermark extractor to guide the fine-tuning of an LDM's decoder to generate images containing the corresponding watermark. Wen et al. [13] proposed Tree-ring watermarks, embedding watermarks into the initial noise vector of diffusion models to align with specific patterns. Feng et al. [3] proposed AquaLoRA, which utilizes LoRA modules [7] to embed watermarks in the UNet of an LDM. However, these methods still require retraining or fine-tuning the generative model when distributing it to new users, which is highly inefficient in practical applications. To address this issue, Kim et al. [9] introduced a weight modulation method [8], called WOUAF, to integrate watermark information into model parameters of Stable Diffusion without retraining.

While existing methods can successfully embed watermarks into LDMs, their loss functions typically employ a single reference image (either the input image or the clean generated image) to optimize image fidelity, often resulting in suboptimal generation quality. As summarized in Table 1, current LDM-based watermarking approaches predominantly focus on one reference image type. Building upon WOUAF [9], our work introduces a dual-reference supervision to simultaneously account for both input and clean generated images during optimization, thereby achieving a better balance between watermark extraction accuracy and generation quality.

3 Proposed Method

This section presents our dual-reference quality preservation framework, with the complete architecture illustrated in Fig. 2. We first

Table 1: Overview of reference-type configurations in LDM-based watermarking methods. The watermark embedding position is also reported.

Method	Embedding position	Input reference	Generated reference
Stable Signature [4]	VAE-Decoder	✘	✔
AquaLoRA [3]	UNet	✘	✔
WOUAF [9]	VAE-Decoder	✔	✘
Ours	VAE-Decoder	✔	✔

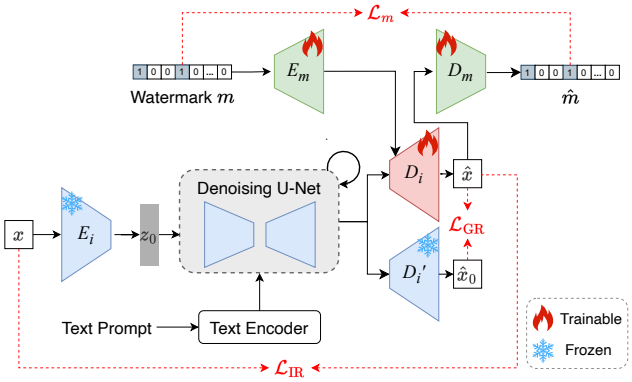


Figure 2: Illustration of our proposed method.

formalize the problem definition, followed by detailed technical descriptions of each component of our proposed framework.

3.1 Problem Definition

To address the generative image watermarking task, the solution usually consists of two key components: (1) the watermarking model for accurate watermark extraction, and (2) the generation model (specifically based on Stable Diffusion in this work) that maintains image quality during watermark embedding. Accordingly, the overall training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_i, \tag{1}$$

where $\mathcal{L}_m$ represents the watermark extraction loss, and $\mathcal{L}_i$ denotes the image quality preservation loss. In our proposed method, $\mathcal{L}_i$ consists of two components: L_{GR} and L_{IR} (see Section 3.3 for details).

Given the watermark decoder D_m, which extracts the predicted watermark $\hat{m}$ from watermarked generated images $\hat{x}$, the objective function aims to minimize the loss between the extracted watermark $\hat{m}$ and the ground truth watermark m. This is typically implemented using the Binary Cross-Entropy (BCE) loss with sigmoid activation, formulated as:

$$\mathcal{L}_m = -\sum_{i=1}^{d_m} \left[m_i \log \sigma(\hat{m}_i) + (1 - m_i) \log(1 - \sigma(\hat{m}_i)) \right], \tag{2}$$

where d_m is the watermark length, and $\sigma(\cdot)$ refers to the sigmoid function. The image quality preservation loss $\mathcal{L}_i$ will be detailed in Section 3.3, where we introduce our dual-reference quality preservation loss.

3.2 Framework Overview

Figure 2 illustrates the overall architecture of our proposed framework, including the Stable Diffusion model, the watermark encoder and decoder. The Stable Diffusion architecture contains several key components: a variational autoencoder (VAE) for latent space representation, a U-Net based denoising network, a CLIP text encoder for conditional generation, and a noise scheduler that controls the diffusion process. During the training stage, the VAE encoder E_i encodes the input image x into a latent representation z_0, which then undergoes gradual Gaussian noise addition over multiple steps (forward process). A text-conditioned U-Net iteratively predicts and subtracts the noise at each step (reverse process), ultimately reconstructing a denoised latent representation. The final denoised latent representation is decoded back to the pixel space by the VAE decoder D_i, generating the output image.

Inspired by recent advances in style-based generation [8] and model watermarking [9], we introduce a weight modulation strategy to integrate watermark into the VAE decoder D_i. Specifically, we apply learnable affine layers to modulate both convolutional and attention layers in D_i, enabling effective watermark embeddings, while maintaining the model's generative capabilities. For a given weight tensor $W^l \in \mathbb{R}^{i \times j \times k}$ at layer l (where i, j, k represent input, output, and kernel dimensions respectively), the weight modulation operation is formulated as:

$$\begin{aligned} \widehat{W}^l_{i,j,k} &= u_i \cdot W^l_{i,j,k} \\ &= A_l(E_m(m)) \cdot W^l_{i,j,k}, \end{aligned} \tag{3}$$

where $A_l(\cdot)$ represents the affine transformation layer for layer l, $E_m(m) \in \mathbb{R}^d$ denotes the d-dimensional watermark embedding for message m, and $u_i \in \mathbb{R}^i$ is the channel-wise modulation vector.

In the context of deepfake generation and attribution, ensuring the traceability and accountability of synthetic content is critical, especially as generative models become increasingly accessible and powerful. Deepfake attribution techniques aim to link generated media back to their source models, which is essential for detecting misuse, verifying authenticity, and enforcing the provenance of the content.

During the model distribution phase, i.e., the phase in which pretrained generative models are distributed to end users, our framework enables direct watermark embedding into the generative model's parameter space without additional fine-tuning or retraining of the watermarking components. This design makes our approach particularly suitable for practical deployment, such as tagging models distributed to different users or institutions for downstream deepfake generation tasks.

3.3 Training Objectives

As discussed in Section 3.1, our framework must simultaneously optimize two competing objectives: watermark extraction accuracy and image quality preservation. To address this, we propose a dual-reference quality preservation loss that leverages both the input image and the clean-generated image as references, each contributing distinct advantages:

(1) **Input image reference**: Ensures watermarked outputs maintain distributional consistency with natural images. For conditional generation models like Stable Diffusion, where generated content may significantly diverge from input images, we employ the Learned Perceptual Image Patch Similarity (LPIPS) [15] metric rather than pixel-wise measures. This perceptual loss better preserves natural image statistics and human visual fidelity.
(2) **Clean-generated reference**: Maintains functional equivalence between watermarked and original generation model outputs. This term minimizes distortions introduced by watermark embedding, preserving the generative model's core capabilities. As illustrated in Fig. 2, during the training phase, our framework incorporates two parallel branches: (a) The trainable watermarked VAE decoder D_i (with watermark embedding); (b) A reference branch containing a fixed copy D_i' of the original generative model's VAE decoder. The reference branch remains completely frozen throughout training, serving exclusively to generate clean (non-watermarked) reference images $\hat{x}_0$. For quality evaluation against this reference, we employ the LPIPS metric to ensure the watermarked outputs maintain perceptual fidelity with the model's original clean generations.

The frozen D_i' branch ensures stable training by providing consistent clean references, while LPIPS captures human-aligned quality metrics beyond pixel-space losses. Both reference paths use LPIPS for quality assessment, yielding the composite loss:

$$\mathcal{L}_i = \mathcal{L}_{\text{IR}}(x, \hat{x}) + \mathcal{L}_{\text{GR}}(\hat{x}_0, \hat{x}), \tag{4}$$

where $\mathcal{L}_{\text{IR}}$ and $\mathcal{L}_{\text{GR}}$ are the LPIPS losses for input and clean-generated references, respectively.

In the subsequent experiments (Section 4), we systematically evaluate the performance of our framework under two distinct configurations of the image quality loss weight (λ_i) in Eq.(5): $\lambda_i = 1.0$ and $\lambda_i = 0.5$. This formulation represents the comprehensive loss function that combines both watermark extraction and image quality objectives:

$$\mathcal{L} = \mathcal{L}_m + \lambda_i \cdot \mathcal{L}_i. \tag{5}$$

The configuration with $\lambda_i = 1.0$ directly integrates the reference image through linear combination in the loss function, placing maximal emphasis on visual fidelity preservation. In contrast, the configuration with $\lambda_i = 0.5$ establishes an optimized equilibrium between image quality preservation and watermark extraction. This balanced approach carefully modulates the relative contributions of these competing objectives during the optimization process.

To provide a clearer and more intuitive understanding, Fig. 3 presents two representative examples. The columns, from left to right, show the prompts from COCO2014, the corresponding input reference images, the images generated by Stable Diffusion, and the watermarked generated results. Visually, it is evident that the content of the generated images with and without watermarks remains largely consistent. A straightforward strategy to reduce watermark-induced distortion is to minimize the perceptual distance between the original and watermarked images. Meanwhile, the input reference image serves as a constraint to ensure that the watermarked outputs remain within the natural image distribution.

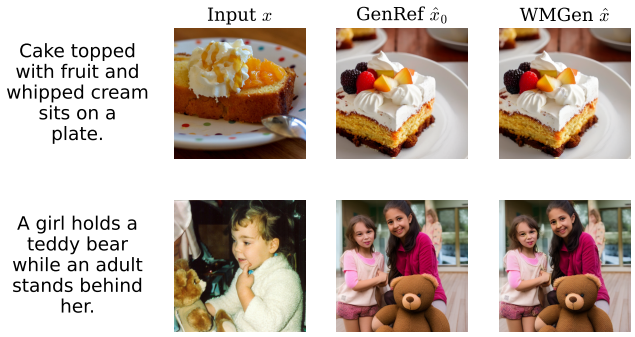


Figure 3: Examples of reference-guided generation and watermark embedding. From left to right: the input text prompt and reference image (in column 1 and 2, respectively), the corresponding generated image (GenRef), and the watermarked generated image (WMGen).

4 Experiments

4.1 Experimental Setup

Datasets: We employ the MS-COCO2014 dataset with Karpathy splits[1] for training, and use the test set, composed of 5,000 samples, for evaluation. To comprehensively evaluate our watermarking method across diverse scenarios, we conducted experiments at multiple image resolutions (256×256 and 512×512) with varying watermark lengths:

- For 256×256 images, we evaluated watermark lengths of 16, 32, and 48 bits to assess performance.
- For 512×512 images, we adopted larger payloads (32 and 64 bits) to leverage their increased spatial dimensions and finer details.

During evaluation, unless otherwise specified, we embed randomly generated watermarks $m \sim U(0,1)^{d_m}$ (d_m represents the watermark length) for each test sample to comprehensively assess model performance under real-world conditions. This approach rigorously tests three critical aspects: (1) encoding reliability through consistent extraction accuracy across diverse watermark patterns, (2) generation stability by maintaining image quality independent of embedded watermarks, and (3) generalization capability against watermark-distribution shifts.

Implementation details: As mentioned at the end of Section 2, our work builds upon WOUAF [9] as the baseline watermarking framework. Specifically, we employ Stable Diffusion v2-base[2] as the backbone generative model. During training, the VAE decoder, watermark encoder and decoder are jointly trained for 25 epochs with a batch size of 16. For optimization, we used AdamW with an initial learning rate of 1×10^{-4} and cosine annealing scheduling. All watermarks are randomly generated binary sequences for both training and evaluation. During T2I generation, we used a classifier-free guidance scale of 7.5 and the Euler scheduler with 20 sampling steps by default to ensure consistent generation quality. For fair comparison, all watermarking solutions are trained and evaluated

[1] https://huggingface.co/datasets/HuggingFaceM4/COCO
[2] https://huggingface.co/stabilityai/stable-diffusion-2-base

under identical configurations (dataset, hardware and evaluation protocols).

Evaluation metrics: To evaluate the proposed watermarking framework, we employ a comprehensive set of metrics assessing both watermark extraction accuracy and image generation quality. For watermark extraction performance, we use the Bit Accuracy (BitAcc), calculated as the percentage of correctly decoded watermark bits against the ground truth embedded sequence. In terms of image quality assessment, we measure the image similarity between watermarked and unmarked generated images using established metrics including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) [15] and Fréchet Inception Distance (FID) [5] to quantify the visual impact of watermark embedding. This multi-faceted evaluation protocol ensures rigorous assessment of both the watermark's robustness and its impact on the generative model's functionality.

4.2 Performance of the Dual-Reference Strategy

In these experiments, we demonstrate the performance of the introduced watermarking method under different training objectives, including: using only the input image as reference (IR, equivalent to WOUAF [9]), using only the clean-generated image as reference (GR), and our proposed dual reference methods with $\lambda_i = 1.0$ (DR-1.0) and $\lambda_i = 0.5$ (DR-0.5). The experimental results are shown in Table 2, which compares the performances under different training objectives across various scenarios, including image resolutions and watermark lengths.

For experiments at 256×256 resolution, DR-1.0 achieved the best image quality performance across all scenarios, particularly with 16-bit watermarks. This is because the training process assigned a higher weight to the image quality objective. However, this emphasis also affected DR-1.0's watermark extraction capability and learning difficulty, resulting in inferior extraction performance compared to GR and DR-0.5. In contrast, DR-0.5 achieved a better balance between watermark extraction and image quality, especially with 32-bit watermarks where it reached 0.984 BitAcc, while maintaining high image quality metrics compared to other single-reference solutions.

For experiments at 512×512 resolution, DR-1.0 expectedly achieved the best image quality results. The results also demonstrate that longer watermarks (64-bit) present greater embedding challenges, leading to lower BitAcc. DR-0.5 maintained a good balance, achieving 0.994 BitAcc with 32-bit watermarks while maintaining competitive image quality. Although IR achieved the highest BitAcc with 64-bit watermarks, it suffered greater losses in image quality, underperforming both DR-1.0 and DR-0.5. This indicates that loss functions using only input reference have greater impact on image quality during optimization.

It should be noted that since the base Stable Diffusion v2 model was pretrained at 512×512 resolution, its performance at this resolution generally surpasses that at 256×256 resolution. Particularly for image quality preservation, the results of 512×512 achieved near-ideal PSNR values around 30 dB, making the watermark-induced distortions perceptually negligible, as shown in Fig. 4(a). Furthermore, the higher resolution provides a larger spatial capacity for embedding watermark information, which further contributes to the generation of high-quality, visually consistent images.

Table 2: Comparison of watermarking performance for various length d_m of the watermark, and different WOUAF [9] configurations. IR: input-reference (WOUAF baseline); GR: generated-reference only; DR-1.0 / DR-0.5: dual-reference method with $\lambda_i = 1.0$ / 0.5.

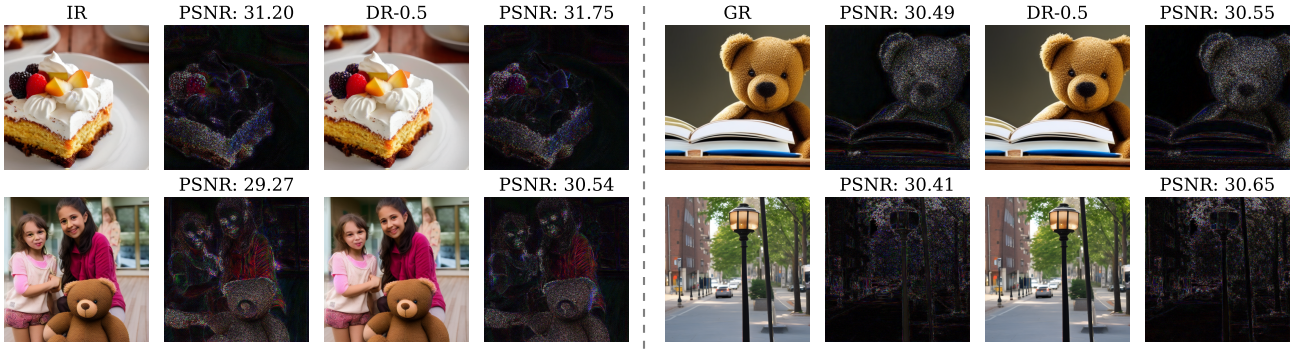
Resolution of 256×256						
d_m	Method	BitAcc↑	FID↓	LPIPS↓	PSNR↑	SSIM↑
16	IR [9]	0.934	6.693	0.099	26.831	0.829
	GR	**0.989**	5.758	0.085	27.375	0.854
	DR-1.0	0.884	**4.987**	**0.075**	**28.107**	**0.861**
	DR-0.5	0.985	6.306	0.093	27.029	0.843
32	IR [9]	0.977	7.596	0.114	25.535	0.795
	GR	0.975	7.459	0.102	25.913	0.820
	DR-1.0	0.795	**5.647**	**0.084**	**27.184**	**0.832**
	DR-0.5	**0.984**	7.631	0.106	25.972	0.817
48	IR [9]	0.801	7.271	0.108	25.771	0.797
	GR	0.871	6.030	0.086	26.780	0.833
	DR-1.0	0.840	**4.738**	**0.072**	**27.792**	**0.847**
	DR-0.5	**0.877**	6.906	0.098	26.404	0.823
Resolution of 512×512						
d_m	Method	BitAcc↑	FID↓	LPIPS↓	PSNR↑	SSIM↑
32	IR [9]	0.968	2.478	0.094	28.766	0.809
	GR	0.991	1.938	0.077	29.737	0.853
	DR-1.0	0.992	**1.606**	**0.068**	**30.172**	**0.856**
	DR-0.5	**0.994**	2.009	0.081	29.539	0.838
64	IR [9]	**0.889**	2.512	0.091	28.520	0.793
	GR	0.875	1.924	0.071	30.026	0.852
	DR-1.0	0.830	**1.429**	**0.060**	**30.683**	**0.864**
	DR-0.5	0.858	1.809	0.073	29.732	0.841

4.3 Robustness Analysis

To enhance robustness against real-world distortions, we integrated a distortion layer (e.g., JPEG compression, Gaussian noise) before the watermark decoder during training.

Building upon the WOUAF [9] framework, we evaluated the robustness of our proposed strategy in terms of BitAcc (the higher the better) against various post-processing attacks: center crop (ratio=0.1), rotation (25°), resizing (scale factor=0.5), JPEG compression (quality factor, QF=50), brightness adjustment (factor=1.5), contrast adjustment (factor=1.5), sharpening (factor=1.5), text overlaying, and a composite attack combining cropping (0.5), brightness adjustment (1.5), and JPEG compression (QF=80). As a baseline reference, we also compared the extraction performance under the "None" setting, i.e., in the absence of any attacks.

As demonstrated in Table 3, the DR-0.5 strategy maintains superior robustness across most attack scenarios. While exhibiting



(a) The procedure based on WOUAF [9].

(b) The procedure based on SSW [4].

Figure 4: Visualization of qualitative results. The 1st and 3rd columns of each subfigure show outputs from the baseline (IR in (a) and GR in (b)) and DR-0.5, respectively. Each watermarked image is followed by its corresponding pixel-wise difference map (×5 magnified) with respect to the clean-generated image. PSNR values are reported above each difference map.

Table 3: Bit-accuracy (↑) comparison under image post-processing attacks on the WOUAF model with image resolution of 256×256 and $d_m = 16$.

Attack Type	IR [9]	GR	DR-1.0	DR-0.5
None	**0.995**	0.992	0.847	0.994
Crop 0.1	0.647	0.646	0.634	**0.663**
Rot. 25	0.976	0.972	0.823	**0.979**
JPEG 50	**0.960**	0.909	0.763	0.920
Bright. 1.5	0.712	0.636	0.651	**0.863**
Contrast. 1.5	0.697	0.926	0.799	**0.933**
Sharp. 1.5	**0.995**	0.991	0.846	0.994
Resize 0.5	**0.928**	0.913	0.774	0.911
Overlay	0.499	0.602	0.589	**0.694**
Comb.	0.603	0.577	0.598	**0.747**

slightly lower performance than the IR baseline under JPEG compression, sharpening, and resizing operations, it still delivers competitive resilience. Notably, DR-0.5 shows particular strength against the composite attacks, validating its balanced design principle.

4.4 Generalization of the Dual-Reference Strategy

To validate the generalization of our dual-reference strategy, we conducted experiments based on the Stable Signature advanced watermarking method [4] (hereinafter referred to as SSW), under the configuration of 256×256 resolution and 48-bit watermark length. Unlike the previous method based on WOUAF [9], the watermarking technique SSW solely utilizes GR as the supervision signal for image quality and fine-tunes the VAE decoder with a fixed watermark.

In Table 4, we systematically evaluate the SSW watermarking framework under distinct reference strategies for image quality optimization: the input-image reference (IR), the original SSW baseline relying solely on clean-generated references (GR), and our proposed

Table 4: Performance comparison of Stable Signature [4] on image resolution 256×256 and 48-bit watermark length. IR: input-reference; GR: generated-reference (SSW baseline); DR-1.0 / DR-0.5: dual-reference methods with $\lambda_i = 1.0$ / 0.5.

Method	BitAcc↑	FID↓	LPIPS↓	PSNR↑	SSIM↑
IR	**0.982**	8.511	0.118	25.505	0.798
GR [4]	0.920	4.058	0.074	28.732	0.869
DR-1.0	0.909	**3.665**	**0.070**	**29.185**	**0.877**
DR-0.5	0.942	4.104	0.074	28.599	0.869

dual-reference strategies with $\lambda_i = 1.0$ (DR-1.0) and $\lambda_i = 0.5$ (DR-0.5). As shown in Table 4, although the IR strategy achieves the best BitAcc, its image quality is suboptimal, especially under fixed watermark conditions. The experimental results clearly demonstrate that the DR-1.0 strategy yields superior image quality due to its stronger optimization weighting on the image preservation objective $\mathcal{L}_i$ with respect to the watermark extraction loss $\mathcal{L}_w$, but this meanwhile comes at the cost of reduced watermark extraction capability. Most significantly, the DR-0.5 strategy achieves the optimal balance between watermark extraction accuracy and image quality preservation. These findings not only confirm the advantages of our proposed method but also strongly support the necessity of incorporating dual-reference supervision in generative watermarking systems.

The last four columns of Fig. 4(b) present the image generation results of SSW with the image resolution of 512×512. Under the DR-0.5 strategy, SSW achieves visually appealing and competitive image quality. Unlike WOUAF, SSW relies solely on the generated reference image for supervision. Compared to the SSW baseline, employing the DR-0.5 strategy not only leads to competitive and even improved image quality, but also provides a better overall balance between watermark extraction accuracy and perceptual fidelity.

5 Concluding Remarks

In this paper, we have proposed a dual-reference loss function for diffusion-based generative watermarking, with specific implementation and validation in the Stable Diffusion framework. Our method demonstrated consistent performance improvements when applied to various state-of-the-art watermarking techniques, achieving superior balance between watermark extraction accuracy and image quality preservation compared to single-reference approaches (using either input images or clean-generated images as exclusive references). This advancement holds significant promise for deepfake attribution, providing a robust mechanism to embed verifiable information within synthetic media.

For future work, we will investigate more intelligent weighting strategies that automatically adjust the trade-off between watermark extraction and image quality objectives during optimization. This direction promises to deliver enhanced performance balance and accelerated convergence, while maintaining the framework's robustness, thereby strengthening its application in the crucial area of deepfake detection and provenance.

Acknowledgments

This work was partially supported by these following projects: AI4Debunk (GA n. 101135757) funded by EU Horizon Europe Programme, SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU and FOSTERER funded by the Italian MUR PRIN-2022.

References

[1] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. 2007. *Digital Watermarking and Steganography* (2 ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[2] Luke Ditria and Tom Drummond. 2023. Hey That's Mine Imperceptible Watermarks are Preserved in Diffusion Generated Outputs. arXiv:2308.11123 [cs.MM] https://arxiv.org/abs/2308.11123

[3] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. 2024. AquaLoRA: toward white-box protection for customized stable diffusion models via watermark LoRA. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. Article 538, 22 pages.

[4] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22466–22477.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6629–6640.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 6840–6851.

[7] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. 2024. 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 (CVPR)*. 8974–8983.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. doi:10.1109/CVPR52688.2022.01042

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

[12] Matthew Tancik, Ben Mildenhall, and Ren Ng. 2020. StegaStamp: Invisible Hyperlinks in Physical Photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 2023. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=Z57JrmubNl

[14] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 2021. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14448–14457.

[15] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595. doi:10.1109/CVPR.2018.00068

[16] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. 2023. A Recipe for Watermarking Diffusion Models. arXiv:2303.10137 [cs.CV] https://arxiv.org/abs/2303.10137

[17] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. HiDDeN: Hiding Data with Deep Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.