

Diffusion-Latent Invisible Watermarking for Proactive Deepfake Provenance Verification

Bibek Das¹, Anurag Deo¹, Chandranath Adak¹, Soumi Chattopadhyay², Zahid Akhtar³, Soumya Dutta⁴, and Abdenour Hadid⁵

¹ Dept. of CSE, Indian Institute of Technology Patna, India– 801106.

² Dept. of CSE, Indian Institute of Technology Indore, India–453552

³ State University of New York Polytechnic Institute, USA–13502

⁴ Dept. of CSE, Indian Institute of Technology Kanpur, India–208016

⁵ SCAI, Sorbonne University Abu Dhabi, UAE
{bibek_2422cs27,chandranath}@iitp.ac.in

Abstract. The rapid advancement of generative face manipulation techniques has exposed fundamental limitations of conventional passive deepfake detection systems, particularly under distribution shift, compression, and pose variation. In this paper, we propose diffusion-latent invisible watermarking as a proactive deepfake provenance verification mechanism, enabling reliable authentication of synthetic content at the point of generation. Our approach embeds structured, imperceptible watermarks directly into the latent space of diffusion models, ensuring minimal visual distortion while preserving robust recoverability under post-processing operations. Unlike existing watermarking methods that are primarily evaluated on near-frontal facial imagery, our framework explicitly addresses non-frontal and side-face deepfake scenarios, where facial geometry and texture statistics differ substantially, and robustness is significantly challenged. We introduce a perceptual-robust latent reconstruction objective that enhances watermark stability under compression, geometric transformations, and learned distortions without sacrificing visual fidelity. To rigorously evaluate the proposed approach, we construct a side-face deepfake benchmark comprising diverse facial identities, multiple state-of-the-art manipulation methods, and controlled pose variations. Extensive experiments demonstrate that our method consistently outperforms prior diffusion-based watermarking baselines in watermark recovery accuracy and provenance verification reliability, particularly under extreme pose and post-processing conditions. These results highlight the effectiveness of proactive, diffusion-latent invisible watermarking as a practical and robust solution for trustworthy deployment of generative face technologies.⁶

Keywords: Deepfake provenance verification · Diffusion models · Invisible watermarking · Proactive deepfake defense

⁶ Codes will be made available upon acceptance of the paper. Institutional ethical clearance for conducting this research was obtained from IIT Patna.

1 Introduction

Recent advances in deep generative modeling have enabled the creation of highly photorealistic synthetic media, with manipulated facial images and videos, commonly referred to as deepfakes, emerging as a particularly concerning modality [1]. As modern generative models continue to improve in realism, controllability, and accessibility, they pose growing threats to the integrity and trustworthiness of digital content. These challenges have significant implications for journalism, digital forensics, law enforcement, and real-time communication platforms, where reliable content authentication is critical.

Most existing deepfake analysis techniques follow a passive detection paradigm, attempting to identify manipulated content by exploiting statistical inconsistencies, visual artifacts, or frequency irregularities introduced during the synthesis process [2]. While such approaches can achieve high accuracy on known manipulation methods, they often generalize poorly to unseen generators, post-processing operations, or distribution shifts [3]. This limitation is particularly severe in non-frontal and side-face deepfakes, where facial landmarks are partially occluded and pose-induced variations significantly degrade the reliability of conventional detection cues.

To address these limitations, we advocate a shift from passive detection toward a proactive provenance verification paradigm, in which synthetic content is authenticated at the point of generation by a cooperative model or platform. In this work, we propose a diffusion-latent invisible watermarking framework that embeds imperceptible yet robust watermark signals directly into the synthesis pipeline of pretrained diffusion models. By operating in the latent space, our approach leverages the structured representations learned by diffusion models while ensuring that the embedded watermark remains visually imperceptible in the generated image domain.

Specifically, we encode a structured concentric ring-based watermark pattern into the Fourier spectrum of the diffusion latent representation, followed by reverse diffusion to reconstruct a high-quality, watermarked image. To balance watermark robustness and visual fidelity, we introduce a composite reconstruction objective that jointly enforces perceptual similarity, resistance to common post-synthesis transformations (e.g., compression, resizing, and rotation), and resilience to high-level semantic perturbations. This design enables reliable watermark recovery even under adversarial or real-world distortions, while preserving the realism of the synthesized content.

A central focus of this work is the robustness of provenance verification under challenging pose conditions. Unlike prior watermarking approaches that predominantly evaluate near-frontal facial imagery, we explicitly study side-face and profile-view deepfakes, where pose-induced geometric and frequency variations present unique challenges. For this purpose, we evaluate our method on the IndicSideFace dataset [4], a benchmark designed for deepfake analysis on side-face perspectives of Indian subjects. The dataset comprises diverse identities, facial poses, lighting conditions, and visual attributes, enabling systematic and realistic evaluation across multiple manipulation scenarios.

The major **contributions** of this work are as follows:

(i) *Proactive deepfake provenance framework*: We introduce a diffusion-latent invisible watermarking approach for proactive provenance verification of synthetic face images, addressing the limitations of passive deepfake detection under pose variation and post-processing.

(ii) *Latent frequency-domain watermark encoding*: We design a structured concentric ring-based watermark and embed it into the Fourier domain of diffusion latent representations, achieving strong imperceptibility and robustness to common image transformations.

(iii) *Perceptual-robust reconstruction objective*: We formulate a multi-objective loss that jointly optimizes perceptual fidelity and watermark recoverability by combining pixel-level reconstruction loss, structural similarity, perceptual loss, and LPIPS, together with transformation-aware and semantic perturbation constraints.

(iv) *Comprehensive evaluation on side-face deepfakes*: We conduct extensive experiments on side-face deepfakes to demonstrate the effectiveness of the proposed approach in preserving watermark integrity and visual quality under extreme pose and post-processing conditions.

The remainder of the paper is organized as follows. Section 2 reviews relevant background and related work. Section 3 details the proposed methodology. Section 4 presents experimental results and analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2 Background and Related Works

This section reviews the foundational concepts and prior research relevant to deepfake generation, passive detection, and proactive provenance verification, with a particular focus on their limitations under pose variation and side-face scenarios.

2.1 Deepfake Generation Models

Recent progress in deepfake generation has been driven by advances in deep generative models that aim to improve photorealism, identity preservation, and controllability. Early image-based deepfakes were predominantly generated using GANs, which remain a foundational paradigm in facial synthesis. StyleGAN-based architectures, including StyleGAN3 [5], significantly improved image fidelity by mitigating aliasing artifacts and enhancing disentangled latent representations. Face-swapping systems such as DF-Platter [6] employ adversarial training to seamlessly blend source identities into target contexts, though they often struggle under low-resolution or extreme pose conditions.

More recently, diffusion models have emerged as the dominant paradigm for high-fidelity image synthesis. Latent diffusion frameworks such as Stable Diffusion and SDXL [7, 8] leverage denoising processes in compressed latent spaces to

enable controllable and scalable generation. Personalized diffusion models further extend this capability by incorporating identity-specific conditioning via test-time fine-tuning or offline training [9, 10], enabling realistic multi-subject and pose-diverse manipulations with minimal visual artifacts.

In parallel, specialized face-swapping models have been proposed to improve identity transfer under challenging conditions. SimSwap [11] injects identity features into target faces using a weak feature matching loss, while SimSwap++ [12] enhances efficiency through conditional dynamic convolution and knowledge distillation. Ghost [13] extends identity transfer to full head swapping, incorporating improved alignment and blending modules to handle complex backgrounds. FaceDancer [14] focuses on pose- and occlusion-aware swapping via adaptive feature fusion, achieving high fidelity without explicit facial segmentation. Despite these advances, such generators increasingly produce artifacts-free outputs, particularly in side-face and profile views, rendering post-hoc detection substantially more difficult.

2.2 Passive Deepfake Detection

Most existing deepfake forensics methods adopt a passive detection paradigm, aiming to classify manipulated content by identifying synthesis-induced artifacts or statistical irregularities. Early studies demonstrated that CNN-generated images exhibit systematic frequency-domain artifacts that can be exploited for cross-model generalization [15]. Transformer-based architectures, such as Vision Transformer (ViT) detectors [16], further improved detection accuracy by modeling long-range dependencies through attention mechanisms.

More recent approaches focus on identifying subtle structural inconsistencies. Tan et al. [17] proposed Neighboring Pixel Relationships (NPR) to capture up-sampling artifacts across a large set of generative models, while Seferbekov et al. [18] combined EfficientNet-based ensembles with pairwise learning for robust video-level detection in the DFDC challenge. To address computational constraints, Romeo et al. [19] introduced Binary Neural Networks with FFT and LBP features, achieving substantial speedups with minimal performance degradation. Baru et al. [20] proposed Wavelet-CLIP, integrating wavelet-domain features with CLIP embeddings to enhance generalization.

Despite these advances, passive detectors remain fundamentally vulnerable to distribution shifts, unseen generators, and post-processing operations. These limitations are exacerbated in non-frontal and side-face deepfakes, where occluded landmarks and pose-induced geometric variations significantly weaken artifact-based cues. As a result, purely post-hoc detection approaches often fail to provide reliable guarantees in realistic deployment settings.

2.3 Proactive Watermarking and Provenance Verification

To overcome the inherent limitations of passive detection, recent research has explored proactive forensic frameworks that embed watermarks into content at generation or capture time. Unlike passive detectors, these methods aim to verify

the provenance and authenticity of synthetic media generated by cooperative systems, enabling reliable authentication even when visual artifacts are absent.

Several approaches leverage facial biometrics to construct content-aware watermarks. FaceProtect [21] introduces a GAN-based one-way dynamic watermark generation mechanism that maps facial feature embeddings to irreversible watermark signals, enabling steganographic verification while preventing reverse engineering. LampMark [22] derives structure-sensitive watermarks from facial landmarks, encoding geometric information to enhance robustness against identity-preserving manipulations.

More recently, AdvMark [23] proposed adversarial watermarking strategies that jointly optimize watermark robustness and detector performance by aligning watermark signals with forgery artifacts. While effective, such approaches remain tightly coupled to specific detector architectures and do not explicitly address robustness under pose variation. Neural identity entanglement methods [24] embed pseudo-random signatures into identity representations via encoder-decoder networks, demonstrating improved resilience against adversarial attacks compared to artifact-based detectors.

However, existing proactive watermarking methods predominantly operate in pixel or feature spaces and are typically evaluated on near-frontal facial imagery. Their robustness under extreme pose changes, particularly side-face and profile-view deepfakes, remains largely unexplored. In contrast, our work leverages diffusion-latent representations to embed invisible watermarks and explicitly investigates provenance verification under side-face scenarios, addressing a critical gap in current deepfake forensics literature.

3 Proposed Methodology

This section formulates the problem statement, and presents the overall framework of our approach, detailing the key design principles.

3.1 Problem Formulation

Conventional deepfake analysis methods operate in a reactive manner, attempting to detect manipulations after a synthesized image has been produced. Given an original image I and a generator $\mathcal{G}$, a deepfake image $I_f = \mathcal{G}(I)$ is analyzed by a detector $\mathcal{D}$ to identify synthesis artifacts. However, as generative models evolve and produce increasingly photorealistic outputs, detectors $\mathcal{D}$ trained on specific manipulation signatures exhibit limited generalization. Additionally, common postprocessing techniques like blurring and compression have a tendency to mask these subtle artifacts, which further reduces detection performance.

To address these challenges, we formulate deepfake analysis as a proactive provenance verification problem. The goal is to embed an imperceptible watermark W at generation time such that the origin of synthetic content can be reliably verified even after post-processing. Formally, given a deepfake image I_f , we generate a visually indistinguishable watermarked image I_w that encodes

W and enables reliable verification from a potentially transformed version I'_w . The detection module $\mathcal{D}$ acts as a verifier that extracts and verifies the presence of the embedded watermark W from I'_w , providing a principled measure of content provenance that remains robust to evolving generative models and post-processing operations.

3.2 Framework Overview

Fig. 1 provides an overview of the proposed diffusion-latent invisible watermarking framework for proactive deepfake provenance verification. The framework consists of three stages: (i) latent-space watermark embedding during image synthesis, (ii) robustness evaluation under post-processing and semantic perturbations, and (iii) watermark extraction and statistical verification.

Diffusion-Latent Watermark Embedding: Let $I_f \in \mathbb{R}^{h \times w \times c}$ denote a synthesized deepfake image. We first invert I_f into the latent space of a pretrained diffusion model using DDIM inversion:

$$Z_T = \mathcal{G}'(I_f), \quad (1)$$

where, $\mathcal{G}'$ denotes the frozen diffusion encoder-decoder operating in inference mode.

The watermark $W \in \mathbb{R}^{h' \times w'}$ is generated via radial sampling from a Gaussian distribution, following the Tree-Ring watermark design [25]. To embed the watermark, the latent representation Z_T is transformed into the Fourier domain, and the watermark is injected into a predefined frequency band using a binary circular mask $\mathcal{M} \in \{0, 1\}^{h' \times w'}$:

$$F(Z_T)[i_c, :, :] = (1 - \mathcal{M}) \odot F(Z_T)[i_c, :, :] + \mathcal{M} \odot W, \quad (2)$$

where, $F(\cdot)$ denotes the Fourier transform, i_c is the designated watermark channel, and $\odot$ represents element-wise multiplication.

The modified latent representation is then decoded via reverse diffusion to obtain the watermarked image:

$$I_w = \mathcal{G}'(Z_T \oplus W), \quad (3)$$

where, $\oplus$ denotes element-wise addition in latent space.

Perceptual-Robust Reconstruction Objective: To ensure that watermark embedding does not degrade visual quality while preserving robustness, we optimize a composite reconstruction loss:

$$\mathcal{L} = \lambda_m \mathcal{L}_m(I_f, I_w) + \lambda_s \mathcal{L}_s(I_f, I_w) + \lambda_p \mathcal{L}_p(I_f, I_w) + \lambda_l \mathcal{L}_l(I_f, I_w), \quad (4)$$

where, $\mathcal{L}_m$ is the mean squared error loss, $\mathcal{L}_s$ is the structural similarity (SSIM) loss [26], $\mathcal{L}_p$ is the Watson-VGG perceptual loss [27], and $\mathcal{L}_l$ is the LPIPS loss [28]. The coefficients λ_m , λ_s , λ_p , and λ_l balance fidelity and robustness.

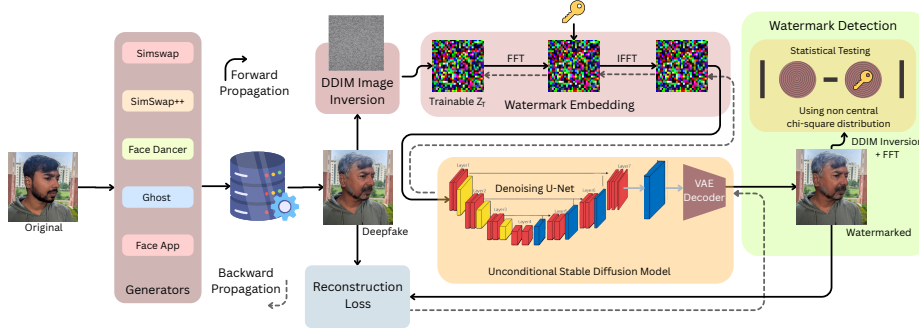


Fig. 1. Overview of the proposed model. The pipeline consists of three stages: (i) *Watermark Embedding*—a synthesized deepfake image I_f is inverted into the diffusion latent space via DDIM, where an imperceptible watermark is embedded before reconstructing the watermarked image I_w through reverse diffusion; (ii) *Post-Processing Simulation*—to evaluate robustness, I_w is subjected to common post-processing operations such as compression and blurring, yielding a perturbed image I'_w ; and (iii) *Watermark Extraction and Verification*—the perturbed image is mapped back to the latent space via DDIM inversion, followed by FFT-based decoding and a non-central χ^2 statistical test to verify watermark presence and establish content provenance. The robustness of the proposed framework stems from watermark embedding in the diffusion latent domain.

During optimization, the diffusion model remains frozen; gradients are propagated only through the latent variables. This design ensures that the embedded watermark remains imperceptible while being resilient to common post-processing operations.

Pose Robustness via Latent-Space Embedding: Embedding the watermark in the diffusion latent space provides intrinsic robustness to pose variation. Unlike pixel-space representations, diffusion latents encode global semantic structure that is comparatively invariant to geometric transformations such as yaw rotation and partial self-occlusion. Consequently, side-face manipulations induce smaller perturbations in the latent-frequency domain, preserving watermark detectability even under extreme pose changes.

Watermark Extraction and Statistical Verification: Given a potentially transformed image I'_w , watermark verification proceeds by first recovering the latent representation via DDIM inversion:

$$Z'_T = \mathcal{G}'(I'_w). \quad (5)$$

The latent is then transformed into the Fourier domain $\hat{f} = F(Z'_T)$.

We perform hypothesis testing under the null hypothesis $\mathcal{H}_0$, which assumes the absence of an embedded watermark:

$$\mathcal{H}_0 : \hat{f} \sim \mathcal{N}(0, \hat{\sigma}^2 \mathbf{I}), \quad (6)$$

where, $\mathbf{I}$ is the identity matrix. The variance $\hat{\sigma}^2$ is estimated using the watermark mask $\mathcal{M}$:

$$\hat{\sigma}^2 = \frac{1}{|\mathcal{M}|} \sum_{i,j} \left(\mathcal{M}_{i,j} \cdot \hat{f}_{i,j} \right)^2. \quad (7)$$

The verification statistic is computed as:

$$\mathcal{D} = \frac{1}{\hat{\sigma}^2} \sum \left(\mathcal{M} \odot (W - \hat{f}) \right)^2. \quad (8)$$

Since DDIM inversion maps natural images into a Gaussian latent distribution, $\mathcal{D}$ follows a non-central χ^2 distribution in the presence of an embedded watermark. A p -value is computed accordingly. If $p < T_h$, where T_h is a predefined threshold, the watermark is verified as present. The verification confidence is defined as $(1 - p)$, providing a principled measure of provenance certainty.

This statistical formulation enables reliable watermark verification independent of the specific generator architecture or post-processing pipeline.

4 Experimental Analysis

In this section, we present the results obtained from our experiments and provide a comprehensive analysis of their implications. We evaluate the performance of our proposed approach using relevant metrics and benchmarks, comparing it against baseline methods where appropriate. The analysis highlights both the strengths and limitations of our method, offering insights into its practical applicability and theoretical significance.

4.1 Employed SideFace Dataset

All experiments in this work are conducted using the IndicSideFace dataset [4], a benchmark designed to facilitate deepfake analysis under side-face and non-frontal viewing conditions of Indian subjects. The dataset was introduced to address a critical gap in existing deepfake benchmarks, which predominantly focus on near-frontal facial imagery.

IndicSideFace comprises 984 genuine images collected from 164 individuals, representing diverse Indian demographics. For each subject, six images are captured under controlled conditions: three indoor and three outdoor samples corresponding to left-profile, frontal, and right-profile facial views. This structured acquisition enables systematic evaluation across varying pose and illumination conditions.

To support a comprehensive evaluation of synthetic face manipulations, the dataset includes 19680 deepfake images generated using four state-of-the-art identity-swapping frameworks (Ghost [13], SimSwap [11], SimSwap++ [12], and FaceDancer [14]) along with one facial attribute manipulation tool (FaceApp [29]). These generators collectively introduce a wide range of spoofing characteristics, identity transformations, and semantic alterations.

The combination of diverse identities, lighting environments, head poses, and manipulation techniques makes IndicSideFace particularly well-suited for evaluating robustness under challenging side-face scenarios. In this work, the dataset serves as a rigorous benchmark for assessing the reliability of diffusion-latent invisible watermarking and provenance verification under pose-induced distribution shifts, rather than as a newly proposed dataset.

4.2 Performance of Passive Deepfake Detectors on the SideFace Dataset

We evaluate the performance of several state-of-the-art passive deepfake detectors on the IndicSideFace dataset [4] to assess their robustness under side-face and non-frontal viewing conditions. As reported in Table 1, a substantial performance degradation is observed when these detectors are applied to side-face deepfake images.

The experimental protocols and evaluation settings follow those described in [4]. The results highlight the inherent limitations of post-hoc, artifact-based detection approaches under pose-induced distribution shifts, where occluded landmarks and geometric variations significantly weaken discriminative cues.

These findings underscore the need for alternative paradigms beyond passive detection and motivate the adoption of proactive deepfake provenance verification strategies, such as the diffusion-latent invisible watermarking framework proposed in this work.

Table 1. Performance of SOTA passive deepfake detectors on IndicSideFace [4]

	Detector	$\mathcal{P}$ %	$\mathcal{R}$ %	$\mathcal{F}_1$ %	$\mathcal{A}$ %	$\mathcal{BA}$ %
Pre-trained	CNN-Detector [15]	50.55	97.18	66.49	50.45	49.90
	NPR [17]	52.98	49.61	51.23	51.01	47.35
	Selim [18]	49.37	89.72	63.68	47.07	47.34
	ViT [16]	50.79	12.92	20.44	24.14	36.01
Transfer learning	CNN-Detector [15]	86.73	87.37	87.04	88.21	84.97
	NPR [17]	61.38	50.80	55.58	60.01	51.69
	Selim [18]	84.71	85.19	84.90	84.79	84.79
	ViT [16]	88.33	87.27	87.79	86.56	85.39

$\mathcal{P}$: Precision, $\mathcal{R}$: Recall, $\mathcal{F}_1$: F1 score, $\mathcal{A}$: Accuracy, $\mathcal{BA}$: Balanced accuracy

4.3 Performance of Proactive Deepfake Provenance Verification on the SideFace Dataset

This section presents a comprehensive qualitative and quantitative evaluation of the proposed proactive deepfake provenance verification framework on IndicSideFace [4]. The analysis focuses on assessing watermark imperceptibility, verification robustness, and reliability under challenging side-face and non-frontal viewing conditions, as well as under common post-processing operations.



Fig. 2. Comparison of original (I), deepfake (I_f), and watermarked (I_w) images across five generators: SimSwap [11], SimSwap++ [12], Ghost [13], FaceDancer [14], and FaceApp [29]. This layout highlights the similarity in deepfake image I_f and watermarked image I_w in the human naked eye. Best viewed in color.

Qualitative Analysis: Fig. 2 presents a qualitative comparison of original images (I), deepfake images (I_f), and their corresponding watermarked counterparts (I_w) across five representative generators: SimSwap [11], SimSwap++ [12], Ghost [13], FaceDancer [14], and FaceApp [29]. The examples span multiple facial poses (frontal, left-profile, and right-profile) and capture conditions (indoor and outdoor), reflecting the diversity of the IndicSideFace dataset.

The primary objective of this visualization is to assess whether diffusion-latent watermark embedding introduces perceptible visual artifacts that could compromise the realism of synthesized deepfakes. Across all generators and pose conditions, the watermarked images (I_w) remain visually indistinguishable from the corresponding deepfake images (I_f) under normal viewing conditions. Facial identity, expression, pose, and global appearance are consistently preserved, indicating that the proposed watermarking strategy does not disrupt semantic or structural content.

A closer inspection reveals only minor, low-level variations between I_f and I_w , primarily manifesting as subtle changes in background luminance or slight tonal adjustments in hair and skin regions. These effects are infrequent, generator-dependent, and visually inconspicuous, and do not affect perceived identity or facial attributes. Importantly, no structured artifacts, spatial distortions, or watermark patterns are observable in the spatial domain.

Overall, the qualitative results confirm that the proposed diffusion-latent invisible watermarking framework preserves perceptual realism across diverse generators and challenging side-face scenarios, supporting its suitability for practical provenance verification without compromising visual quality.

Quantitative Analysis: Table 2 presents a comprehensive evaluation of various watermarking techniques applied to generative AI-based generated deep-

Table 2. Comparison of watermarking methods across various deepfake generators under different attack scenarios

Generator	Watermarking Method	SSIM	Pre-Attack	Post-Attack							
				Brightness	Contrast	JPEG	G-Noise	G-Blur	Rotation	Bmshj18 [30]	Cheng20 [31]
FaceApp [29]	DwtDct	0.9815	0.7551	0.7483	0.7485	0.7478	0.7444	0.7550	0.7364	0.7352	0.7266
	DwtDctSvd [32]	0.9816	0.8730	0.7452	0.7454	0.7569	0.8281	0.8281	0.7500	0.7417	0.7428
	SSL [33]	0.9937	0.4966	0.4988	0.4980	0.4844	0.4918	0.4910	0.4928	0.4818	0.4814
	StegaStamp [34]	0.9668	0.9999	0.9995	0.9995	0.8097	0.9515	0.9998	0.4995	0.9715	0.9485
	RivaGAN [35]	0.9959	1.0000	0.8816	0.8641	0.9606	0.9934	0.9997	0.4441	0.5512	0.5559
	Zodiac [36]	0.9242	1.0000	1.0000	1.0000	0.9999	0.9995	1.0000	0.8468	0.9978	0.9989
	Ours	0.9994	1.0000	1.0000	1.0000	0.9957	0.9891	1.0000	0.8870	0.9931	0.9941
FaceDancer [14]	DwtDct	0.9861	0.7500	0.7472	0.7479	0.7492	0.7462	0.7500	0.7362	0.7447	0.7426
	DwtDctSvd [32]	0.9881	0.8571	0.7457	0.7464	0.7560	0.8175	0.8056	0.7500	0.7452	0.7495
	SSL [33]	0.9930	0.4953	0.4943	0.4928	0.4822	0.4922	0.4850	0.4971	0.4807	0.4872
	StegaStamp [34]	0.9690	0.9999	0.9996	0.9991	0.8110	0.9507	0.9998	0.4976	0.9694	0.9472
	RivaGAN [35]	0.9964	0.985	0.8716	0.8588	0.9488	0.9906	1.0000	0.4278	0.5494	0.5534
	Zodiac [36]	0.92665	1.0000	1.0000	1.0000	0.9999	0.9999	1.0000	0.859	0.9983	0.9988
	Ours	0.9994	1.0000	1.0000	1.0000	0.9989	0.9912	1.0000	0.9264	0.9976	0.9954
Ghost [13]	DwtDct	0.9856	0.7507	0.7472	0.7472	0.7478	0.7387	0.7527	0.7321	0.7421	0.7350
	DwtDctSvd [32]	0.9883	0.8596	0.7488	0.7500	0.7540	0.8192	0.8036	0.7500	0.7474	0.7475
	SSL [33]	0.9933	0.4964	0.4961	0.4939	0.4799	0.4934	0.4865	0.4963	0.4796	0.4875
	StegaStamp [34]	0.9690	0.9999	0.9998	0.9997	0.8128	0.9526	0.9999	0.4951	0.9708	0.9460
	RivaGAN [35]	0.9963	0.9850	0.8638	0.8566	0.9553	0.9816	0.9894	0.4431	0.5522	0.5528
	Zodiac [36]	0.9288	1.0000	0.9998	1.0000	0.9998	0.9999	0.9999	0.8674	0.9984	0.9984
	Ours	0.9993	1.0000	1.0000	1.0000	0.9969	0.9971	1.0000	0.8746	0.9898	0.9966
SimSwap [11]	DwtDct	0.9860	0.7500	0.7459	0.7452	0.7485	0.7539	0.7507	0.7397	0.7436	0.7397
	DwtDctSvd [32]	0.9880	0.8581	0.7500	0.7500	0.7525	0.8224	0.8056	0.7500	0.7464	0.7459
	SSL [33]	0.9926	0.4977	0.4965	0.4960	0.4820	0.4957	0.4892	0.4997	0.4856	0.4862
	StegaStamp [34]	0.9680	0.9999	0.9998	0.9996	0.8163	0.9542	0.9999	0.4954	0.9739	0.9509
	RivaGAN [35]	0.9963	0.9650	0.8919	0.8756	0.9656	0.9919	1.0000	0.435	0.5581	0.5528
	Zodiac [36]	0.925	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8161	0.9991	0.9904
	Ours	0.9994	1.0000	1.0000	1.0000	0.9975	0.9944	1.0000	0.9156	0.9940	0.9960
SimSwap++ [37]	DwtDct	0.9856	0.7472	0.7444	0.7429	0.7485	0.7430	0.7522	0.7328	0.7417	0.7416
	DwtDctSvd [32]	0.9878	0.8591	0.7488	0.7494	0.7545	0.8234	0.8036	0.7490	0.7380	0.7400
	SSL [33]	0.9931	0.4965	0.4952	0.4924	0.4827	0.4964	0.4874	0.4979	0.4812	0.4783
	StegaStamp [34]	0.9678	0.9999	0.9995	0.9994	0.8141	0.9547	0.9999	0.4972	0.9729	0.9471
	RivaGAN [35]	0.9962	0.9900	0.8766	0.8669	0.9584	0.9928	1.0000	0.4341	0.5481	0.5481
	Zodiac [36]	0.9254	1.0000	1.0000	1.0000	0.9972	0.9989	1.0000	0.8604	0.9982	0.998
	Ours	0.9993	1.0000	1.0000	1.0000	0.9950	0.9932	1.0000	0.9196	0.9916	0.9948

fake content, focusing on the fidelity and robustness of the proposed model in comparison to six baseline methods: DwtDct, DwtDctSvd [32], SSL [33], StegaStamp [34], RivaGAN [35], and Zodiac [36]. The analysis spans five distinct generators, such as FaceApp [29], FaceDancer [14], Ghost [13], SimSwap [11], and SimSwap++ [12], which represent various synthetic face generation techniques mainly used to generate deepfake images. To simulate realistic adversarial conditions, watermarked outputs were subjected to a range of post-processing attacks, including adjustments in brightness or contrast with a factor of 0.5, JPEG compression with a quality setting of 50, image rotation by 90 degrees, addition of Gaussian noise (standard deviation of 0.05), and Gaussian blur (kernel size of 5 and standard deviation of 1). In addition, more advanced transformations were employed: VAE-based image compression models Bmshj18 [30] and Cheng20 [31] with a quality level of 3.

In this evaluation, image quality was quantified using the Structural Similarity Index (SSIM) metric, which is measured in decibels (dB), where higher values of SSIM indicate greater structural similarity between the watermarked and original images. The proposed model consistently demonstrates superior performance across all generated deepfake datasets and attack types. Notably, it achieves near-perfect SSIM values of 0.999, indicating negligible perceptual distortion relative to the original images. Furthermore, the proposed method

consistently achieves perfect pre-attack bit accuracy near 0.9999 to 1 across all generator models.

In post-attack scenarios, our proposed model, consistently demonstrates superior robustness across all evaluated deepfake data generated by five different generators. Notably, our model maintains exceptionally high bit recovery rates, frequently reaching 1.000 even under challenging conditions such as Gaussian Blur, Brightness, and Contrast distortions. It also demonstrates remarkable resilience to aggressive transformations such as rotation (90°), a condition under which all competing methods compared here struggle to achieve acceptable recovery. For instance, on FaceApp, StegaStamp and SSL degrade to 0.4995 and 0.4928, respectively, under Rotation, while RivaGAN’s watermark is completely destroyed while rotating the image, whereas our model retains a bit recovery rate of 0.8870, showcasing its robustness to affine distortions.

Across a variety of fake images generated by multiple deepfake generators, our model consistently outperforms all baselines under various attack scenarios. For example, under JPEG compression, our model achieves bit recovery rates close to 0.99, whereas DwtDctSvd and StegaStamp drop to 0.7569 and 0.8097 on FaceApp, and to 0.7560 and 0.8110 on FaceDancer. Furthermore, our model exhibits near-perfect robustness across all additive noise attacks (e.g., Gaussian Noise, Blur), with recovery rates often exceeding 0.99, emphasizing its generalization across spatial and frequency-domain perturbations. These results confirm that our model not only preserves high-fidelity watermark invisibility as evidenced by its superior SSIM (~ 0.999) and Pre-Attack recovery scores (1.0) but also delivers state-of-the-art resilience against a diverse spectrum of adversarial attacks. Leveraging these capabilities, our approach integrates a proactive deepfake detection mechanism that reliably flags and mitigates the fake content.

4.4 Performance of Proactive Deepfake Provenance Verification on Public Benchmark Datasets

We further evaluate the proposed proactive deepfake provenance verification framework on three widely used public datasets: FaceForensics++ [2], CelebA [38], and FFHQ [39]. The objective of this evaluation is to assess watermark imperceptibility and visual fidelity across diverse data distributions beyond the Indic-SideFace benchmark.

As reported in Table 3, the proposed method achieves consistently high Structural Similarity Index Measure (SSIM) scores across all datasets, with values exceeding 0.999 under unperturbed conditions. This indicates that diffusion-latent watermark embedding preserves structural content and visual quality across a wide range of facial images. Under common post-processing operations, including brightness and contrast adjustment, JPEG compression, Gaussian noise, and blurring, the SSIM values remain largely stable, demonstrating minimal perceptual degradation.

Notably, the method maintains strong visual fidelity even under learned compression schemes such as Bmshj18 and Cheng20, with SSIM values remaining

above 0.996. More challenging geometric transformations, such as rotation, induce a larger but expected drop in SSIM (approximately 0.87), reflecting the inherent sensitivity of pixel-level similarity metrics to spatial misalignment rather than perceptual failure.

Overall, these results confirm that the proposed diffusion-latent invisible watermarking framework preserves visual realism and structural integrity across diverse public datasets and post-processing conditions, supporting its applicability for practical provenance verification without compromising image quality.

Table 3. Performance of our method on public benchmark datasets

Public Datasets	SSIM	Pre-Attack	Post-Attack							
			Brightness	Contrast	JPEG	G-Noise	G-Blur	Rotation	Bmshj18	Cheng20
FaceForensics [2]	0.9994	1.0000	1.0000	1.0000	0.9966	0.9963	1.0000	0.8746	0.9875	0.9963
Celeba [38]	0.9993	1.0000	1.0000	1.0000	0.9967	0.9963	1.0000	0.8743	0.9887	0.9965
FFHQ [1]	0.9997	1.0000	1.0000	1.0000	0.9967	0.9971	1.0000	0.8749	0.9889	0.9967

4.5 Ablation Study

To assess the contribution of each component in the proposed reconstruction objective (Eqn. 4), we conduct an ablation study on FaceApp-generated samples from the IndicSideFace dataset. The results are summarized in Table 4.

We evaluate three loss configurations: (i) a baseline combining the pixel-level and perceptual losses ($\mathcal{L}_m, \mathcal{L}_s, \mathcal{L}_p$); (ii) a configuration using the LPIPS loss $\mathcal{L}_l$ alone; and (iii) the full composite loss ($\mathcal{L}_m, \mathcal{L}_s, \mathcal{L}_p, \mathcal{L}_l$).

Post-attack evaluation results indicate that the full composite loss consistently achieves the highest robustness across all considered distortions. In particular, incorporating the LPIPS term significantly improves the stability of watermark verification while maintaining high perceptual fidelity. These findings confirm that the complementary nature of pixel-level, structural, and perceptual constraints is essential for achieving a favorable trade-off between visual quality and watermark robustness in diffusion-latent embedding.

Table 4. Ablation study with various loss functions

Loss configurations	SSIM	Pre-Attack	Post-Attack							
			Brightness	Contrast	JPEG	G-Noise	G-Blur	Rotation	Bmshj18	Cheng20
combining $\mathcal{L}_m, \mathcal{L}_s, \mathcal{L}_p$	0.9697	1.0000	0.9962	0.9985	0.9852	0.9742	0.9975	0.8869	0.9845	0.9854
$\mathcal{L}_l$	0.9884	1.0000	0.9942	0.9913	0.9845	0.9743	0.9965	0.8866	0.9824	0.9469
combining $\mathcal{L}_m, \mathcal{L}_s, \mathcal{L}_p, \mathcal{L}_l$	0.9994	1.0000	1.0000	1.0000	0.9957	0.9891	1.0000	0.8870	0.9931	0.9941

5 Conclusion

We presented a diffusion-latent invisible watermarking framework for proactive deepfake provenance verification that overcomes key limitations of passive de-

tection approaches. By embedding imperceptible watermarks in the latent space of diffusion models, the proposed method enables reliable authentication of synthetic facial content under post-processing operations and pose-induced distribution shifts. Extensive experiments on the IndicSideFace benchmark and multiple public datasets demonstrate that the framework preserves high visual fidelity while achieving robust watermark verification, including under challenging side-face conditions. These results underscore the effectiveness of diffusion-latent embedding combined with perceptual-robust optimization for trustworthy content authentication. This work advocates a shift from post-hoc deepfake detection to proactive provenance verification and provides a scalable foundation for secure deployment of generative face technologies. Future work will extend the framework to video diffusion models and real-time verification pipelines.

Acknowledgments

Partial support from IHUB-NTIHAC Foundation, IIT Kanpur, under proposal no. 3092 is gratefully acknowledged.

References

1. Tero Karras and Others. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, pages 4396–4405, 2019.
2. Andreas Rössler et al. FaceForensics++: Learning to Detect Manipulated Facial Images, 2019.
3. Junyi Cao et al. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *CVPR*, pages 4103–4112, 2022.
4. Anurag Deo et al. IndicSideFace: A Dataset for Advancing Deepfake Detection on Side-Face Perspectives of Indian Subjects. In *IEEE FG*, pages 1–10, 2025.
5. Tero Karras et al. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.
6. Narayan, Vatsa, Mayank and others. Df-platter: Multi-face heterogeneous deepfake dataset. In *CVPR*, pages 9739–9748, 2023.
7. Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.
8. Dustin Podell et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023.
9. Yuwei Guo et al. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv*, 2023.
10. Jing Gu et al. Swapanything: Enabling arbitrary object swapping in personalized image editing. In *ECCV*, pages 402–418. Springer, 2024.
11. Renwang Chen et al. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, pages 2003–2011, 2020.
12. Xuanhong Chen et al. SimSwap++: Towards faster and high-quality identity swapping. *IEEE TPAMI*, 46(1):576–592, 2023.
13. Alexander Groshev et al. GHOST—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022.

14. Felix Rosberg et al. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *WACV*, pages 3454–3463, 2023.
15. Sheng-Yu Wang et al. CNN-generated images are surprisingly easy to spot... for now, 2020.
16. Wvolf. ViT Deepfake Detection. https://huggingface.co/Wvolf/ViT_Deepfake_Detection. Accessed: 2025-01-31.
17. Chuangchuan Tan et al. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, pages 28130–28139, 2024.
18. Selim Seferbekov. DeepFake Detection (DFDC) Solution. https://github.com/selimsef/dfdc_deepfake_challenge, 2020. Accessed: 2025-01-31.
19. Lanzino Romeo et al. Faster Than Lies: Real-time Deepfake Detection using Binary Neural Networks. *arXiv:2406.04932*, 2024.
20. Lalith Bharadwaj Baru et al. Wavelet-Driven Generalizable Framework for Deepfake Face Forgery Detection. In *WACV*, pages 1661–1669, 2025.
21. Shulin Lan et al. Facial Features Matter: a Dynamic Watermark based Proactive Deepfake Detection Approach. *arXiv*, 2024.
22. Tianyi Wang et al. LampMark: Proactive Deepfake Detection via Training-Free Landmark Perceptual Watermarks. In *ACM MM*, pages 10515–10524, 2024.
23. Xiaoshuai Wu et al. Are watermarks bugs for deepfake detectors? rethinking proactive forensics. *arXiv*, 2024.
24. Yuan Zhao and others Liu. Proactive deepfake defence via identity watermarking. In *WACV*, pages 4602–4611, 2023.
25. Yuxin Wen et al. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv:2305.20030*, 2023.
26. Hang Zhao et al. Loss Functions for Image Restoration With Neural Networks. *IEEE TCI*, 3(1):47–57, 2017.
27. Steffen Czolbe et al. A loss function for generative neural networks based on watson’s perceptual model. In *NIPS ’20*, NIPS ’20. Curran Associates Inc., 2020.
28. Richard Zhang et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, 2018.
29. FaceApp Technology Ltd. FaceApp. <https://www.faceapp.com>.
30. Johannes Ballé et al. Variational image compression with a scale hyperprior, 2018.
31. Zhengxue Cheng et al. Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules. In *CVPR*, 2020.
32. K. A. Navas et al. DWT-DCT-SVD based watermarking. In *COMSWARE ’08*, pages 271–274, 2008.
33. Watermarking Images in Self-Supervised Latent Spaces, author=Fernandez, Pierre and others. In *ICASSP*. IEEE, 2022.
34. Matthew Tancik et al. StegaStamp: Invisible Hyperlinks in Physical Photographs. In *CVPR*, 2020.
35. Kevin Alex Zhang et al. Robust Invisible Video Watermarking with Attention. 2019.
36. Lijun Zhang et al. Attack-Resilient Image Watermarking Using Stable Diffusion. In *NeurIPS*, 2024.
37. Xuanhong Chen et al. SimSwap++: Towards Faster and High-Quality Identity Swapping. *IEEE TPAMI*, 46(1):576–592, 2024.
38. Ziwei Liu et al. Deep learning face attributes in the wild. In *ICCV*, December 2015.
39. Tero Karras et al. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4396–4405, 2019.