DeepForgeSeal: Latent Space-Driven Semi-Fragile Watermarking for Deepfake Detection Using Adversarial Reinforcement Learning

Tharindu Fernando, *Member, IEEE,* Clinton Fookes, *Senior Member, IEEE,* and Sridha Sridharan, *Life Senior Member, IEEE.*

***Abstract*—Rapid advances in generative AI have led to increasingly realistic deepfakes, posing growing challenges for law enforcement and public trust. Existing passive deepfake detectors struggle to keep pace, largely due to their dependence on specific forgery artifacts, which limits their ability to generalize to new deepfake types. Proactive deepfake detection using watermarks has emerged to address the challenge of identifying high-quality synthetic media. However, these methods often struggle to balance robustness against benign distortions with sensitivity to malicious tampering. This paper introduces a novel deep learning framework that harnesses high-dimensional latent space representations and the Adversarial Reinforcement Learning (ARL) paradigm to develop a robust and adaptive watermarking approach. Specifically, we develop a learnable watermark embedder that operates in the latent space, capturing high-level image semantics, while offering precise control over message encoding and extraction. The ARL paradigm empowers the learnable watermarking module to pursue an optimal balance between robustness and fragility. This is achieved through interaction with a dynamic curriculum of benign and malicious image manipulations simulated by an adversarial attacker agent. Comprehensive evaluations on the CelebA and CelebA-HQ benchmarks reveal that our method consistently outperforms state-of-the-art approaches, achieving improvements of over 4.5% on CelebA and more than 5.3% on CelebA-HQ under challenging manipulation scenarios.**

***Index Terms*—Deepfake Detection, Semi-Fragile Watermarks, Adversarial Reinforcement Learning, Digital Forensics**

I. Introduction

Generative AI has rapidly transformed the way we create and edit multimedia, making it faster and easier than ever to produce high-quality images, audio, and video [1]. While these innovations unlock exciting possibilities, they also introduce serious risks. Among the most concerning threats are image and video deepfakes, highly realistic, AI-generated images and videos that mimic real individuals. These deepfakes pose significant challenges, not only for their role in spreading misinformation and disinformation at an alarming pace but also for their implications on privacy, consent, security, and societal trust [2], [3], [4], [5], [6].

To combat the rapid spread of deepfakes, watermarking [7], [8] has emerged as a technique that allows users to validate the authenticity and integrity of the multimedia they consume. A robust watermarking technique should be resilient to distortions such as JPEG compression, cropping, and colour changes to keep the embedded watermark intact. While being resilient to such incidental distortions, a good watermarking technique should demonstrate fragility towards content-altering modifications, such as face-swapping-type malicious edits that alter the semantic-level meaning of the image. Recently, deep learning techniques [9], [10], [11], [12], [13] have made significant progress towards achieving this paradoxical combination of resilience and fragility. For instance, SepMark [11] introduced an encoder–decoder architecture with a unified encoder for watermark embedding. It employs two distinct decoders to recover the watermark under different robustness constraints, enabling selective resilience to varying levels of adversarial distortion. However, most existing deep watermarking techniques for deepfake detection rely on a fixed, predefined watermark location for extraction. This design choice

Fig. 1. Watermarking and watermark retrieval performance of proposed *DeepForgeSeal* framework. The visualisations include low-resolution and high-resolution bonafide images (row 1), numerous learned attacks against watermarking (row 2), and watermarked images that have undergone semantic-altering manipulations (row 3).

T. Fernando, C. Fookes, and S. Sridharan are with The Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT), Queensland University of Technology, Australia.

makes them vulnerable, as perpetrators can easily detect and tamper with the watermark once its position is known. In addition, these methods often lack robustness against benign image manipulations such as brightness adjustments or heavy compression, leading to a high rate of false positives in deepfake detection. Furthermore, current supervised and adversarial learning approaches provide only limited fidelity in addressing the resilience–fragility paradox.

Most recently, OmniGuard [12] introduced a more flexible approach for image manipulation detection and localisation by incorporating a degradation-aware tamper extraction network designed for blind, robust copyright protection and tamper localization. However, the framework embeds watermarks directly in the pixel space, which makes them more susceptible to attacks. Even minor changes in image orientation or scale can significantly alter pixel values. As a result, pixel-based systems become highly sensitive to trivial modifications, even when the semantic content of the image remains unchanged. Consequently, the OmniGuard framework is less effective in detecting deepfakes, where the primary objective is to identify only malicious edits that alter the semantic-level meaning of the image.

In contrast to existing methods, we embed watermarks in the latent space rather than the pixel space, leveraging semantic-level manipulations of the image's representation. Since the latent space encodes high-level meaning, the watermark becomes tightly coupled with the image's semantics. Any malicious attempt to alter the image's meaning disrupts this coupling, effectively breaking the watermark and revealing tampering. Most importantly, through a learnable watermark embedder network, we identify less perceptually salient directions to embed information without significantly altering the human-perceived semantics of the input image. We propose to leverage a spherical latent space from CLIP [14] for the embedding of the proposed watermark. CLIP features lie on a unit hypersphere and offer normalised operations for the embedding of our watermark, constraining our watermark perturbations to keep features on this sphere. This prevents the watermark from introducing large deviations in the feature norm, thus preserving image quality. It effectively bounds the perturbation in a way that minor pixel shifts (like rotation, scale) that keep semantics unchanged will not move the feature off the sphere, whereas semantic changes will cause larger movements around the sphere.

Embedding the watermark in this space offers two key advantages: (1) inherent resilience to benign transformations, such as resizing, compression, or brightness adjustments, that do not significantly alter semantic features, and (2) heightened sensitivity to malicious manipulations that change the semantic meaning of the image, such as identity swaps or expression alterations in deepfakes. Latent space embedding reduces artefacts and ties the watermark to semantic content.

We propose a novel Adversarial Reinforcement Learning (ARL) paradigm for learning watermarks that are both resilient to benign manipulations and fragile to semantic alterations, enabling robust deepfake detection. The framework is structured as an adversarial game between two networks: the watermarker (i.e., the watermark embedder, extractor), and the attacker agent. The watermarking network aims to embed latent space watermarks that survive traditional image transformations but fail under semantic changes. The attacker, in turn, seeks to break the watermark by maximizing extractor loss. It can dynamically adjust the strength and type of attacks, ranging from conventional (e.g., compression, resizing) to semantic (e.g., facial attribute editing via StyleGAN), and combine them into complex, evolving curricula. This adversarial setup forces the watermarker to learn a sophisticated strategy that balances robustness and fragility, avoiding both overly weak and excessively strong watermarking schemes, allowing us to arrive at the right combination for this paradoxical objective. The result is an architecture that achieves unprecedented levels of resilience across bonafide transformations and greater levels of fragility towards malicious manipulations. Fig. 1 illustrates examples of our method's performance: row 1 shows genuine images, row 2 shows those images after numerous learned attacks against watermarking were applied. In both cases, the watermark remains intact. Row 3 shows images after malicious semantic edits where the watermark fails, and the system declares them fake.

The main technical contributions of this paper, through which we introduce the proposed *DeepForgeSeal* framework, can be summarised as follows:

1) We introduce a novel deep watermarking framework for deepfake detection that operates in the high-dimensional semantic space of input images.
2) We design a learnable watermarker with explicit control over message encoding and extraction, achieving semantic stealth.
3) We propose a new ARL (Adversarial Reinforcement Learning) paradigm that enables the agent to learn a strategy balancing watermark resilience and fragility.
4) We develop a reward function that guides the watermark attacker to discover failure regions in the latent space and image manipulations that induce significant latent shifts, allowing it to devise a structured curriculum of attacks.

II. Related Work

A. Watermarking Approaches for Deepfake Detection

Passive deepfake detection refers to techniques that analyse the media without requiring any prior knowledge of the generation process or embedding of security features such as watermarks. These approaches have demonstrated remarkable performance over the past decade, primarily by identifying artefacts introduced during the deepfake generation process. Examples include unsynchronised lip movements [15], irregular facial landmarks [16], or physiological inconsistencies like abnormal blood oxygen concentration [17]. Other forensic methods extend this idea by using deep neural networks to recognise the sequence of image manipulation operations applied to media [18]. However, their effectiveness is inherently tied to the presence of such artefacts, which vary depending on the generation technique used. Consequently, the generalisability of these detectors to novel or unseen deepfake generation methods remains limited. Moreover, the

rapid evolution of deepfake technologies poses a significant challenge to maintaining detection robustness, raising concerns about the long-term reliability of passive detection strategies. Such limitations have led to the development of proactive deepfake detection approaches, with watermarking being a prominent example. This technique embeds invisible signals into benign images before any manipulation occurs. These signals act as verification markers, enabling systems to actively determine whether the content has been tampered with based on the presence and integrity of the embedded watermark.

One of the earliest works in watermarking based deepfake detection can be attributed to [19], which embeds a robust, invisible watermark into facial images before any manipulation occurs. These watermarks are resilient to various deepfake transformations and image processing operations. The embedded watermark acts as a unique identifier, allowing the origin of a manipulated image to be traced. The works [20], [21] have also leveraged the concept of robust watermarking as a line of defense against deepfakes. Specifically, in contrast to [19], which embeds identity-linked tags using a learned network which is resilient to deepfake transformations, [20], [21] introduce training-free watermarking mechanisms based on facial landmarks. For example, instead of relying on learned embeddings, the authors of [21] propose to transform structural facial representations into binary perceptual watermarks, enabling robust and imperceptible embedding without extensive model training. Beyond face-specific designs, the broader steganography literature has explored optimising watermark embedding for improved imperceptibility and resilience. Liao et al. [22], [23] propose adaptive payload partitioning strategies across colour channels and multiple images based on texture complexity, enhancing robustness against statistical detection. In addition, recent work has investigated watermark robustness under extreme real-world distortions; for example, partial screen-shooting watermarking schemes embed redundant watermarks across spatially distributed regions to remain detectable even when only a portion of the image is captured [24].

Despite these advances, robust watermarking-based deepfake detection approaches require validating the embedded watermark against a known source or identity, which can be logistically complex in real-world applications, such as in social media platforms. Each image or video must be checked to ensure the watermark matches a trusted source identity, which assumes access to and maintenance of such a database.

Semi-fragile watermarks, which leverage both the fragility and robustness of watermarking, simplify the detection process by serving as self-contained integrity checks. Specifically, if the watermark is broken or unreadable, it directly signals tampering, without needing to compare against a source. This makes semi-fragile watermarking approaches more efficient and scalable for deployment in environments where rapid and autonomous verification is critical. One of the early works in semi-fragile watermarks for deepfake detection can be attributed to FaceGuard [25], which encodes a semi-fragile identity-specific watermark. Along similar lines, Zhao et al. [9] proposed an identity-based semi-fragile watermarking approach that is sensitive to face swap manipulations while remaining resilient to benign image transformations. Despite its advances, the reliance on identity-specific watermarks for each individual makes these techniques less appealing for real-world applications in which the true identity of the individuals is not available or cannot be readily extracted. WaterLo [26] addresses these limitations by introducing a localised, fixed semi-fragile watermark that disappears in manipulated regions, allowing not only detection but also precise localisation of tampered areas. Following this approach, the authors of FaceSigns [27] have introduced a semi-fragile watermarking framework that embeds a 128-bit secret message directly into the image pixels.

Most recently, the EditGuard [13] introduces a novel approach for decoupling the training process from specific tampering types. Traditional watermarking systems often require retraining or fine-tuning for each manipulation scenario. In contrast, EditGuard's image-to-image steganography framework generalises across diverse tampering methods, including face swaps, background replacements, and AI-generated edits, without needing retraining. OmniGuard [12] further enhances this approach by introducing a hybrid forensic framework that combines proactive embedding of the semi-fragile watermark with passive blind extraction. In a different line of work, the authors of [10] leverage the mathematical properties of fractals to generate watermark patterns through a parameter-driven pipeline. Moreover, watermarks are embedded using an entry-to-patch strategy, where each watermark matrix entry is mapped to a specific image patch, enabling precise localisation of manipulated regions. Though significant strides have been made, the majority of existing semi-fragile watermarking approaches still rely on the pixel space for embedding, rather than leveraging the latent space, which could offer greater flexibility and robustness. Furthermore, current adversarial training pipelines used when training the watermarking frameworks are often simplistic and lack the sophistication needed to simulate complex attack curricula. As a result, the resilience of watermarking techniques under realistic and increasingly sophisticated adversarial conditions remains limited. In contrast, our framework leverages the latent space to obtain more structured and disentangled representations. This enhances watermark embedding consistency, robustness, and sensitivity to manipulations. Additionally, the integration of adversarial reinforcement learning enables the system to simulate and defend against complex, curriculum-based adversarial attacks. This significantly improves the robustness and generalisability of the watermarking strategy across diverse manipulation scenarios.

B. Adversarial Reinforcement Learning

In a learning setting, the relationships among agents can be categorised into cooperative, competitive, and both [28]. However, most of the prior works [29], [30], [31], [32], [33] in Multi Agent Reinforcement Learning (MARL) have focused on cooperative tasks, in which the cumulative reward is maximized as a group. Among the limited number of works extending traditional MARL into adversarial settings, researchers have introduced adversarial elements, either as

competing agents or perturbations, to enhance the robustness and generalisation capabilities of the overall algorithm. For example, in [34] the authors have shown the potential of learning a robust generalised policy using Multi Agent Adversarial Reinforcement Learning (MAARL), whereas the authors of [35] have used MAARL to avoid overfitting. In a similar line of work, Bukharin et. al [36] address the lack of robustness and sensitivity to environment changes in MARL through adversarial training. Specifically, the authors of [36] introduce adversarial regularization to enforce Lipschitz continuity in policies, improving robustness against noisy observations. Most recently, Yuan et. al [37] have proposed an approach based on evolutionary learning to enhance robustness in message-passing for improving the communication efficiency of agents.

While adversarial training has shown promise in both reinforcement learning and watermarking independently, to the best of our knowledge, none of the prior works have investigated the integration of Adversarial Reinforcement Learning (ARL) into watermarking. Through empirical evaluations, we demonstrate that ARL offers a compelling framework for training watermarking agents that can dynamically adapt to image manipulations. This enables watermarking systems that are both robust and semi-fragile, providing resistance to benign transformations while remaining sensitive to malicious tampering.

III. Methods

In this section, we discuss our proposed approach. We first provide an overview of the main components that constitute our DeepForgeSeal framework in Sec. III-B. Our watermarker, watermark attacker, and watermark extraction and deepfake detection processes are discussed in Secs. III-C, III-D, and III-E, respectively. The implementation details of the framework are provided in the supplementary material.

A. Problem Formulation

Let x denote an input image with CLIP semantic feature $f(x) \in \mathbb{R}^{\zeta}$, $\|f(x)\|_2 = 1$, and let K be a shared secret key from which a watermark message M is deterministically generated. We seek a watermarker $\theta = \{\mu, \phi, \pi\}$ that embeds M in the latent space to produce a watermarked image $x' = \pi(f(x)+p)$, an attacker η that maps x' to an adversarial image x_a, and an extractor δ that recovers a message M' from any query image x''. The objective is semi-fragility: M must be recoverable under benign transformations (low BER) yet fail under malicious semantic edits (high BER). At verification, integrity is decided by $f_{\text{DEEPFAKE}}(x'') = \left[\text{BER}(M', M) > \lambda\right]$ (Eq. 19), where λ is a predefined threshold. The full procedure is summarized in Alg. 1. Furthermore, a summary of notation is provided in Table VI in the supplementary material.

B. Framework Overview

Given an input image x, we define a watermarker θ that learns the distribution of the semantic features $f(x)$ in a latent semantic space $\mathbb{S}$. The watermarker inserts a watermark M into the latent space, ensuring that it does not significantly alter the semantic meaning of x, and decodes the new image x' with the watermark from the latent space back to the image space. The objective of the watermarker is to identify the region in $\mathbb{S}$ where it could embed the watermark without significantly altering the human-perceived meaning of x. The watermarker must also learn to maintain an optimal balance between resilience and fragility. Specifically, the embedded watermark should remain robust to benign operations such as JPEG compression, cropping, and color adjustments, while being fragile to malicious edits, such as face swapping or facial attribute editing, that alter the semantic content of the image.

The objective of the watermark attacker agent, η, is to generate an adversarial image x_a from the watermarked image x', i.e., $\eta(x') \rightarrow x_a$, by applying transformations that degrade or remove the embedded watermark. To achieve this, the attacker can choose from a range of image manipulations, including benign operations and malicious edits. These manipulations may also be combined, and for each type, the agent can control the associated parameters to vary the strength of the attack. The watermarker is trained via supervised losses, not via an RL algorithm, whereas the attacker agent is trained via reinforcement learning to maximise its reward.

The watermark extractor, δ, is responsible for retrieving the embedded watermark from a given image x'', which may or may not carry a watermark. Since the watermarker θ dynamically adjusts the location of the watermark, the extractor must co-adapt and learn in tandem with the watermarker to ensure reliable extraction. In our framework, the success or failure of watermark extraction serves as a heuristic for deepfake detection. Specifically, if the extractor δ fails to recover a valid watermark from the image x'', the image is flagged as a potential deepfake. Our approach leverages the learned consistency between the watermarking and extraction processes as an indirect signal of authenticity, enabling the detection of semantic-level manipulations without relying on a fixed watermark pattern. These modules, and the flow of information between them, are illustrated in Fig. 2.

C. Watermarker

The watermarker is responsible for encoding and embedding the watermark into the image. Its architecture is designed to operate within the semantic latent space of a pre-trained CLIP model, ensuring that the embedding process is guided by high-level image representations. The procedure comprises two key stages: (i) directional embedding in the latent sphere, and (ii) perturbative watermark embedding. These stages are discussed in detail in the following sub-sections.

1) Directional Embedding in Latent Sphere: To extract semantically meaningful information from the input image, x, our watermarker, θ, operates in the feature space of the CLIP image encoder [14], which is denoted as E_{CLIP}. Specifically, the feature vector $f(x) = E_{\text{CLIP}}(x) \in \mathbb{R}^{\zeta}$ is L2-normalized, such that $\|f(x)\|_2 = 1$, constraining it to the surface of a ζ-dimensional hypersphere. The watermarker has explicit control over the embedding directions that it leverages to embed the message (i.e., watermark), M. However, allowing

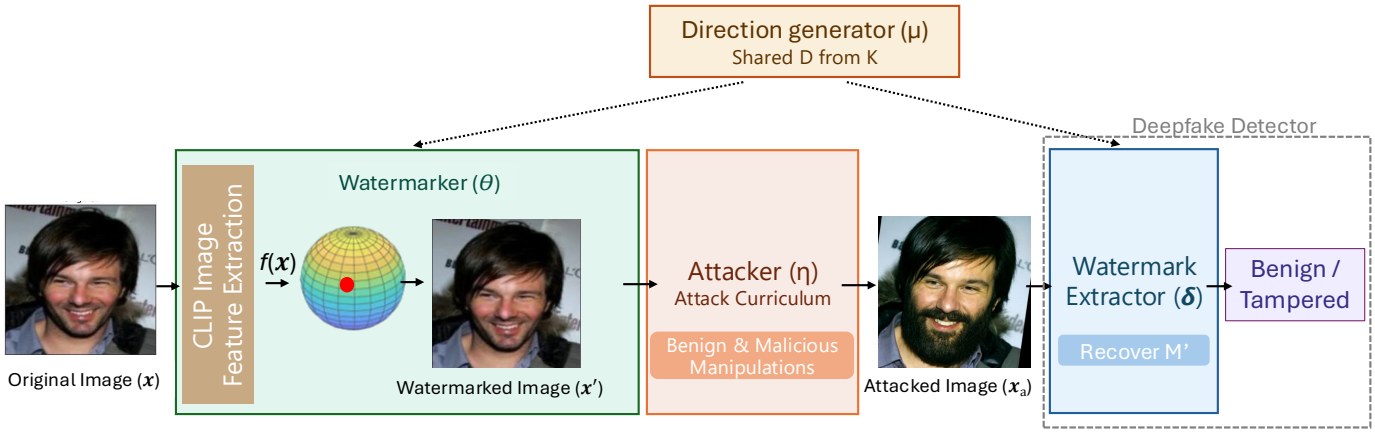

Fig. 2. Method Overview: Given an input image x, the watermarker θ embeds a watermark M into the latent space $\mathbb{S}$ derived from semantic (CLIP Image) features $f(x)$, producing a watermarked image x'. The direction generator μ generates this watermark from a secret key K, which is shared by both the watermarker θ and the extractor δ. An attacker η generates an adversarial image x_a to disrupt watermark integrity. η uses both benign edits (e.g., JPEG compression, cropping) and semantic-altering manipulations (e.g., face swaps) when generating its attack curriculum. The extractor δ attempts to recover M from any image x''; failure to extract a valid watermark flags x'' as a potential tampered image, leveraging watermark consistency as a proxy for semantic authenticity.

the watermarker to freely modify both the embedded message and the latent space directions makes it infeasible for the watermark extractor to recover the message without access to either the original image x or the watermark M. To address the challenge of extracting dynamically embedded messages, we propose deriving the latent space directions through a learnable **direction generator**, μ, while relying on a **shared secret key** K. The watermark message M is not a fixed bit string but is deterministically generated from a secret key. This design is analogous to cryptographic pseudo-random number generation, where a secret key is used as the seed to produce a reproducible message. Consequently, M does not need to be stored with each image; the secret key alone is sufficient for the detector to regenerate the watermark message during verification. The key is first projected into the CLIP text embedding space [14] to obtain a feature vector $f(K)$. The direction generator network μ then uses this feature to produce a canonical set of L direction vectors $D = \{d_1, d_2, \ldots, d_L\}$. Formally,

$$\begin{aligned} D &= \text{reshape}(\mu(f(K))) \in \mathbb{R}^{L\times\zeta}, \qquad \|d_i\|_2 = 1,\ \forall i, \\ P &= Df(x') \in \mathbb{R}^{L}, \qquad P_i = \langle f(x'), d_i\rangle. \end{aligned} \tag{1}$$

Since D depends only on K, it remains consistent across all images. However, the directions are still **learnable**, as the weights ω of μ are optimised during training. For each bit $m_i \in \{0,1\}$ of a dynamically generated message M, the objective is to modify the image x to produce x' such that the projection of its new feature vector $f(x')$ onto each direction d_i matches a target value,

$$\langle f(x'), d_i\rangle \approx P_i^{\text{target}} = \begin{cases} \xi_1 & \text{if } m_i = 1 \\ \xi_0 & \text{if } m_i = 0 \end{cases} \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product, and ξ_1, ξ_0 are predefined projection magnitudes that encode binary values.

2) Perturbative Watermark Embedding: Once the $m_i \in \{0,1\}$ of the message M have been identified, the watermarker employs an MLP layer, ϕ, that receives semantic features, $f(x)$, of x and the watermark, M, and synthesizes a perturbation vector p over ζ-dimensional feature vector $f(x)$ such that:

$$p = \phi([f(x); M]). \tag{3}$$

The perturbed latent code is then given by:

$$q = f(x) + p, \tag{4}$$

and a decoder π reconstructs the watermarked image x' from the perturbed representation:

$$x' = \pi(q). \tag{5}$$

The watermarker's objective is to minimize a composite loss function $\mathcal{L}_W$ that balances imperceptibility ($\mathcal{L}_{\text{clip}}$), embedding accuracy ($\mathcal{L}_{\text{dir}}$), and the conditional fragility objective ($\mathcal{L}_{\text{ext}}$). Formally, this can be written as:

$$\mathcal{L}_W = \alpha\mathcal{L}_{\text{clip}} + \beta\mathcal{L}_{\text{dir}} + \gamma\mathcal{L}_{\text{ext}}, \tag{6}$$

where,

$$\mathcal{L}_{\text{clip}} = \|f(x) - f(x')\|_2^2 \tag{7}$$

$$\mathcal{L}_{\text{dir}} = \text{CrossEntropy}(M, \varpi), \tag{8}$$

$$\mathcal{L}_{\text{ext}} = -\text{BER}(M, M'), \tag{9}$$

where M' is the recovered message obtained by decoding the projection vector $P = \{\langle f(x'), d_i\rangle\}_{i=1}^{L}$, ϖ denotes predicted message logits by the watermark extractor, from which the final binary message M' is recovered, and BER(.,.) is Bit Error Rate. See Sec. III-E for details.

D. Watermark Attacker

The objective of the watermark attacker, η, is to learn an optimal policy to destroy the embedded watermark. One of our key innovations is its ability to compose complex attacks using

a predefined set of benign/traditional image transformations and malicious/generative transformations. This enables the agent to learn a dynamic attack curriculum that jointly models both the type and intensity of each attack. The following subsections illustrate: (i) the process of composing a combinatorial attack, (ii) how the watermark attacker agent is trained, and (iii) how it is incentivized to identify attack curricula that lead to substantial semantic drifts and are closer to historically known failure regions in the latent space, where the watermark extractor fails to recover the watermark.

1) Composing a Combinatorial Attack: Let the set of available attacks, including the benign and malicious transformations, be denoted as $\mathcal{A} = \{a_1, \ldots, a_\Gamma\}$. The watermark attacker agent, η, receives the CLIP features, $f(x')$, of the watermarked image, x', as input and outputs two vectors: logits for attack selection and normalized strength parameters τ as:

$$(\text{logits}, \tau) = \eta(f(x')). \tag{10}$$

These logits are passed through a sigmoid function (denoted by σ to convert them into probabilities:

$$B = \sigma(\text{logits}) \in [0,1]^\Gamma. \tag{11}$$

Then, for each probability, B_l , we sample a binary action, a_l, from a Bernoulli distribution as:

$$a_l \sim \text{Bernoulli}(B_l) \quad \forall l \in \{1, ..., \Gamma\} \tag{12}$$

The result is a binary action vector $a \in \{0,1\}^\Gamma$ in which each element is independently sampled, creating a **combinatorial attack strategy**.

Simultaneously, the strength vector $\tau \in [0,1]^\Gamma$ determines the intensity for each attack. Each normalized value τ_l is mapped to a meaningful parameter, param_l, in the selected attack type, a_l, and within a predefined range $[\min_l, \max_l]$ as:

$$\text{param}_l = \min_l + (\max_l - \min_l) \cdot \tau_l. \tag{13}$$

2) Learning an Attack Curriculum: We train our watermark attacker agent using **reinforcement learning** to learn an optimal policy ψ that maximizes the likelihood of watermark extraction failure, thereby ensuring the success of the attack. Specifically, the primary reward R for the watermark attacker agent can be formulated as the failure of the Watermark Extractor, measured by the Binary Cross-Entropy (BCE) loss between the extracted message $\tilde{M}$ and the original message M as:

$$R_{\text{failure}} = \text{BCE}(\tilde{M}, M). \tag{14}$$

A strong and adaptive attack policy would drive the watermarker to learn more resilient and sophisticated watermarking strategies. In contrast, the simplistic objective in Eq. 14, which focused solely on extractor failure, lacks the distinction required to drive meaningful improvements. As such, an advanced reward signal formulation is required to incentivize the attacker to pursue diverse and exploratory policies that provide richer adversarial signals.

3) Promoting Semantic Drift and Targeted Shifts Toward Known Failure Regions: We propose to augment the reward with bonuses to encourage exploration and exploitation. Specifically, we define a **curiosity reward** which is proportional to the squared Euclidean distance between the CLIP features of the image before and after the attack. This could be written as:

$$R_{\text{curiosity}} = \Delta \|f(x') - f(x_a)\|_2^2, \tag{15}$$

where x_a is the attacked image and Δ is a scaling factor. This reward incentivizes the attacker to discover attacks that cause the most significant semantic disruption, which are often the most challenging for a watermark to survive.

In addition, we introduce a **proximity reward** that encourages the attacker to perturb the image toward regions of the latent space known to challenge the watermark extractor. In particular, we introduce a memory buffer, J, that stores the latent representations, $f(x_a)$, of images that have historically caused failures in watermark extraction. Then, the potential, ρ, of a new attacked image x_a can be defined as its proximity to its closest feature vector stored in the failure memory. Now the attacker's proximity reward can be defined as inversely proportional to the distance of its attacked image from the closest failure memory embedding. Formally, this could be written as:

$$\begin{aligned} \rho(x_a) &= \min_{f_j \in J} \|f(x_a) - f_j\|_2, \\ R_{\text{proximity}} &= \frac{\nu}{\rho(x_a) + \epsilon}, \end{aligned} \tag{16}$$

where ν is a scaling factor and ϵ is a small constant to ensure numerical stability.

Now, the final objective of the watermark attacker agent can be written as:

$$\begin{aligned} \mathcal{L}_\eta = -\mathbb{E}_{a\sim\psi}\Big[R_{\text{failure}} + R_{\text{proximity}} + R_{\text{curiosity}} - o\sum_{l=1}^{\Gamma} a_l\Big] \\ -r\sum_{l=1}^{\Gamma} H(B_l) \end{aligned} \tag{17}$$

where $\sum_{l=1}^{\Gamma} a_l$ is a small regularization term penalizing the number of active actions, $H(B_l)$ denotes the entropy of the Bernoulli action distribution for the l^{th} attack, and o and r are hyperparameters. It should be noted that $R_{\text{curiosity}}$ encourages goal-directed exploration while $\sum_{l=1}^{\Gamma} H(B_l)$ is used to encourage policy-level exploration such that the agent maintains diversity in action selection, serving complementary purposes in the learning process.

E. Watermark Extraction and Deepfake Detection

The watermark extractor, δ, is responsible for recovering the message M from a given image x'', which may or may not carry a watermark and could potentially be corrupted. Specifically, given the input x'', it first extracts CLIP features $f(x'')$, and independently generates the direction vectors, D, from the secret key, K, using the direction generator, μ. Then

it produces the projection of $f(x'')$ onto each of the direction vector $d_i \in D$ as:

$$P_i'' = \langle f(x''), d_i \rangle. \tag{18}$$

The resulting vector of projections $P'' = \{P_1'', ..., P_L''\}$ is passed through an MLP, which outputs the message logits ϖ, from which the recovered binary message M' is obtained.

In the proposed framework, the fragility of the watermark serves as the deepfake detection mechanism. Given the direction generator, μ, and the shared secret key K, the integrity of an input image can be verified at inference time by computing the Bit Error Rate (BER) between the extracted message M' and the known original message M regenerated from the shared secret key, K. We assume that the watermark embedder and the extractor share a secret key K securely, which is a common assumption in watermarking schemes [38]. Since the same trained direction generator μ and secret key K are used during both watermark embedding and extraction, the resulting direction vectors should remain consistent. A high BER indicates potential generative manipulation, enabling us to leverage the fragility of the watermark as a deepfake detection signal. Formally, this could be written as:

$$f_{\text{DEEPFAKE}}(x'') = \begin{cases} \text{True}, & \text{if BER}(M', M) > \lambda \\ \text{False}, & \text{otherwise} \end{cases} \tag{19}$$

where λ is a predefined classification threshold.

Algorithm 1 DeepForgeSeal: Training and Inference

Require: Image set $\mathcal{X}$; secret key K; CLIP image/text encoders; failure memory $\mathcal{J} \leftarrow \emptyset$
Require: Modules: direction generator μ, perturbation MLP ϕ, decoder π, extractor δ, attacker η
// Training
for each training iteration **do**
  Sample $x \in \mathcal{X}$; $f(x) = E_{\text{CLIP}}(x)$, $\|f(x)\|_2 = 1$
  Project K to CLIP-text space to get $f(K)$; $D = \text{reshape}(\mu(f(K)))$ ▷ Eq. 1
  Generate message M from K; set targets P^{target} ▷ Eq. 2
  $p = \phi([f(x); M])$; $q = f(x) + p$; $x' = \pi(q)$ ▷ Eqs. 3–5
  $(\text{logits}, \tau) = \eta(f(x'))$; $B = \sigma(\text{logits})$; $a_l \sim \text{Bernoulli}(B_l)$ ▷ Eqs. 10–12
  Map strengths param_l; apply selected attacks sequentially: $x_a = \eta(x')$ ▷ Eq. 13
  Recover message(s) via δ from x' and x_a ▷ Eq. 18
  Compute $\mathcal{L}_W = \alpha\mathcal{L}_{\text{clip}} + \beta\mathcal{L}_{\text{dir}} + \gamma\mathcal{L}_{\text{ext}}$ ▷ Eqs. 6–9
  Update $\{\mu, \phi, \pi, \delta\}$ by descending $\nabla\mathcal{L}_W$ ▷ supervised
  Compute $R_{\text{failure}}, R_{\text{curiosity}}, R_{\text{proximity}}$ ▷ Eqs. 14–16
  Update η via REINFORCE on $\mathcal{L}_\eta$ ▷ Eq. 17
  if watermark extraction fails on x_a **then** add $f(x_a)$ to $\mathcal{J}$
  end if
end for
// Inference (deepfake detection)
Require: Query image x''; secret key K; threshold λ
$f(x'') = E_{\text{CLIP}}(x'')$; regenerate D and M from K
$P_i'' = \langle f(x''), d_i \rangle$; $\delta \to \varpi \to M'$ ▷ Eq. 18
return DEEPFAKE $= \big[\,\text{BER}(M', M) > \lambda\,\big]$ ▷ Eq. 19

IV. Experiments

In this section, we present the results of our experiments designed to evaluate and compare the effectiveness of the proposed DeepForgeSeal framework against existing state-of-the-art methods. We begin by detailing the datasets used in our evaluations (Sec. IV-A), followed by a description of the evaluation metrics employed to assess model performance (Sec. IV-B). The main experimental results, where we benchmark our method against current state-of-the-art approaches, are provided in Sec. IV-C. Additionally, ablation studies conducted to validate the contributions of the learnable directional embedding strategy, adversarial learning pipeline, and novel reward function are discussed in Sec. IV-D. For implementation details, hyperparameter evaluations, qualitative evaluations, time complexity analysis, and additional ablation studies, please refer to the supplementary materials.

A. Datasets

Following prior works [9], we use the Flickr-Faces-HQ (FFHQ) dataset [39] to train our model. The CelebA [40] and CelebA-HQ [41] datasets are used for evaluation and to demonstrate generalisability. In particular, our method is trained solely on the FFHQ dataset and exposed only to face swapping and attribute-level manipulations (i.e., change hair color, expression, and age) as defined in [42]. Therefore, evaluation on the unseen CelebA and CelebA-HQ datasets demonstrates a strict cross-dataset generalization.

The FFHQ dataset contains 70,000 images at a resolution of 1024 × 1024 pixels. In accordance with the dataset authors' guidelines, the first 60,000 images are designated for training, while the remaining 10,000 are reserved for validation. We follow this protocol and use the training subset of FFHQ to train our model. CelebA comprises 10,000 distinct identities, with each identity represented by 20 images at a 128x128 resolution. CelebA-HQ is a high-quality version of the CelebA dataset, consisting of 30,000 images at a resolution of 1024 × 1024 pixels. Specifically, we followed the official test split provided by the CelebA and CelebA-HQ dataset authors to enable direct comparisons.

B. Evaluation Metics

We evaluate our method from two key perspectives. First, we assess the visual fidelity of the watermarked images compared to the original images using two widely adopted metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR provides a quantitative measure of the pixel-level differences, where higher values indicate less distortion and better preservation of image quality. SSIM, on the other hand, evaluates perceptual similarity by considering factors such as luminance, contrast, and structural information, offering a more human-aligned assessment of image quality. Together, these metrics help us determine how effectively our method maintains the integrity of the original image while embedding the watermark.

To evaluate the performance of deepfake detection, following prior works [9], we compute image-level Accuracy (ACC) and F1-Score. Accuracy reflects the overall proportion of correctly classified images, encompassing both correctly identified real and fake samples. The F1-Score, which balances precision and recall, is particularly informative in capturing

the trade-off between false alarms (incorrectly classifying real images as fake) and miss detections (failing to identify fake images). Together, these metrics provide a reliable measure of the system's ability to detect deepfakes while minimizing classification errors. In addition, for the proposed DeepForgeSeal method and the considered proactive baselines, we report the Area Under the Receiver Operating Characteristic Curve (AUC) metric.

C. Comparisons with Existing State-of-the-art Methods

In this section, we report results for the proposed model and compare with the existing state-of-the-art methods considering the visual quality of the watermarking (See Sec. IV-C1), deepfake detection performance (See Sec. IV-C2), and watermark's resilience and fragility (See Sec. IV-C3).

1) Visual Quality of Watermarking Approaches: As baselines we use state-of-the-art methods, including FaceSigns [27], EditGard [13], and Zhao et. al [9]. When choosing our baselines, we ensured coverage of a diverse range of deep learning approaches. These include adversarially trained models [27], dual-purpose watermarking methods for tamper detection and copyright protection [13], and identity-entangled watermarking methods [9], enabling a comprehensive comparison.

TABLE I
VISUAL FIDELITY OF DIFFERENT WATERMARKING TECHNIQUES. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	Quality Metrics	
	PSNR (↑)	SSIM (↑)
FaceSigns [27]	32.03	0.92
Zhao et. al [9]	38.32	0.94
EditGard [13]	45.07	0.96
Omniguard [12]	46.15	**0.97**
VINE [43]	46.83	**0.97**
DeepForgeSeal (ours)	**48.39**	**0.97**

Quantitative comparisons for the CelebA-HQ dataset are presented in Tab. I. When comparing the proposed DeepForgeSeal with the previous state-of-the-art methods, we observe that our method has the highest PSNR and SSIM, demonstrating our framework's ability to embed watermarks stealthily, preserving the image's visual quality. We attribute this to operating within the high-dimensional semantic space of the input images, which enables subtle yet effective modifications. Notably, these results are obtained using only the CLIP feature-similarity loss L_clip to supervise the decoder π, without any explicit pixel-domain or perceptual reconstruction loss. In addition to the above quantitative comparisons, we provide qualitative comparisons in the supplementary materials.

2) Deepfake Detection Performance: We compare the proposed DeepForgeSeal method with both passive and proactive deepfake detection methods. Following [9], we employ the CelebA and CelebA-HQ datasets to represent low and high-resolution real image cases. Deepfakes are generated using a variety of deepfake generation methods, including AttGAN, StarGAN, InfoSwap and SimSwap, CIAGAN, and DeepPrivacy. These methods allow us to evaluate the performance of the proposed deepfake detection framework against a wide range of malicious manipulations, including identity swapping and face anonymization techniques. In our experimental pipeline, for passive deepfake detection methods, we use non-watermarked images from the CelebA and CelebA-HQ datasets as real images. The above deepfake generation techniques are then applied to create their fake counterparts. For proactive deepfake detection methods, including the proposed DeepForgeSeal approach, we first watermark the images from the CelebA and CelebA-HQ datasets, and then generate the corresponding fake images using the same deepfake generation methods.

Following [9], as baselines for passive deepfake detection we use the state-of-the-art BTS [44], CD [45], Bonettini et. al [46], PF [47], RFM [48], SBI [49], FakeShiled [50], and CBO-DD [51]. In addition, the state-of-the-art FaceSigns [27], EditGard [13], Zhao et al. [9], Omniguard [12], and VINE [43], which are proactive deepfake detection methods, have also been used for comprehensiveness. Results of all passive baselines except FakeShiled [50], and CBO-DD are obtained from Zhao et al. [9] under their stated evaluation protocol, i.e., using pre-trained models provided by the authors of the passive baseline models, without retraining or threshold tuning on the test data. FakeShiled and CBO-DD models were re-trained using the exact hyperparameters and thresholds provided by the authors to ensure reproducibility of the results. For all proactive deepfake detection baseline methods, we re-evaluated the baselines using the provided public implementations and the thresholds.

Tab. II summarises the comparison results for deepfake detection in CelebA and CelebA-HQ datasets. It is evident that passive deepfake detection methods generally struggle to generalise across different deepfake generation techniques. Among the approaches evaluated, only methods such as SBI [49] and CBO-DD [51] consistently demonstrate strong performance across all considered generation methods. In contrast, most other passive detection techniques perform well against a limited subset of generation methods but show significant performance degradation when applied to others. This limitation stems from their training approach, which focuses on identifying artefacts specific to the generation techniques used during training. Consequently, these methods tend to perform poorly when tested on deepfakes produced by unfamiliar or diverse generation techniques.

When comparing the results of the proactive detection methods FaceSigns [27], Zhao et. al [9], EditGard [13], Omniguard [12], and VINE [43] with the proposed DeepForgeSeal method, we see that our method has been able to achieve significant detection accuracy across all considered generation methods. Although state-of-the-art methods such as Zhao et. al [9], EditGard [13], Omniguard [12], and VINE [43] have been able to achieve impressive performance against deepfake generation methods such as AttGAN and DeepPrivacy, their performance degrades when tested against DeepPrivacy and StarGAN2 generation methods. In contrast, our proposed DeepForgeSeal method consistently demonstrates superior detection performance across all evaluated deepfake generation techniques. We attribute this to the learnable watermarking strategy embedded within the framework,

TABLE II
ACCURACY (↑) AND F1 SCORES (↑) OF DIFFERENT METHODS' DEEPFAKE DETECTION RESULTS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Detection methods	Low Resolutions (CelebA)						High Resolutions (CelebA-HQ)					
	AttGAN	CIAGAN	DeepPrivacy	InfoSwap	SimSwap	StarGAN2	AttGAN	CIAGAN	DeepPrivacy	InfoSwap	SimSwap	StarGAN2
BTS [44]	0.86/0.87	0.51/0.66	0.50/0.66	0.49/0.66	0.51/0.67	0.53/0.67	0.86/0.87	0.50/0.66	0.50/0.66	0.49/0.65	0.50/0.66	0.55/0.68
CD [45]	0.88/0.86	0.51/0.03	0.51/0.01	0.54/0.17	0.51/0.01	0.78/0.71	0.81/0.77	0.51/0.04	0.52/0.07	0.52/0.07	0.52/0.06	0.84/0.81
Bonettini et. al [46]	0.59/0.69	0.62/0.62	0.58/0.68	0.57/0.64	0.60/0.71	0.63/0.63	0.53/0.68	0.65/0.62	0.56/0.69	0.51/0.66	0.55/0.69	0.49/0.65
PF [47]	0.76/0.79	0.51/0.66	0.52/0.65	0.57/0.68	0.54/0.67	0.99/0.98	0.75/0.79	0.51/0.66	0.55/0.68	0.56/0.69	0.54/0.68	**0.98/0.97**
RFM [48]	0.50/0.67	0.51/0.67	0.51/0.67	0.50/0.67	0.51/0.67	0.50/0.67	0.50/0.67	0.50/0.67	0.51/0.67	0.50/0.67	0.50/0.67	0.50/0.67
SBI [49]	0.79/0.82	0.77/0.8	0.78/0.82	0.77/0.81	0.78/0.81	0.72/0.75	0.80/0.80	0.72/0.78	0.78/0.80	0.76/0.77	0.83/0.84	0.69/0.70
FakeShield [50]	0.84/0.83	0.92/0.93	0.85/0.86	0.78/0.85	0.84/0.87	0.83/0.83	0.83/0.85	0.91/0.93	0.84/0.84	0.81/0.83	0.88/0.90	0.81/0.83
CBO-DD [51]	0.84/0.84	0.91/0.93	0.86/0.87	0.79/0.85	0.84/0.88	0.85/0.85	0.83/0.86	0.92/0.93	0.86/0.87	0.80/0.85	0.89/0.91	0.82/0.84
FaceSigns [27]	0.91/0.93	0.83/0.84	0.91/0.95	0.93/0.95	0.95/0.95	0.96/0.98	0.87/0.88	0.92/0.93	0.80/0.83	0.91/0.91	0.82/0.82	0.84/0.87
Zhao et. al [9]	0.94/0.94	0.87/0.86	0.98/0.98	0.98/0.98	0.97/0.98	0.82/0.84	0.94/0.94	0.85/0.82	0.99/0.98	**0.99/0.99**	0.98/0.98	0.85/0.87
EditGard [13]	0.96/0.97	0.88/0.88	0.96/0.98	0.94/0.97	0.97/0.98	0.91/0.93	0.95/0.96	0.87/0.91	0.97/0.99	**0.99/0.99**	0.98/0.99	0.93/0.95
Omniguard [12]	0.96/0.97	0.87/0.87	0.96/0.98	0.93/0.91	0.97/0.98	0.92/0.94	0.95/0.96	0.88/0.92	0.97/0.98	**0.99/0.99**	**0.99/0.99**	0.94/0.95
VINE [43]	0.95/0.96	0.89/0.88	0.97/0.98	0.95/0.96	0.97/0.97	0.96/0.97	0.96/0.97	0.88/0.92	0.98/0.99	0.98/0.99	**0.99/0.99**	0.95/0.95
DeepForgeSeal (Ours)	**0.96/0.98**	**0.92/0.95**	**0.99/0.99**	**0.98/0.99**	**0.98/0.98**	**0.99/0.99**	**0.96/0.98**	**0.95/0.96**	**0.99/0.99**	**0.99/0.99**	**0.99/0.99**	**0.98/0.97**

which enables explicit control over message encoding and extraction. Additionally, the adversarial reinforcement learning pipeline allows the watermarking agent to actively observe and respond to malicious attacks, dynamically adapting its watermarking strategy. Together, these components contribute to DeepForgeSeal's strong generalisation capability across a wide range of deepfake generation methods. In addition to these comparisons, we compare the proposed method and the considered proactive baselines using the AUC metric, which is available in the supplementary material.

3) Watermark Resilience and Fragility Against Attacks: In this section, we compare the resilience of our DeepForgeSeal watermarking methods with the current state-of-the-art methods under different attacks against watermarks. For comprehensiveness, we consider both traditional benign attacks, such as cropping, compression, and color changes, as well as the learnable benign attacks proposed by Zhao et. al [52] and Lukas et. al [53]. In addition, to quantitatively assess the fragility against malicious image manipulations, we generate attacks using state-of-the-art GAN-based face swapping and reenactment techniques, including SimSwap [54], UniFace [55], FaceDancer [56], and HyperReenact [57].

The evaluation results are presented in Tab. III. As baselines, we use the state-of-the-art FaceSigns [27] and EditGard [13] semi-fragile watermarking methods. Following prior works [27], [13], we measure the Bit Recovery Accuracy (BRA) as the evaluation metric, where we expect a higher BRA against benign and lower BRA against malicious transforms to achieve the goal of semi-fragile watermarking. As reported in Tab. III, we observe that both baseline models perform poorly against benign image manipulations, demonstrating that these watermarking techniques are too fragile to withstand the benign image transformations that users can apply in the real world. At the same time, both baseline techniques demonstrate less fragility towards the malicious, content-altering manipulations that the considered face swapping and reenactment techniques have generated. These evaluations clearly demonstrate the limitations of the existing state-of-the-art watermarking techniques and the superiority of the DeepForgeSeal paradigm that enables us to learn a strategy that optimally balances watermark resilience and fragility.

D. Ablation Evaluations

We conducted a series of ablation studies to systematically analyse the impact of the individual innovations that our DeepForgeSeal framework proposes. Several design choices contribute to the robustness of our model: i) the proposed learnable directional embedding strategy that uses a latent sphere; ii) the proposed adversarial learning pipeline with composed attacks with a learnable attack curriculum; and iii) the novel reward function that promotes semantic drift and targeted shifts toward known failure regions. All ablation experiments were conducted on the CelebA-HQ dataset, and for testing the ablation models, we used the validation set of CelebA-HQ. As evaluation metrics, we use PSNR and average BRA against benign and malicious manipulations.

1) Effects of Learnable Directional Embedding in Latent Sphere: To study the effect of the proposed learnable directional embedding strategy that uses a latent sphere, we generated two ablation variants of the proposed DeepForgeSeal model: i) DeepForgeSeal - w/o [D] - a model with naive latent embedding in which we embed the message directly in the latent space without directional encoding, and ii) DeepForgeSeal - FD - a model with fixed directional embeddings.

Results of this comparison are shown in Tab. IV. We observe that the learnable directional embedding strategy is an integral part of the proposed framework. Specifically, we observe that the naive latent embedding approach, where the message is directly embedded into the latent space, leads to a reduction in PSNR, distorts the natural image characteristics, and diminishes the watermark's resilience against benign image transformations. The introduction of directional embedding enhances the robustness of the watermark (see the row corresponding to DeepForgeSeal - FD). Notably, the experimental results demonstrate that the learnable directional embedding strategy, which encodes the message within a latent sphere, achieves the most favorable balance between resilience and fragility, while preserving image fidelity. These experiments confirm the utility of our learnable directional embedding procedure.

TABLE III
BIT RECOVERY ACCURACY (BRA) OF BASELINE FACESIGNS [27] AND EDITGARD [13] WATERMARKING METHODS AND PROPOSED DEEPFORGESEAL METHOD UNDER BENIGN AND MALICIOUS TRANSFORMATIONS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	BRA (%) - Benign Transforms (↑)							BRA (%) - Malicious Transforms (↓)			
	JPG	Noise	Crop	Jitter	Affine	Zhao et. al [52]	Lukas et. al [53]	SimSwap	UniFace	FaceDancer	HyperReenact
FaceSigns [27]	0.87	0.73	0.72	0.91	0.96	0.87	0.88	0.52	0.49	0.62	0.51
EditGard [13]	0.95	0.78	0.96	0.96	0.99	0.92	0.95	0.48	0.49	0.44	0.50
Omniguard [12]	0.96	0.83	0.96	0.98	**1.0**	0.95	0.96	0.36	0.28	0.25	0.42
VINE [43]	0.96	0.81	0.97	0.95	0.99	0.94	0.98	0.31	0.25	0.29	0.48
DeepForgeSeal (ours)	**1.00**	**0.98**	**0.98**	**0.99**	**1.00**	**0.98**	**0.99**	**0.24**	**0.11**	**0.10**	**0.06**

TABLE IV
EFFECT OF LEARNABLE DIRECTIONAL EMBEDDING IN LATENT SPHERE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	PSNR (↑)	SSIM (↑)	BRA - Benign (↑)	BRA - Malicious (↓)
DeepForgeSeal - w/o [D]	46.23	0.95	0.78	0.42
DeepForgeSeal - FD	46.58	0.95	0.86	0.19
DeepForgeSeal	**48.12**	**0.97**	**0.99**	**0.12**

2) Effects of Composing a Combinatorial Attack and Learning an Attack Curriculum: The effectiveness of the proposed adversarial learning pipeline with composed attacks with a learnable attack curriculum is evaluated in this experiment. We generated eight ablation variants of the proposed DeepForgeSeal model: i) DeepForgeSeal - w/o η - a model without a watermark attacker agent, ii) DeepForgeSeal - w/o $\mathcal{A}$ - a model in which one fixed attack type (no composition or learning of an attack curriculum) is applied, iii) DeepForgeSeal - $|\mathcal{A}|$ - a model in which multiple attacks are applied in a fixed sequence, iv) DeepForgeSeal - $\tilde{\mathcal{A}}$ - a model in which multiple random combinations of attacks applied, v) DeepForgeSeal - $a_1, \tau^{\uparrow}$ - a model in which a single attack type with gradually increasing strength is applied, vi) DeepForgeSeal - $\mathcal{A}^{\rightarrow}$ - a model which progressively introduces more complex combinations of attacks, vii) DeepForgeSeal - a_1, τ - a model with single attack type with learnable strength parameter, viii) DeepForgeSeal - $[\mathcal{A}]$ - a model with a predefined attack curriculum without adaptation to agent performance.

TABLE V
EFFECTS OF COMPOSING A COMBINATORIAL ATTACK AND LEARNING AN ATTACK CURRICULUM. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	PSNR (↑)	SSIM (↑)	BRA - Benign (↑)	BRA - Malicious (↓)
DeepForgeSeal - w/o η	32.11	0.92	0.61	0.59
DeepForgeSeal - w/o $\mathcal{A}$	41.58	0.93	0.72	0.55
DeepForgeSeal - $\lvert\mathcal{A}\rvert$	46.20	0.95	0.82	0.37
DeepForgeSeal - $\tilde{\mathcal{A}}$	47.15	0.96	0.87	0.31
DeepForgeSeal - $a_1, \tau^{\uparrow}$	47.12	0.96	0.76	0.42
DeepForgeSeal - $\mathcal{A}^{\rightarrow}$	48.09	0.96	0.92	0.22
DeepForgeSeal - a_1, τ	48.05	0.96	0.79	0.38
DeepForgeSeal - $[\mathcal{A}]$	47.11	0.95	0.89	0.28
DeepForgeSeal	**48.12**	**0.97**	**0.99**	**0.12**

From the results in Tab. V, we can confirm that there are benefits to incorporating both combinatorial attacks with a learnable attack curriculum as an adversary in the proposed framework, which is evidenced by the rows in Tab. V corresponding to models DeepForgeSeal - $\mathcal{A}^{\rightarrow}$, and DeepForgeSeal - a_1, τ. Most importantly, when the attacker agent has the knowledge of the current vulnerabilities of the victim model (i.e. watermarking agent), it helps the attacker dynamically adapt its strategy and generate a more complex curriculum of attacks, in turn helping the victim to understand the vulnerabilities of its watermarking strategy and rectifying those. We believe this is the reason for the poor performance of the DeepForgeSeal - $[\mathcal{A}]$ model, in which a predefined attack curriculum was used without adaptation to the victim's performance. Furthermore, rows corresponding to DeepForgeSeal - $|\mathcal{A}|$, and DeepForgeSeal - $\tilde{\mathcal{A}}$ confirm the importance of using a combinatorial attack in which a learnable agent determines the attack sequence. Therefore, using this experiment, we can confirm the necessity of both combinatorial attacks and learning an attack curriculum to enhance the robustness of the watermark, while simultaneously preserving image fidelity.

3) Effects of Promoting Semantic Drift and Targeted Shifts Toward Known Failure Regions: To better establish the contributions of the proposed reward function that promotes semantic drift and targeted shifts toward known failure regions, we conducted a further ablation experiment. Specifically, three ablation variants of the proposed model were generated: i) DeepForgeSeal - w/o $[R_{\text{proximity}}, R_{\text{curiosity}}]$ - proposed model without both curiosity reward and proximity reward, ii) DeepForgeSeal - w/o $R_{\text{proximity}}$ - proposed model with curiosity reward but without proximity reward, iii) DeepForgeSeal - w/o $R_{\text{curiosity}}$ - proposed model without curiosity reward but with proximity reward.

TABLE VI
EFFECTS OF THE PROMOTING SEMANTIC DRIFT AND TARGETED SHIFTS TOWARD KNOWN FAILURE REGIONS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	PSNR (↑)	SSIM (↑)	BRA - Benign (↑)	BRA - Malicious (↓)
DeepForgeSeal - w/o $[R_{\text{proximity}}, R_{\text{curiosity}}]$	48.10	0.96	0.82	0.35
DeepForgeSeal - w/o $R_{\text{proximity}}$	48.12	0.96	0.91	0.24
DeepForgeSeal - w/o $R_{\text{curiosity}}$	48.08	0.96	0.94	0.21
DeepForgeSeal	**48.12**	**0.97**	**0.99**	**0.12**

The results presented in Tab. VI confirm the need for the proposed innovative reward function. In particular, the ablation model without both the proximity and the curiosity rewards (see the row corresponding to DeepForgeSeal - w/o $[R_{\text{proximity}}, R_{\text{curiosity}}]$ in Tab. VI) struggles to achieve a good watermark robustness. This is because the attacker agent has learned a sub-optimal attacking policy, marking the watermarking agent unprepared for real-world attacks against watermarks. Comparing rows corresponding to DeepForgeSeal - w/o $R_{\text{proximity}}$ and DeepForgeSeal - w/o $R_{\text{curiosity}}$ models in Tab. VI, we can confirm that proximity and curiosity rewards complement each other. By leveraging underexplored regions of the latent space alongside known failure zones encountered during watermark retrieval, the attacker agent can formulate a sophisticated attack policy. This, in turn, drives the watermarker agent to refine and strengthen its watermarking strategy. These observations strongly validate the importance of the proposed reward function that promotes semantic drift

and targeted shifts toward known failure regions

For details of hyperparameter evaluation, sensitivity analysis, qualitative evaluations, time-complexity and parameter-count comparison against FaceSigns and EditGuard, and a discussion on limitations of *DeepForgeSeal*, please refer to the supplementary material.

V. Conclusion

This paper presented *DeepForgeSeal*, a Latent Space-Driven Semi-Fragile Watermarking using Adversarial Reinforcement Learning (ARL) for accurate detection of face deepfakes. We demonstrate that the proposed learnable watermarking agent achieves enhanced robustness and semantic stealth through a directional embedding strategy in the latent space. The ARL paradigm enabled the agent to dynamically balance resilience and fragility by interacting with a curriculum of benign and adversarial manipulations introduced by an attacker agent. To intensify this adversarial interaction, we designed reward functions that encouraged semantic drift and targeted perturbations toward known failure regions. This incentivised the attacker agent to develop increasingly sophisticated attack strategies. In turn, this drove the watermarking agent to refine and strengthen its embedding approach, resulting in a more resilient and adaptive watermarking system. Extensive experiments were conducted on two public benchmarks: CelebA and CelebA-HQ, which demonstrated the ability of the proposed framework to outperform the current state-of-the-art algorithms by significant margins.

Acknowledgment

The research was supported by the Australian Government through the Office of National Intelligence Postdoctoral Grant awarded to the primary author under Project NIPG-2024-022.

References

[1] T. Qiao, S. Xie, Y. Chen, F. Retraint, and X. Luo, "Fully unsupervised deepfake video detection via enhanced contrastive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4654–4668, 2024.

[2] T. Hunter, "Princess catherine cancer video spawns fresh round of ai conspiracies," 2024, accessed on 31 09, 2025. [Online]. Available: https://www.washingtonpost.com/technology/2024/03/27/kate-middleton-video-cancer-ai/

[3] W. Galston, "Is seeing still believing? the deepfake challenge to truth in politics," 2020, accessed on 09 23, 2025. [Online]. Available: https://www.brookings.edu/articles/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/

[4] N. S. Agenc, "Nsa, u.s. federal agencies advise on deepfake threats," 2023, accessed on 08 23, 2025. [Online]. Available: https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/

[5] T. Fernando, D. Priyasad, S. Sridharan, A. Ross, and C. Fookes, "Face deepfakes–a comprehensive review," *arXiv preprint arXiv:2502.09812*, 2025.

[6] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *Journal of Business Research*, vol. 154, p. 113368, 2023.

[7] Z. Zhang, J. Zhang, W. Zhou, X. Zhou, Q. Guo, W. Zhang, T. Zhang, and N. Yu, "Facetracer: Unveiling source identities from swapped face images and videos for fraud prevention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[8] Z. Qu, W. Lu, X. Luo, Q. Wang, and X. Cao, "Id-guard: A universal framework for combating facial manipulation via breaking identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[9] Y. Zhao, B. Liu, M. Ding, B. Liu, T. Zhu, and X. Yu, "Proactive deepfake defence via identity watermarking," in *Proceedings of the IEEE/CVF WACV*, 2023, pp. 4602–4611.

[10] T. Wang, H. Cheng, M.-H. Liu, and M. Kankanhalli, "Fractalforensics: Proactive deepfake detection and localization via fractal watermarks," *arXiv preprint arXiv:2504.09451*, 2025.

[11] X. Wu, X. Liao, and B. Ou, "Sepmark: Deep separable watermarking for unified source tracing and deepfake detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1190–1201.

[12] X. Zhang, Z. Tang, Z. Xu, R. Li, Y. Xu, B. Chen, F. Gao, and J. Zhang, "Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking," in *Proceedings of the CVPR*, 2025, pp. 3008–3018.

[13] X. Zhang, R. Li, J. Yu, Y. Xu, W. Li, and J. Zhang, "Editguard: Versatile image watermarking for tamper localization and copyright protection," in *Proceedings of the IEEE/CVF CVPR*, 2024, pp. 11 964–11 974.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PmLR, 2021, pp. 8748–8763.

[15] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of IEEE/CVF CVPR*, 2021, pp. 5039–5049.

[16] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes." in *CVPR workshops*, vol. 1, no. 38, 2019.

[17] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, "Predicting heart rate variations of deepfake videos using neural ode," in *Proceedings of the IEEE/CVF CVPR workshops*, 2019, pp. 0–0.

[18] X. Liao, K. Li, X. Zhu, and K. R. Liu, "Robust detection of image operator chain with two-stream convolutional neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 955–968, 2020.

[19] R. Wang, F. Juefei-Xu, M. Luo, Y. Liu, and L. Wang, "Faketagger: Robust safeguards against deepfake dissemination via provenance tracking," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3546–3555.

[20] T. Wang, M. Huang, H. Cheng, B. Ma, and Y. Wang, "Robust identity perceptual watermark against deepfake face swapping," *arXiv preprint arXiv:2311.01357*, 2023.

[21] T. Wang, M. Huang, H. Cheng, X. Zhang, and Z. Shen, "Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 515–10 524.

[22] X. Liao, Y. Yu, B. Li, Z. Li, and Z. Qin, "A new payload partition strategy in color image steganography," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 3, pp. 685–696, 2019.

[23] X. Liao, J. Yin, M. Chen, and Z. Qin, "Adaptive payload distribution in multiple images steganography based on image texture features," *IEEE transactions on dependable and secure computing*, vol. 19, no. 2, pp. 897–911, 2020.

[24] M. Chen, X. Liao, H. Fang, J. Guo, Y. Chen, and X. Wu, "Flexible partial screen-shooting watermarking with provable robustness," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[25] Y. Yang, C. Liang, H. He, X. Cao, and N. Z. Gong, "Faceguard: Proactive deepfake detection," *arXiv preprint arXiv:2109.05673*, 2021.

[26] N. Beuve, W. Hamidouche, and O. Déforges, "Waterlo: Protect images from deepfakes using localized semi-fragile watermark," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 393–402.

[27] P. Neekhara, S. Hussain, X. Zhang, K. Huang, J. McAuley, and F. Koushanfar, "Facesigns: Semi-fragile watermarks for media authentication," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 11, pp. 1–21, 2024.

[28] A. Wachi, "Failure-scenario maker for rule-based agent using multi-agent adversarial reinforcement learning and its application to autonomous driving," *Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[29] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI*, vol. 32, no. 1, 2018.

[30] J. Wang, W. Xu, Y. Gu, W. Song, and T. C. Green, "Multi-agent reinforcement learning for active voltage control on power distribution networks," *NeurIPS*, vol. 34, pp. 3271–3284, 2021.

[31] Z. Peng, Q. Li, K. M. Hui, C. Liu, and B. Zhou, "Learning to simulate self-driven particles system with coordinated policy optimization," *NeurIPS*, vol. 34, pp. 10 784–10 797, 2021.

[32] M. Kouzehgar, M. Meghjani, and R. Bouffanais, "Multi-agent reinforcement learning for dynamic ocean monitoring by a swarm of buoys," in *Global Oceans 2020: Singapore–US Gulf Coast*. IEEE, 2020, pp. 1–8.

[33] J. Wang, Z. Ren, B. Han, J. Ye, and C. Zhang, "Towards understanding cooperative multi-agent q-learning with value factorization," *NeurIPS*, vol. 34, pp. 29 142–29 155, 2021.

[34] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient," in *Proceedings of the AAAI*, vol. 33, no. 01, 2019, pp. 4213–4220.

[35] T. van der Heiden, C. Salge, E. Gavves, and H. van Hoof, "Robust multi-agent reinforcement learning with social empowerment for coordination and communication," *arXiv preprint arXiv:2012.08255*, 2020.

[36] A. Bukharin, Y. Li, Y. Yu, Q. Zhang, Z. Chen, S. Zuo, C. Zhang, S. Zhang, and T. Zhao, "Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms," *NeurIPS*, vol. 36, pp. 68 121–68 133, 2023.

[37] L. Yuan, F. Chen, Z. Zhang, and Y. Yu, "Communication-robust multi-agent learning by adaptable auxiliary multi-agent adversary generation," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186331, 2024.

[38] J. Cao, Q. Li, Z. Zhang, and J. Ni, "Secure and robust watermarking for ai-generated images: A comprehensive survey," *arXiv preprint arXiv:2510.02384*, 2025.

[39] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF CVPR*, 2019, pp. 4401–4410.

[40] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE ICCV*, 2015, pp. 3730–3738.

[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[42] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Styleganex: Stylegan-based manipulation beyond cropped aligned faces," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 21 000–21 010.

[43] S. Lu, Z. Zhou, J. Lu, Y. Zhu, and A. W.-K. Kong, "Robust watermarking using generative priors against image editing: From benchmarking to advances," in *ICLR*, 2025.

[44] Y. He, N. Yu, M. Keuper, and M. Fritz, "Beyond the spectrum: Detecting deepfakes via re-synthesis," *arXiv preprint arXiv:2105.14376*, 2021.

[45] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF CVPR*, 2020, pp. 8695–8704.

[46] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *ICPR*. IEEE, 2021, pp. 5012–5019.

[47] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *European conference on computer vision*. Springer, 2020, pp. 103–120.

[48] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF CVPR*, 2021, pp. 14 923–14 932.

[49] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF CVPR*, 2022, pp. 18 720–18 729.

[50] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, and J. Zhang, "Fakeshield: Explainable image forgery detection and localization via multi-modal large language models," *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[51] T. Fernando, C. Fookes, S. Sridharan, and S. Denman, "Cross-branch orthogonality for improved generalization in face deepfake detection," *arXiv preprint arXiv:2505.04888*, 2025.

[52] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li, "Invisible image watermarks are provably removable using generative ai," in *NeurIPS*, 2024.

[53] N. Lukas, A. Diaa, L. Fenaux, and F. Kerschbaum, "Leveraging optimization for adaptive attacks on image watermarks," *ICLR 2024*, 2024.

[54] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *MM '20: The 28th ACM International Conference on Multimedia*, 2020.

[55] C. Xu, J. Zhang, Y. Han, G. Tian, X. Zeng, Y. Tai, Y. Wang, C. Wang, and Y. Liu, "Designing one unified framework for high-fidelity face reenactment and swapping," in *European conference on computer vision*. Springer, 2022, pp. 54–71.

[56] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund, "Facedancer: Pose-and occlusion-aware high fidelity face swapping," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3454–3463.

[57] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, "Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 7149–7159.

Tharindu Fernando received his BSc (special degree in computer science) from the University of Peradeniya, Sri Lanka, and his PhD from Queensland University of Technology (QUT), Australia. He is currently a Postdoctoral Research Fellow in the Signal Processing, Artificial Intelligence, and Vision Technologies (SAIVT) research program at the School of Electrical Engineering and Robotics at Queensland University of Technology (QUT). He is a recipient of the 2019 QUT University Award for Outstanding Doctoral Thesis and the 2024 National Intelligence Post-Doctoral Grant.

Clinton Fookes (Senior Member, IEEE) received the B.Eng. in Aerospace/Avionics, the MBA degree, and the Ph.D. degree in computer vision. He is currently the Associate Dean Research, a Professor of Vision and Signal Processing, and Co-Director of the SAIVT Lab (Signal Processing, Artificial Intelligence and Vision Technologies) with the Faculty of Engineering at the Queensland University of Technology, Brisbane, Australia. His research interests include computer vision, machine learning, signal processing, and artificial intelligence. He serves on the editorial boards for IEEE TRANSACTIONS ON IMAGE PROCESSING and Pattern Recognition. He has previously served on the Editorial Board for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He is a Fellow of the International Association of Pattern Recognition, a Fellow of the Australian Academy of Technological Sciences and Engineering, and a Fellow of the Asia-Pacific Artificial Intelligence Association. He is a Senior Member of the IEEE and a multi-award winning researcher including an Australian Institute of Policy and Science Young Tall Poppy, an Australian Museum Eureka Prize winner, Engineers Australia Engineering Excellence Award, Australian Defence Scientist of the Year, and a Senior Fulbright Scholar.

Sridha Sridharan has obtained an MSc (Communication Engineering) degree from the University of Manchester, UK, and a PhD degree from the University of New South Wales, Australia. He is currently with the Queensland University of Technology (QUT) where he is a Professor in the School of Electrical Engineering and Robotics. He has published over 600 papers consisting of publications in journals and in refereed international conferences in the areas of Image and Speech technologies during the period 1990-2023. During this period he has also graduated 85 PhD students in the areas of Image and Speech technologies. Prof Sridharan has also received a number of research grants from various funding bodies including the Commonwealth competitive funding schemes such as the Australian Research Council (ARC) and the National Security Science and Technology (NSST) unit. Several of his research outcomes have been commercialised.