특집논문 (Special Paper)
방송공학회논문지 제31권 제4호, 2026년 7월 (JBE Vol.31, No.4, July 2026)
https://doi.org/10.5909/JBE.2026.31.4.687
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

생성형 미디어 워터마킹의 잠재·픽셀 도메인 분류 체계 및 상충 관계 분석

이 성 주[a)], 조 남 익[a)‡]

A Taxonomy and Trade-off Analysis of Generative Media Watermarking across Latent and Pixel Domains

Sung Ju Lee[a)] and Nam Ik Cho[a)‡]

요 약

텍스트-이미지 확산 모델의 발전으로 고품질 미디어 생성이 보편화되면서, 콘텐츠의 무단 도용과 프롬프트 기반 편집에 의한 변조를 검증하는 문제가 중요해졌다. 전통적 픽셀 도메인 워터마크는 편집 도구에 내재된 확산 사전 지식에 의해 쉽게 제거되어, 잠재 공간에 신호를 심는 워터마킹이 대안으로 부상했다. 본 논문은 삽입 메커니즘(학습 기반·최적화 기반·무학습)과 삽입 표현공간(초기 노이즈·중간 상태·VAE 잠재·픽셀)의 2축 분류 체계를 제시한다. 그 위에서 검증·식별 태스크를 형식화하고, 주파수 무결성이 검출 성능과 시각적 품질에 미치는 영향, 시드 독립형 편집에서의 생존 원리(의미론적 각인 가설), VAE 잠재 위상 변조에 의한 연산 효율화를 사례로 분석한다. 또한 신호처리·크롭·생성형 편집·재생성·파라미터 튜닝의 다섯 공격 범주로 확장해 강건성·품질·효율의 상충을 종합하고, 적응적 공격·재생성·모델 파라미터 변형·동영상 확장·교차 아키텍처·평가 표준화의 여섯 미해결 과제를 제시한다.

Abstract

Recent advances in text-to-image diffusion models have made high-quality media generation ubiquitous, raising the need to verify whether content has been misappropriated or altered by prompt-based editing. Traditional pixel-domain watermarks are easily erased by the diffusion prior embedded in modern editing tools, motivating watermarks that embed signals in the latent space. This paper proposes a taxonomy that organizes such methods along two independent axes: insertion mechanism (learning-based, optimization-based, training-free) and insertion representation space (initial noise, intermediate state, VAE latent, pixel). Within this coordinate system, we formalize the verification and identification tasks and analyze, as case studies, the effect of frequency integrity on detection performance and visual quality, the survival mechanism under seed-independent editing (the Semantic Imprinting Hypothesis), and computational efficiency achieved through phase modulation in the VAE latent space. We further extend the threat model to five attack categories—signal processing, cropping, generative editing, regeneration, and parameter tuning—and synthesize the resulting trade-offs among robustness, quality, and efficiency. Finally, we identify six open challenges: adaptive attacks, regeneration, model-parameter modification, video extension, cross-architecture generalization, and evaluation standardization.

Keyword : Diffusion Model Watermarking, Watermarking Taxonomy, Semantic Watermarking, Generative Editing Robustness, Frequency Integrity

Ⅰ. 서 론

생성형 AI, 특히 텍스트-이미지 잠재 확산 모델(Latent Diffusion Models, LDMs)의 발전으로 고품질 미디어 콘텐츠 생성이 보편화되고 있다. 이에 따라 특정 미디어의 AI 생성 여부를 판별하는 기술과 더불어, 해당 콘텐츠의 일부 또는 전체가 무단 도용되거나 프롬프트 기반 인페인팅(Inpainting) 및 의미론적 편집(Semantic Editing) 기술에 의해 변조되었는지를 검증하는 것이 미디어 보호 측면의 핵심 과제로 대두되었다. 이를 해결하기 위해 전통적인 픽셀 도메인 워터마킹[1]이나 미세 노이즈 삽입 방식이 적용되어 왔으나, 이들은 최신 생성형 미디어 편집 도구에 내재된 확산 사전 지식(Diffusion prior)에 의해 외부 노이즈로 간주되어 쉽게 제거된다는 치명적인 한계를 가진다[2]. 이에 대한 대안으로, 확산 모델의 초기 잠재 노이즈 시드(z_T)에 직접 신호를 삽입하는 의미론적 워터마킹(Semantic Watermarking) 기법이 주목받고 있다. 이러한 방식은 설계 관점에 따라 샘플링 기반(Sampling-based) 워터마킹 또는 역산 기반(Inversion-based) 생성 이미지 워터마킹 등으로 구분된다. 대표적으로는 초기 시드 벡터의 주파수 영역에 신호를 삽입하는 기법들[3,4]과 가우시안 노이즈의 통계적 분포를 보존하면서 워터마크 정보를 암호화하는 방식[5] 등이 제안되었다. 의미론적 워터마킹 기술은 생성형 재합성(Generative re-synthesis) 과정에서도 원본 잠재 궤적이 유지되는 특성 덕분에 높은 강건성을 보이는 반면, 주파수 스펙트럼의 왜곡이나 탐지 연산 비용의 증가와 같은 고유한 한계점 역시 보고되고 있다.

이러한 배경에서 본 논문은 생성형 미디어를 위한 워터마킹 기술을 비교하며, 비교 대상으로 선정한 개별 알고리즘의 성능 비교에 그치지 않고, 생성·편집·검출 과정 전반에서 워터마크 신호가 보존되는 조건을 분석하는 관점에서 각 방법들을 조망한다. 이를 통해 기존 픽셀 도메인 방식의 한계 이후 제안된 잠재 공간 기반 접근들이 어떤 문제를 해결하려 하며, 동시에 어떤 새로운 제약을 남기는지를 체계적으로 정리하고자 한다. 이를 위해 기존 연구를 단순히 기법별로 나열하는 대신, 삽입 메커니즘과 삽입 표현공간이라는 두 축으로 재구성하고, 각 설계 선택이 강건성, 시각적 품질, 연산 효율 사이에서 형성하는 상충 관계를 비교한다. 특히 잠재 공간 워터마킹이 생성형 편집 및 재생성 환경에서 신호를 보존하는 원리를 중심으로, 향후 방송·미디어 보호 기술 설계에 필요한 위협 모델과 연구 과제를 도출하고자 한다. 이에 따라, 본 조사는 확산 모델 기반 생성 미디어에 대한 잠재 공간 워터마킹 기술을 중심으로, 최근 제안된 핵심 문헌들을 집중적으로 다룬다. 문헌 선정은 Ⅱ장에서 정의하는 '삽입 메커니즘'과 '표현공간'의 2축 분류 체계를 바탕으로 각 범주를 대표할 수 있는 기법을 우선하였으며, 비교 기준점으로서 각 범주의 초기 랜드마크 연구와 고전적 신호처리 기반 기법을 포함하였다. 단, 확산 모델이 아닌 타 생성 모델(예: GAN)을 대상으로 한 워터마킹 기법은 본 조사의 범위에서 제외하였다.

본 논문의 구성은 다음과 같다. 먼저 분류 체계를 통해 픽셀 도메인부터 초기 노이즈 영역에 이르는 전반적인 설계 공간을 조망한다(Ⅱ장). 이어 의미론적 워터마킹을 형식적으로 정의하고(Ⅲ장), 이 좌표계 위에서 주파수 무결성이 탐지 성능과 시각적 품질에 미치는 영향(Ⅳ장), 그리고 시드 독립형 편집 환경에서의 워터마크 생존 메커니즘을 분석한다(Ⅴ장). 아울러, VAE 잠재 공간의 위상 변조를 통한 연산 효율화 방안을 살펴보고(Ⅵ장), 이를 신호처리·크롭·생성형 편집·재생성뿐만 아니라 모델 파라미터 튜닝 및 영상 도메인까지 아우르는 확장된 위협 모델 체계 속에서 다각도로 검증한다(Ⅶ장). 이 과정에서 서로 다른 연구들의 실험 조건 차이를 명시함으로써 데이터셋, 생성 모델, 공격 조건에 따른 독자적인 기술적 비평을 제공하고자 한다. 최종적으로 이러한 다각적 분석과 기술적 비평을 종합하여 잠재 공간 워터마킹 기술의 한계를 진단하고 연구의 결론을 맺는다(Ⅷ장).

a) 서울대학교 전기·정보공학부 뉴미디어통신공동연구소(Department of ECE, INMC, Seoul National University)
‡ Corresponding Author : 조남익(Nam Ik Cho)
E-mail: nicho@snu.ac.kr
Tel: +82-2-880-8420
ORCID: https://orcid.org/0000-0001-5297-4649
· Manuscript received July 6, 2026; Revised July 9, 2026; Accepted July 9, 2026.

Ⅱ. 잠재 공간 워터마킹의 분류 체계

생성형 미디어 워터마킹 기술은 삽입 메커니즘과 삽입 표현공간이라는 서로 독립적인 두 가지 기준에 따라 다양한 설계 선택지를 가진다. 그러나 기존 서술은 학습 기반, 초기 노이즈(z_T) 기반, VAE 잠재 공간(z_0) 기반, 픽셀 도메인 기반처럼 이 두 기준을 하나의 축에 혼재하여 나열하는 경우가 많아 개별 기법 간의 근본적인 공통점과 차이점을 체계적으로 조망하기 어렵다. 예컨대 Stable Signature[6]는 학습 기반이면서도 워터마크 신호는 픽셀 도메인에서 발현되는데, 이는 두 기준이 독립적임을 보여준다. 본 장에서는 삽입 메커니즘과 삽입 표현공간을 독립된 두 축으로 정의하여 주요 워터마킹 기법을 재분류하고, 설계 공간 전체의 구조를 명확히 제시하고자 한다.

1. 삽입 메커니즘

삽입 메커니즘은 워터마크 신호가 어떤 절차로 구현되는가를 기준으로 하며, 본 논문은 이를 세 범주로 구분한다. 학습 기반(learning-based)은 모델 파라미터를 학습 또는 미세조정하여 신호를 각인하는 방식이다. 최적화 기반(optimization-based)은 모델 자체는 고정한 채 입력 잠재 텐서만을 역전파를 통해 목표 패턴에 부합하도록 최적화하는 방식이다. 무학습(training-free)은 고정된 규칙이나 명시적 변환을 통해 신호를 삽입하며, 별도의 학습이나 최적화 과정을 거치지 않는다.

2. 삽입 표현공간

삽입 표현공간은 신호가 실제로 위치하는 계산 단계를 기준으로 초기 노이즈(z_T), 중간 상태(z_t), VAE 잠재 공간(z_0), 픽셀 도메인의 네 범주로 구분하며, 각 범주는 다시 주파수(frequency)와 공간(spatial) 영역으로 세분된다. 다만 이러한 구분은 엄격한 이산 분류라기보다는 생성 파이프라인상의 연속적인 스펙트럼에 가깝다. 우선 초기 노이즈는 확산 생성의 시작점이 되는 잠재 시드로, 신호가 전체 생성 과정을 거치며 발현된다. 중간 상태는 연속적인 생성 디노이징 과정의 중간 단계에 신호를 심는 방식으로, 대표적으로 $t \in [200, 300]$ 구간에서 적대적 최적화를 통해 워터마크를 삽입하는 ROBIN[7]이 이에 속한다. VAE 잠재 공간은 확산 생성의 전후 단계에서 VAE 잠재 텐서에 직접 신호를 변조하여 편입시키는 지점이며, 픽셀 도메인은 최종 생성 완료된 이미지 자체에 신호가 잔존하도록 처리하는 경우이다.

3. 배정 원칙: 신호의 실제 주입 지점 판별

개별 기법을 배치할 때 유의할 점은, 파이프라인상에 특정 표현공간이 등장한다는 사실과 신호가 실제로 그 공간에 주입 및 검출된다는 사실을 혼동하지 않는 것이다. 예컨대 Stable Signature[6]는 VAE 잠재 벡터 z_0를 거치지만, 워터마크 신호 자체는 z_0에 삽입되지 않는다. VAE 디코더를 미세조정하여 그 매핑 함수에 신호를 각인시키므로 신호는 디코더 통과 이후의 픽셀 단계에서 발현되며, 따라서 본 분류는 이를 픽셀·학습 기반으로 배정한다.

역방향의 사례도 존재한다. ZoDiac[8]은 표면적으로 z_0 및 픽셀 이미지를 다루는 후처리 기법으로 보이나, 실제 최적화와 검출 대상은 초기 노이즈의 푸리에 표현이다. 입력 이미지를 역산해 얻은 잠재 시드가 Tree-Ring[3]과 동일한 저주파 스펙트럼 패턴에 부합하도록 최적화하므로, 배정 기준은 목표 패턴이 정의 및 검출되는 실제 지점인 초기 노이즈·주파수 영역이다.

표현공간 내 주파수와 공간 하위 구분 역시 균일한 판별력을 갖지 않는다. 고정된 규칙이나 명시적 변환으로 신호를 심는 무학습 및 최적화 기반 메커니즘에서는 이 구분이 삽입 원리를 그대로 반영하지만, 학습 기반 메커니즘에서는 판별 경계가 모호해진다. RivaGAN[9]의 어텐션 기반 인코더가 픽셀에 삽입하는 패턴은 학습으로 결정되는 분포이므로 고전적 공간 삽입과 성격이 다르다. 이러한 모호함은 분류의 결함이 아니라, 두 축의 교차 과정에서 각 메커니즘의 고유한 특성이 드러나는 지점이다.

4. 2축 분류 체계: 종합과 함의

표 1은 본 조사가 다루는 주요 기법을 앞서 정의한 두

표 1. 삽입 메커니즘과 표현공간에 따른 잠재 공간 워터마킹 기법의 분류 체계
Table 1. Taxonomy Matrix of Latent Space Watermarking Methods by Insertion Mechanism and Representation Space

	← Mechanism →		
↓Representation Space	Training-free	Optimization-based	Learning-based
Initial Noise (z_T)	Tree-Ring[3] (freq.), RingID[4] (freq.), G.Shading[5] (spatial), HSTR/HSQR[11] (freq.), WIND[12] (freq., seed)	ZoDiac[8] (freq.)	DiffuseTrace[13] (spatial)
Intermediate (z_t)	-	ROBIN[7]	-
VAE Latent (z_0)	PhaseMark[14] (freq., phase)	FreqMark[10] (freq.)	RoSteALS[15] (spatial), Latent Watermark[16] (spatial)
Pixel	Classical DCT/DWT, DwtDctSvd[17] (freq.)	-	Stable Signature[6] (spatial, VAE-FT), RivaGAN[9] (spatial, attention), HiDDeN[18] (spatial)

축으로 배치한 결과이다. 배치된 기법들은 특정 연구에 국한되지 않고 생성 모델 워터마킹 분야 전반의 설계 공간을 포괄한다. 일부 셀이 비어 있는 현상은 기술적 제약과 설계적 트레이드오프를 반영한다. 초기 노이즈의 직접 최적화는 전체 생성 궤적을 통과하는 그래디언트 전파 비용이 커 무학습 규칙이나 중간 상태 최적화로 대체된다. 중간 상태를 고정 규칙이나 파라미터 학습으로 겨냥하는 사례는 드물며, 픽셀 직접 최적화는 재생성에 취약하다는 점[10] 때문에 최적화가 주로 잠재 공간에서 이루어진다.

삽입 메커니즘과 표현공간 축이 워터마크의 주입 방법과 위치를 다룬다면, 본 조사가 함께 고려하는 또 다른 축은 외부 위협에 대한 공격 강건성이다. 본 논문은 이를 신호처리, 크롭, 생성형 편집, 재생성, 파라미터 튜닝의 다섯 범주로 구분하며 상세는 Ⅶ장에서 다룬다.

Ⅲ. 의미론적 워터마킹의 수학적 정식화 및 탐지 태스크

Ⅱ장에서 정의한 삽입 표현공간 축 가운데, 본 장부터 Ⅴ

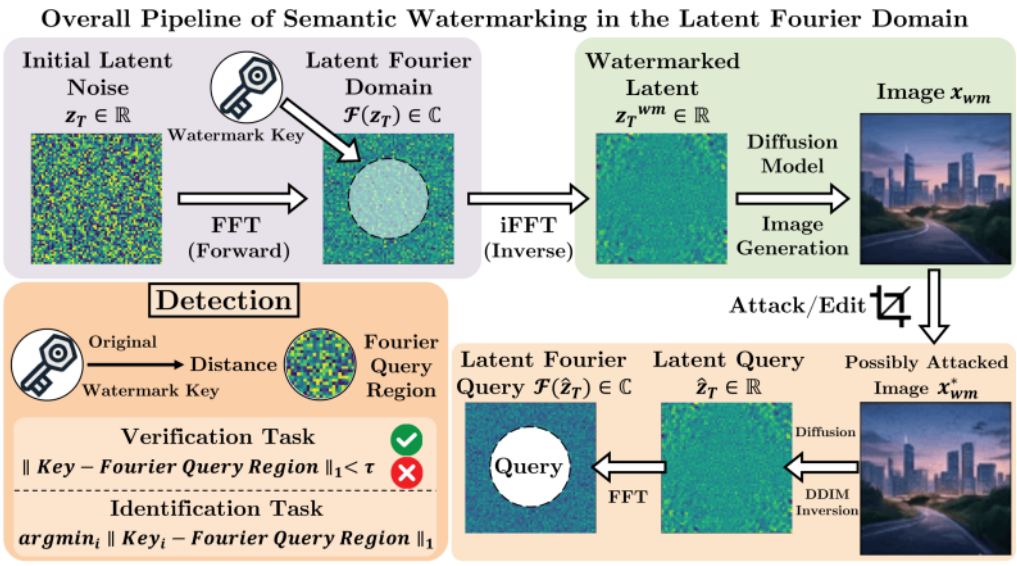

그림 1. 잠재 푸리에 영역에서의 의미론적 워터마킹 전체 파이프라인. 초기 노이즈 시드에서의 삽입, 이미지 생성, 역산 기반 추출 및 거리 기반 탐지 과정을 포함한다.
Fig. 1. Overall pipeline of semantic watermarking in the latent Fourier domain. The process includes embedding in the initial noise seed, image generation, inversion-based extraction, and distance-based detection.

장까지는 초기 노이즈(z_T)의 주파수 영역에 신호를 삽입하는 의미론적 워터마킹(semantic watermarking)을 구체적 사례로 삼아 그 수학적 정식화, 주파수 무결성의 효과, 편집 생존 메커니즘을 순서대로 규명한다.

의미론적 워터마킹의 강건성과 성능을 평가하기 위해서는 삽입 및 추출 과정과 더불어, 두 가지 주요 탐지 태스크인 검증(Verification) 및 식별(Identification)에 대한 수학적 정의가 필요하다. 그림 1은 의미론적 워터마킹의 전체 절차를 잠재 푸리에 영역에서의 워터마크 삽입, 이미지 생성, 역산 기반 추출, 그리고 거리 기반 탐지 단계로 구분하여 나타낸다.

1. 잠재 공간 워터마크 삽입 및 추출

텍스트-이미지 확산 모델 G는 초기 노이즈 시드 $z_T \sim N(0,I)$와 소스 프롬프트 p_s를 입력받아 이미지 $x = G(z_T, p_s)$를 생성한다. 의미론적 워터마킹은 z_T의 푸리에 변환(Fourier Transform) 영역 내 특정 주파수 대역(Key Region)에 워터마크 패턴 w를 삽입하여 워터마크가 포함된 시드 z_T^{wm}을 생성한다. 이후 $x_{wm} = G(z_T^{wm}, p_s)$를 통해 이미지를 얻는다.

탐지 과정에서는 검사 대상 이미지 x_{query}에 대해 상미분방정식(ODE) 기반의 DDIM Inversion 프로세스를 수행하여 역산된 잠재 시드 $\widehat{z_T}$를 추출한다. 이후 $\widehat{z_T}$를 푸리에 변환하여 해당 주파수 대역의 키 $\hat{w}$를 얻어낸다.

2. 검증(Verification) 태스크

검증은 주어진 이미지 내에 워터마크가 존재하는지 여부를 판단하는 이진 분류 문제이다. 추출된 키 $\hat{w}$와 기준이 되는 원본 워터마크 키 w 사이의 거리 $d(\hat{w}, w)$ (주로 푸리에 영역에서의 L_1 거리)를 계산한다. 워터마크가 없는 경우의 거리를 $d(\hat{u}, w)$ (여기서 $\hat{u}$는 Unwatermarked query key)라고 할 때, 사전에 정의된 임계값(Threshold) τ를 기준으로 워터마크 존재 여부를 판별한다. 주로 1%의 오탐률(False Positive Rate, FPR)을 허용하는 기준에서 정탐률(True Positive Rate, TPR)을 측정하는 TPR@1%FPR 지표가 활용된다.

3. 식별 및 사용자 속성(Identification / User Attribution) 태스크

식별은 이미지가 "어떤 사용자(또는 모델)"에 의해 생성되었는지를 판별하는 다중 클래스 분류 문제이다. 워터마크가 삽입되어 있다는 전제하에, N개의 참조 키 풀 $\{w_1, w_2, \ldots, w_N\}$ 중에서 추출된 키 $\hat{w}$와 거리가 가장 가까운 키의 인덱스 $\hat{i}$를 탐색하며, 이를 식으로 나타내면 다음과 같다.

$$\hat{i} = \arg\min_i d(\hat{w}, w_i) \quad (1)$$

이 태스크는 단순히 워터마크 유무를 판단하는 것보다 훨씬 높은 수준의 신호 보존력을 요구하므로, 강한 왜곡 환경에서 워터마킹 기법의 성능을 평가하는 주요 지표(Accuracy 또는 Perfect Match Rate)로 활용된다[11].

Ⅳ. 의미론적 워터마킹과 주파수 무결성의 중요성

초기 의미론적 워터마킹 기법인 Tree-Ring[3] 등은 초기 잠재 노이즈 벡터의 주파수 성분을 특정 패턴으로 변조하여 정보를 삽입한다. 그러나 역 푸리에 변환 과정에서 실수값 공간(real-valued domain)을 유지하기 위해 허수 성분이 제거되며, 이로 인해 주파수 무결성 손실이 발생한다. 이러한 잠재 가우시안 노이즈의 분포 외(Out-of-Distribution, OOD) 특성은 워터마크 탐지 성능 저하뿐만 아니라 생성 이미지의 시각적 품질 저하를 초래한다.

1. 주파수 무결성을 통한 생성 품질 및 식별력 강화

최근 제안된 HSTR(Hermitian Symmetric Tree-Ring) 및 HSQR(Hermitian Symmetric QR code) 방식[11]은 이 문제

를 에르미트 대칭(Hermitian Symmetry) 조건을 강제함으로써 기존 방식의 주파수 무결성 문제를 보완하였다. 해당 대칭성은 주파수 성분의 손실 없는 복원을 가능하게 하며, 이를 통해 탐지 성능 향상에 기여한다.

표 2와 표 3은 Ⅱ장의 분류 체계에 속하는 주요 기법들을 선행 연구[14]의 통일된 벤치마크 조건에 맞추어 비교 분석한 결과이다. Tree-Ring과 동일한 패턴을 사용하면서 주파수 무결성을 추가로 달성한 HSTR은 Tree-Ring 대비 검증(0.699→0.993)과 식별(0.125→0.946) 성능 모두에서 유의미한 향상을 보인다. 한편, RingID와 HSQR은 구조화된 이진 패턴을 공유하지만 세부 설계가 상이하므로 두 기법의 격차를 대칭성 유무의 효과로만 한정 짓기는 어렵다. 그럼에도 불구하고 무결성을 보존한 HSQR은 검증 및 식별 성능 측면에서 RingID와 유사한 수준(각 0.999 내외)을 기록한 반면, 품질 지표(표 3의 ΔFID-1k)에서는 HSQR이 -0.721, RingID가 +2.054를 기록하여 명확한 차이를 보인다. 즉, 검출력이 이미 포화된 지점에서는 무결성 보존 여부가 생성 품질의 손실을 방지하는 핵심 변수로 작용함을 알 수 있다.

2. 구조적 설계를 통한 공간적 공격(Cropping) 방어

워터마크의 강건성을 확보하기 위해서는 주파수 무결성

표 2. 통일 벤치마크 기준 워터마킹 기법 검출 성능 비교
Table 2. Detection Performance Comparison of Watermarking Methods under a Unified Benchmark

Method	Mechanism / Space	Post-hoc	Payload	Verif. (T@1%F)	User IDs	Ident. (Acc.)
DwtDctSvd[17]	Training-free / Pixel·freq.	✓	32	0.494	10^6	0.206
RivaGAN[9]	Learning-based / Pixel	✓	32	0.692	10^6	0.484
Stable Signature[6]	Learning-based (VAE-FT) / Pixel	✗	48	0.819	10^6	0.489
Tree-Ring[3]	Training-free / z_T ·freq.	✗	zero-bit	0.699	2048	0.125
HSTR[11]	Training-free / z_T ·freq.	✗	zero-bit	0.993	2048	0.946
RingID[4]	Training-free / z_T ·freq.	✗	zero-bit	0.999	2048	0.977
HSQR[11]	Training-free / z_T ·freq.	✗	zero-bit	0.999	8192	0.995
G.Shading[5]	Training-free / z_T ·spatial	✗	256	1.000	10^6	0.999
ZoDiac[8]	Optimization-based / z_T ·freq.	✓	zero-bit	0.982	64	0.031
PhaseMark-PCQ[14]	Training-free / z_0 ·freq.	✓	128	0.970	10^6	0.816
PhaseMark-APM[14]	Training-free / z_0 ·freq.	✓	128	0.999	10^6	0.988

표 3. 통일 벤치마크 기준 이미지 생성 품질 비교
Table 3. Image Generative Quality Comparison under a Unified Benchmark

Method	Mechanism / Space	ΔFID-1k	Note
RivaGAN[9]	Learning-based / Pixel	-0.785	-
HSQR[11]	Training-free / z_T ·freq.	-0.721	Integrity-preserving
PhaseMark-PCQ[14]	Training-free / z_0 ·freq.	-0.517	Quality-centric variant
DwtDctSvd[17]	Training-free / Pixel·freq.	-0.389	-
G.Shading[5]	Training-free / z_T ·spatial	-0.317	Distribution-preserving
Stable Signature[6]	Learning-based (VAE-FT) / Pixel	-0.241	-
HSTR[11]	Training-free / z_T ·freq.	-0.099	Integrity-preserving
ZoDiac[8]	Optimization-based / z_T ·freq.	+0.091	-
PhaseMark-APM[14]	Training-free / z_0 ·freq.	+0.042	Performance-centric variant
Tree-Ring[3]	Training-free / z_T ·freq.	+1.184	Integrity not preserved
RingID[4]	Training-free / z_T ·freq.	+2.054	Integrity not preserved; artifact

표 4. 공간적 왜곡(Random Crop) 공격 하에서의 식별(Identification) 정확도 비교
Table 4. Identification Accuracy under Spatial Distortion (Random Crop Attack)

Method	Crop Scale 0.3	Crop Scale 0.4	Crop Scale 0.5	Crop Scale 0.6
RingID[4]	0.559	0.774	0.919	0.971
HSTR[11]	0.903	0.992	0.999	1.000
HSQR[11]	0.999	0.999	1.000	1.000

뿐만 아니라 크롭(Cropping)과 같은 공간적 왜곡에 대한 내성 또한 중요하다. 기존 방식들은 이미지 전체 영역에 푸리에 변환을 적용하므로, 이미지의 일부 영역이 제거될 경우 워터마크 정보 손실이 크게 발생하는 한계를 지닌다. HSTR 및 HSQR은 잠재 벡터의 중앙 영역(center-aware)에 선택적으로 워터마크를 삽입하는 전략을 통해 공간적 왜곡에 대한 강건성을 향상시켰다.

표 4는 크롭 비율에 따른 식별 정확도를 나타내며, 해당 데이터 역시 선행 연구[11]의 벤치마크 환경에서 산출되었다. 분석 결과, 주파수 무결성과 중앙 영역 기반 삽입 전략을 결합한 HSTR 및 HSQR 방식은 랜덤 크롭 공격 환경에서도 높은 강건성을 유지하는 것으로 나타났다.

V. 생성형 편집 공격에 대한 생존 메커니즘 규명

기존 의미론적 워터마킹 기법의 강건성은 주로 인버전 과정을 통해 원래의 생성 궤적을 복원할 수 있다는 전제, 즉 시드 연계형(Seed-linked) 패러다임에서 입증되어 왔다. 그러나 InfEdit[19]과 같이 생성 과정을 새로운 랜덤 시드로 재초기화하는 시드 독립형(Seed-independent) 편집 기법이 등장함에 따라, 생성형 재합성에 대한 보다 엄밀한 강건성 평가가 요구되고 있다.

1. 편집 방식에 따른 워터마크 강건성 차이

표 5는 시드 연계형 편집(P2P)[20]과 시드 독립형 편집(InfEdit) 환경에서 다양한 워터마킹 기법의 생존율을 비교한 결과이다. 주파수 대역에 정렬된(frequency-aligned) HSQR과 분포 보존형 암호화 기법인 Gaussian Shading은 무작위 추출된 새로운 시드에서 생성 과정이 시작되는 InfEdit 환경에서도 0.98 이상의 높은 강건성을 유지한다. 반면, 주파수 무결성 손실로 스펙트럼 불일치가 큰 Tree-Ring은 평균 탐지율이 P2P에서 0.25, InfEdit에서 0.32 수준으로 낮게 측정된다[21]. 그러나 이러한 탐지율 저하는 생성형 편집 과정에서 워터마크가 직접적으로 파괴되었기 때문만은 아니다. 세부 실험 결과, Tree-Ring은 편집 이전의 원본 이미지에서도 이미 낮은 검출률을 보였으며, 편집 후 검출률 또한 원본과 유사한 수준을 유지하였다.

이는 의미론적 편집 과정이 워터마크 신호를 일괄적으로 제거한다기보다, 원본 이미지 단계에서 형성된 검출 가능성이 편집 이후에도 상당 부분 유지될 수 있음을 시사한다. 그러나 Tree-Ring은 OOD 패턴으로 인한 구조적 한계로 인해 원본 이미지 단계에서도 신호가 안정적으로 복원되지 못하며, 이러한 한계가 post-edit attack이 추가된 환경에서 전체 성능 저하로 이어진다. 따라서 매니폴드에 적절히 정렬된 워터마크만이 원본 이미지에서의 높은 신호 보존력을 바탕으로 강한 생성형 편집 환경에서도 실질적인 강건성을

표 5. 생성형 편집 메커니즘에 따른 워터마크 검증(TPR@1%FPR) 성능 비교
Table 5. Verification Robustness (TPR@1%FPR) under Generative Editing Mechanisms

Method Category	Method	P2P (Inversion-based)	InfEdit (Seed-independent)
Spectral Modification	Tree-Ring[3]	0.25	0.32
Optimization-based	ZoDiac[8]	0.75	0.86
Frequency-aligned	HSQR[11]	0.99	1.00
Cryptographic	G.Shading[5]	1.00	0.98

확보할 수 있다.

표 5의 결과는 SD2.1로 생성된 이미지를 SD1.5 기반의 InfEdit으로 변조하는 교차 모델(cross-model) 환경에서 산출된 것으로, 앞선 실험들과는 데이터셋 및 평가 세팅이 상이하다[21]. 한편 주파수 및 매니폴드 정렬 방식이 아키텍처 전이까지 포함된 가혹한 조건에서 높은 검출 성능을 유지한다는 점은 시드 독립형 공격에 대한 실질적인 강건성을 나타낸다.

2. 의미론적 각인 가설(Semantic Imprinting Hypothesis)

매니폴드에 정렬된 워터마크가 원본 시드와 독립적인 생성형 편집 환경에서도 보존되는 원인은 무엇인가? 최신 연구[21]는 이를 '의미론적 각인 가설(Semantic Imprinting Hypothesis)'을 통해 설명한다. 그림 2는 이러한 가설을 OOD 특성이 강한 의미론적 워터마크와 주파수·매니폴드에 정렬된 의미론적 워터마크의 대비를 통해 개념적으로 나타낸 것이다. 상단의 경우, Tree-Ring과 같이 초기 잠재 시드의 주파수 성분을 직접 변조하는 방식은 워터마크 패턴이 확산 모델의 자연스러운 생성 매니폴드와 충분히 정렬되지 못한다. 이때 워터마크는 이미지의 의미적·시각적 구조에 안정적으로 흡수되기보다 분포 외(OOD) 잡음 성분에 가깝게 남게 되며, 그 결과 편집 이전의 원본 이미지 단계에서도 워터마크 신호가 일관되게 복원되지 못한다. 프롬프트 기반 편집 이후의 낮은 탐지율은 이러한 신호 불안정성이 편집 과정을 통해 새롭게 발생했다기보다는, 원본 생성 단계에서 이미 형성된 낮은 검출 가능성이 재합성 이후에도 거의 그대로 이어진 결과로 해석할 수 있다. 따라서 Tree-Ring의 성능 저하는 생성형 편집이 워터마크를 직접적으로 제거했음을 의미하기보다, 매니폴드에 정렬되지 못한 워터마크가 처음부터 안정적인 의미론적 각인으로 내재화되지 못했음을 보여준다.

반면, 그림 2의 하단은 HSQR이나 Gaussian Shading과 같이 주파수 무결성 또는 분포 보존성을 고려한 방식의 경우를 나타낸다. 이러한 워터마크는 단순히 초기 노이즈 공간에 은닉된 코드에 머무르지 않고, 생성 과정에서 텍스처, 스타일, 국소적 구조와 같은 안정적인 시각적 속성으로 내재화된다. 특히 InfEdit과 같이 새로운 랜덤 시드에서 편집을 시작하는 시드 독립형 환경에서는 원래의 생성 궤적이 보존되지 않으므로, 워터마크가 단순한 시드 코드라면 탐지가 불가능해야 한다. 그럼에도 주파수·매니폴드 정렬 방식이 높은 검출률을 유지한다는 점은, 워터마크 신호가 원본 초기 노이즈 시드 z_T 자체가 아니라 편집 모델의 가이던스

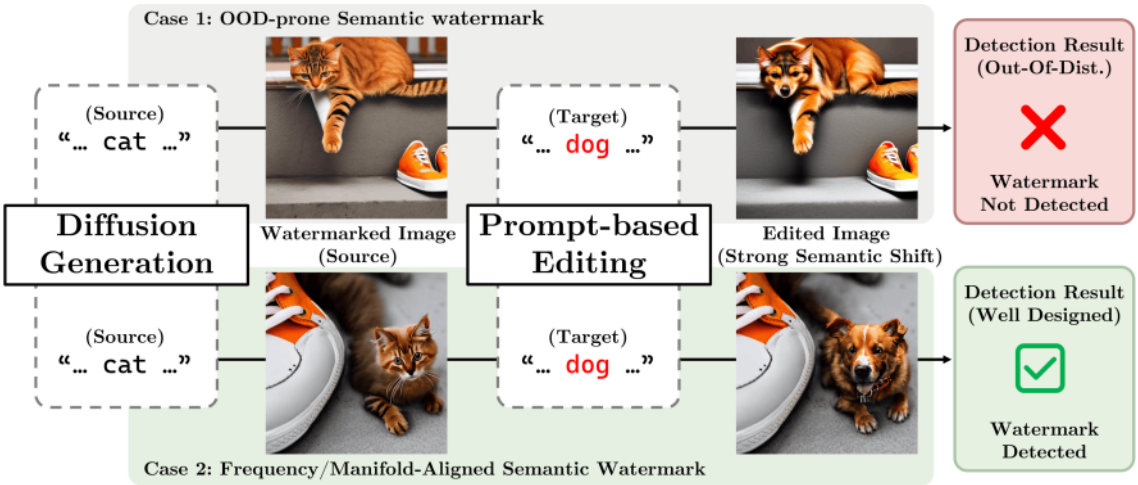

그림 2. 의미론적 각인 가설의 개념도. OOD 특성이 강한 방식은 원본 생성 단계에서부터 안정적인 검출 특성을 형성하지 못하는 반면, 잘 설계된 워터마크는 안정적인 시각적 속성으로 내재화되어 시드 독립적 생성형 재합성 환경에서도 높은 탐지 강건성을 유지한다.

Fig. 2. Visualization of the Semantic Imprinting hypothesis. OOD-prone methods fail to form stable detectable signals from the original generation stage, whereas well-designed watermarks are internalized as stable visual attributes, ensuring robust detection even under seed-independent generative re-synthesis.

가 참조하는 이미지 특징 공간으로 전이되었음을 시사한다.

따라서 의미론적 각인 가설은 강건한 의미론적 워터마크를 “초기 노이즈에 삽입된 은닉 신호”로만 해석하지 않고, 확산 모델의 생성 과정을 거치며 이미지의 의미론적 특징 공간에 반영된 안정적 시각 속성으로 해석한다. 이 관점에서 시드 독립적 생성형 편집은 워터마크가 원본 초기 시드나 생성 궤적에만 종속되는지를 검토하기 위한 조건으로 볼 수 있다. Tree-Ring과 같이 원본 단계에서부터 검출력이 낮은 방식은 편집 이후에도 낮은 검출률이 유지되는 반면, HSQR 및 Gaussian Shading과 같이 원본 단계에서 안정적으로 검출되는 방식은 새로운 시드 기반의 재합성 이후에도 높은 탐지 성능을 유지한다. 이는 워터마크의 편집 후 생존 여부가 편집 과정에서의 단순한 제거 여부만으로 설명되기보다, 생성 단계에서 형성된 신호의 안정성과 이미지 특징 공간으로의 내재화 정도에 의해 좌우될 수 있음을 시사한다.

Ⅵ. 최적화 없는 고효율 위상 변조

초기 노이즈 시드(z_T) 기반의 의미론적 워터마킹은 높은 강건성을 제공하지만, 검출 시마다 계산 비용이 큰 DDIM Inversion 과정을 수행해야 하므로 실시간 서비스 적용에는 한계가 있다. 특히 이미지당 수 초의 처리 시간이 요구된다는 점은 대규모 검출 환경에서 주요 병목으로 작용한다. 이에 따라 확산 과정 이후의 VAE 잠재 공간(z_0)에서 워터마크를 삽입하는 사후 처리(Post-hoc) 방식이 대안으로 재조명되고 있다. 기존 사후 처리 기법인 ZoDiac[8] 등은 최적화 과정에 수 분 단위의 처리 시간이 요구되었으나, 최근 제안된 PhaseMark[14]는 VAE 잠재 공간의 중간 주파수 대역 위상만을 단일 단계(single-shot)로 변조하는 방식을 채택한다.

표 6은 워터마킹 레벨과 삽입·탐지 방식에 따라 처리 시간과 재생성(regeneration) 공격 강건성이 어떻게 달라지는지를 비교한 결과이다. 여기서 재생성 공격은 VAE 압축 및 확산 기반 재생성과 같이 이미지의 잠재 표현이나 생성 통계를 다시 구성하는 변환을 포함한다. PhaseMark와 같은 VAE 잠재 공간 기반 위상 변조 기법은 역산이나 최적화 과정을 요구하지 않음에도, 기존 초기 노이즈 시드 기반 방식에 필적하는 0.998의 높은 강건성을 보인다. 이는 재생성 공격에 대한 내성이 반드시 초기 노이즈 시드에만 의존하는 것이 아니라, 잠재 공간 주파수의 위상 특성을 효과적으로 제어함으로써 확보될 수 있음을 시사한다.

추가적으로, z_0 잠재 공간에서의 위상 변조는 강제적 변조(Hard Modulation)와 유연한 양자화(Soft Modulation) 전략에 따라 탐지 성능과 이미지 품질 사이의 명확한 상충 관계를 형성한다. 표 7은 PhaseMark의 다양한 위상 변조 전

표 6. 삽입 레벨 및 최적화 여부에 따른 탐지 속도 및 재생성 강건성 비교
Table 6. Comparison of Latency and Regeneration Robustness by Target Space and Optimization

Methods	Target Space	Optimization-Free (Embed / Detect)	Latency (Embed / Detect)	Regen. Verif. (TPR@1%FPR)
ZoDiac[8]	z_T	Emb.: X (Opt. z_T) Det.: X (Inversion)	~7.3mins / 4.11s	0.958
Semantic Watermarks[3,4,5,11]	z_T	Emb.: O (Sampling) Det.: X (Inversion)	- / 4.11s	0.999[5], 0.997[11]
PhaseMark[14]	z_0	O (Single-Shot)	0.14s / 0.05s	0.998

표 7. VAE 잠재 공간 위상 변조 전략에 따른 성능-품질 트레이드오프
Table 7. Performance vs. Image Quality Trade-off across Phase Modulation Variants

Modulation	Strategy Type	TPR@1%FPR (Verif.)	Match Rate (Ident.)	PSNR ↑	LPIPS ↓
APM (Robust)	Hard / Absolute	0.999	0.988	31.715	0.129
IPS	Hard / Relative	0.998	0.980	32.437	0.090
SPS	Soft / Relative	0.996	0.960	32.952	0.068
PCQ (Quality)	Soft / Absolute	0.970	0.816	34.156	0.037

략, 즉 절대 위상 변조(APM)와 위상 성좌 양자화(PCQ) 사이의 성능 변화를 나타낸다. 이에 따라 적용 환경의 요구사항에 맞추어 검출 성능을 우선하는 APM 방식 또는 이미지 품질을 우선하는 PCQ 방식을 선택적으로 활용할 수 있다.

Ⅶ. 공격 범위의 확장

앞선 장들(Ⅳ, Ⅴ, Ⅵ장)의 강건성 평가는 주로 모델 배포 이후 발생하는 사후적 변조 위협을 정적 모델(Static Model) 환경에서 검증하였다. 그러나 이는 모델 가중치 자체가 수정되는 미세조정(Fine-tuning) 시나리오나 종합적인 위협 체계를 반영하기 어렵다. 본 장에서는 기존의 파편화된 위협 요인들을 하나의 통일된 분류 체계로 구조화하고, 모델 파라미터 튜닝이라는 잠재적 보안 위협을 정식 편입하여 공격 범위를 확장하고자 한다.

1. 공격 유형의 종합 분류

본 논문은 잠재 공간 워터마킹이 직면하는 공격 패러다임을 신호처리, 크롭, 생성형 편집, 재생성, 파라미터 튜닝의 다섯 가지 범주로 분류한다(표 8). 이전 장들을 통해 실증적으로 분석된 선행 범주들에 더해, 다섯 번째인 파라미터 튜닝은 사용자가 공개된 가중치를 직접 미세조정하여 워터마크 신호를 제거하려는 능동적 위협을 의미한다.

표 8에 제시된 공격별 강건성 특성은 Ⅱ장의 삽입 표현 공간 분류(표 1)와 밀접한 상관관계를 가질 것으로 예상된다. 픽셀 도메인 및 모델 학습 기반 기법들은 재생성과 파라미터 튜닝 공격에 취약할 여지가 있는 반면, 초기 노이즈 및 VAE 잠재 공간 기반 기법들은 상대적으로 높은 생존율을 유지할 것으로 기대된다. 다만 파라미터 튜닝 공격 하에서의 실증적 비교 데이터는 아직 부족한 상태이며, 이러한 한계는 각 연구가 가정한 데이터셋, 생성 모델, 공격 조건의 상이함으로부터 기인하므로 개별 기법의 상대적 우위를 정량적으로 단정하기 어렵다.

2. 모델 파라미터 튜닝 공격

공격자가 모델 가중치를 확보한 뒤 수행할 수 있는 대표적인 미세조정 기법으로는 전체 미세조정, LoRA, DreamBooth, Textual Inversion 등이 있다. 비록 이들이 워터마크 제거만을 목적으로 활용된 실제 연구 사례는 아직 부재하나, 학습 과정에서 발생하는 가중치 변화는 워터마크의 생존력을 위협하는 잠재적 공격 표면(Attack Surface)으로서 분석 가치가 크다.

각 방법론은 적용 방식에 따라 가중치가 변경되는 레이어의 범위가 상이하다. 전체 미세조정(Full Fine-Tuning)은 U-Net 및 텍스트 인코더 전반의 가중치를 갱신하므로, 가중치 자체에 신호를 체화하는 모델 학습 기반 워터마크에 가장 치명적일 수 있다. 반면 LoRA[25]는 어텐션 레이어의 저랭크 증분만을 학습하여 특정 레이어 의존형 워터마크에 선택적인 영향을 미칠 수 있고, DreamBooth[26]는 사전 보존 손실을 통해 과적합을 방지하며 U-Net 가중치를 갱신하므로 가중치 각인형 워터마크에 영향을 줄 수 있으나 그 범위는

표 8. 공격 유형 분류와 주 위협 대상
Table 8. Taxonomy of Attack Types and Primary Vulnerable Representation

Attack Category	Sub-types	Primary Vulnerable Representation	Representative Literature
Signal Processing (Ⅳ)	Brightness, Contrast, JPEG, Blur, BM3D, Noise	Pixel spatial domain methods	-
Cropping (Ⅳ)	Center crop, Random crop	Globally-embedded methods	Lee & Cho[11]
Generative Editing (Ⅴ)	Seed-linked (P2P/NTI), Seed-independent (InfEdit)	OOD-prone watermarks	Lee & Cho[21]
Regeneration (Ⅵ)	VAE compression, Diffusion regeneration	Pixel-domain collapses; z_T/z_0 largely robust	Zhao et al[2], Ni et al[22]
Parameter Tuning (Ⅶ)	Full fine-tuning, LoRA, DreamBooth, Textual Inversion	Learning-based, pixel methods	Feng et al[23], Wang et al[24]

전체 미세조정보다 국소적일 것으로 예상된다. 마지막으로 Textual Inversion[27]은 신규 토큰의 임베딩 벡터만 학습하고 기존 가중치는 보존하므로 파라미터 의존형 워터마크에 미치는 영향이 극히 제한적일 것으로 예측된다.

최근 연구들은 이러한 미세조정 특성을 방어 설계에 반영하고 있다. AquaLoRA[23]는 워터마크를 U-Net의 핵심 구조에 결합하여, 미세조정으로 워터마크를 제거할 경우 모델의 미디어 생성 성능(Utility)도 함께 훼손되도록 상충관계(Trade-off)를 의도적으로 설계하였다. 다만 이러한 설계는 VAE 교체나 샘플링 변경 등 화이트박스 우회 공격 방어에 초점을 맞춘 것으로, 다운스트림 미세조정에 대한 근본적인 강건성을 보장하지는 않는다. 한편 Sleeper-Mark[24]는 워터마크 정보를 모델이 학습하는 의미 개념으로부터 명시적으로 분리한다. 이를 통해 LoRA 기반 스타일 적응, DreamBooth 개인화, ControlNet 기반 조건 추가 등 다양한 다운스트림 미세조정 태스크 환경에서도 워터마크 강건성을 유지하며, 파라미터 변조 대응이 핵심 설계 목표로 구현될 수 있음을 보여준다.

이러한 특성은 앞서 언급한 표현공간의 내재적 특성과 연결된다. 학습 기반 및 픽셀 도메인 기법들은 신호가 디코더나 백본의 가중치 자체에 반영되어 있어 파라미터 튜닝에 원리적으로 취약할 것으로 예상된다. 반면 무학습 기반 초기 노이즈 기법들은 신호가 생성 시점의 z_T 영역에 존재하여 직접적인 영향에서 상대적으로 자유로울 여지가 있다. 그러나 초기 노이즈 기법 역시 화이트박스 환경에서 무력화될 수 있다는 지적이 존재하며[24], 검출 시에는 DDIM 역산(Inversion) 등 모델 가중치를 경유하는 절차에 의존하므로 파라미터 갱신이 역산 궤적을 왜곡해 검출 성능을 저하시킬 가능성이 존재한다. 결과적으로 파라미터 튜닝 공격 시나리오 하에서 각 표현공간별 내성 차이에 대한 체계적인 비교 및 실증은 향후 독자적인 위협 모델 검증을 통해 구체화되어야 할 부분이다.

3. 최신 편집·동영상 도메인 워터마킹 사례

최근 잠재 공간 워터마킹은 이미지 편집 방어를 넘어 동영상(Video) 도메인으로 확장되는 추세이다. 대표적 연구인 VideoShield[28]는 기존 Gaussian Shading 구조를 시간축으로 확장하고 고유한 template bits를 도입하여 영상의 시간적·공간적 조작 국지화(Localization)를 달성하였다. 프레임 교환, 삽입, 삭제 등의 시간적 변조를 탐지하고 원래 순서로 복원하는 강점이 있으나, DDIM 역산에 따른 높은 연산 비용과 검출 시 동일 해상도 및 프레임 수를 요구하는 한계가 있다. 또 다른 기법인 VideoMark[29]는 영상 도메인의 세 가지 핵심 난제, 즉 (i) 낮은 역산 정확도, (ii) 시간적 왜곡 취약성, (iii) 가변 길이 영상과 고정 초기화 간의 불일치를 명시하였다. 이를 해결하기 위해 VideoMark는 편집 거리 기반의 시간적 정합 모듈을 제안하였다. 이러한 동영상 도메인 확장 연구들은 기존 분류 체계가 정의한 정적 위협 모델을 넘어, 영상 고유의 시간적 공격 범주에 대한 추가적인 방어 설계가 필요함을 시사한다.

Ⅷ. 결 론

본 논문에서는 삽입 메커니즘과 표현공간을 축으로 하는 분류 체계를 통해 잠재 공간 워터마킹 기술의 전체 설계 공간을 조망하고, 이 안에서 최신 확산 모델을 대상으로 한 워터마킹 기술의 발전 동향과 생성형 미디어 편집 공격 환경에서의 강건성을 종합적으로 분석하였다. 먼저, 초기 노이즈 시드의 주파수 무결성과 중앙 영역 삽입 구조를 유지하는 것이 워터마크 검출 성능, 생성 이미지 품질, 그리고 공간적 공격에 대한 강건성을 향상시키는 핵심 요소임을 확인하였다. 또한, 잘 설계된 의미론적 워터마크는 궤적 정보가 유지되지 않는 시드 독립형 편집 환경에서도 시각적 특징 공간에 안정적으로 반영되어 높은 생존율을 유지함을 확인하였다. 마지막으로, 잠재 공간 기반의 최적화 없는 위상 변조 기법은 기존 의미론적 워터마킹의 높은 연산 비용 문제를 완화하면서도 높은 강건성을 유지할 수 있는 효과적인 대안임을 보였다. 이에 더해 제안한 2축 분류 체계는 픽셀부터 초기 노이즈까지 전 설계 공간을 통일된 좌표계 위에서 조망할 수 있는 기반을 제공하였으며, 모델 파라미터 튜닝을 포함하도록 공격 범위를 확장하여 개별 연구들의 실험 조건 상이성을 인지한 독자적인 기술적 비평을 수행하였다.

향후 연구에서는 적응적·화이트박스 공격에 대한 강건성 검증, 재생성 강도에 따른 강건성-품질 곡선의 체계화, 모델 파라미터 변형에 불변한 임베딩 설계, 동영상 도메인으로의 확장, 교차 아키텍처 일반화, 그리고 잠재 공간 워터마킹 전용의 통합 평가 벤치마크 구축이 필요하다. 특히 확산 모델 내부의 어텐션 맵 등 중간 표현을 분석하는 것은 이러한 과제들을 가로질러 활용될 수 있는 공통 방법론이 될 수 있다.

참 고 문 헌 (References)

[1] M. Barni, F. Bartolini, V. Cappellini, and A. Piva, “A DCT-domain System for Robust Image Watermarking,” Signal Processing, Vol.66, No.3, pp.357-372, May 1998.
doi: https://doi.org/10.1016/S0165-1684(98)00015-2

[2] 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,” Advances in Neural Information Processing Systems, Vol.37, pp.8643-8672, 2024.
doi: https://doi.org/10.52202/079017-0276

[3] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, “Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images,” Advances in Neural Information Processing Systems, Vol.36, pp.58047-58063, 2023.
doi: https://doi.org/10.52202/075280-2529

[4] H. Ci, P. Yang, Y. Song, and M. Z. Shou, “RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-Key Identification,” Proceedings of the European Conference on Computer Vision, Milan, Italy, pp.338-354, 2024.
doi: https://doi.org/10.1007/978-3-031-73390-1_20

[5] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, “Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, pp.12162-12171, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01156

[6] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, “The Stable Signature: Rooting Watermarks in Latent Diffusion Models,” Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp.22466-22477, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.02053

[7] H. Huang, Y. Wu, and Q. Wang, “ROBIN: Robust and Invisible Watermarks for Diffusion Models with Adversarial Optimization,” Advances in Neural Information Processing Systems, Vol.37, pp.3937-3963, 2024.
doi: https://doi.org/10.52202/079017-0129

[8] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun, and H. Guan, “Attack-Resilient Image Watermarking Using Stable Diffusion,” Advances in Neural Information Processing Systems, Vol.37, pp.38480-38507, 2024.
doi: https://doi.org/10.52202/079017-1215

[9] K. A. Zhang, L. Xu, A. Cuesta-Infante, and K. Veeramachaneni, “Robust Invisible Video Watermarking with Attention,” arXiv preprint arXiv:1909.01285, Sep. 2019.
doi: https://doi.org/10.48550/arXiv.1909.01285

[10] Y. Guo, R. Li, M. Hui, H. Guo, C. Zhang, C. Cai, L. Wan, and S. Wang, “FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space,” Advances in Neural Information Processing Systems, Vol.37, pp.112237-112261, 2024.
doi: https://doi.org/10.52202/079017-3564

[11] S. J. Lee and N. I. Cho, “Semantic Watermarking Reinvented: Enhancing Robustness and Generation Quality with Fourier Integrity,” Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, USA, pp.18759-18769, 2025.
doi: https://doi.org/10.1109/ICCV51701.2025.01743

[12] K. Arabi, B. Feuer, R. T. Witter, C. Hegde, and N. Cohen, “Hidden in the Noise: Two-Stage Robust Watermarking for Images,” Proceedings of the 13th International Conference on Learning Representations, Singapore, 2025, arXiv:2412.04653.
doi: https://doi.org/10.48550/arXiv.2412.04653

[13] L. Lei, K. Gai, J. Yu, and L. Zhu, “DiffuseTrace: A Transparent and Flexible Watermarking Scheme for Latent Diffusion Model,” arXiv preprint arXiv:2405.02696, May 2024.
doi: https://doi.org/10.48550/arXiv.2405.02696

[14] S. J. Lee and N. I. Cho, “PhaseMark: A Post-Hoc, Optimization-Free Watermarking of AI-Generated Images in the Latent Frequency Domain,” Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 2026.
doi: https://doi.org/10.1109/ICASSP55912.2026.11463051

[15] T. Bui, S. Agarwal, N. Yu, and J. Collomosse, “RoSteALS: Robust Steganography Using Autoencoder Latent Space,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Vancouver, Canada, pp.933-942, 2023.
doi: https://doi.org/10.48550/arXiv.2304.03400

[16] Z. Meng, B. Peng, and J. Dong, “Latent Watermark: Inject and Detect Watermarks in Latent Diffusion Space,” IEEE Transactions on Multimedia, Vol.27, pp.3399-3410, 2025.
doi: https://doi.org/10.1109/TMM.2025.3535300

[17] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, 2nd Edition, Morgan Kaufmann Publishers, Burlington, USA, 2008.
doi: https://doi.org/10.1016/B978-0-12-372585-1.X5001-3

[18] J. Zhu, R. Kaplan, J. Johnson and L. Fei-Fei, “HiDDeN: Hiding Data With Deep Networks,” Proceedings of the European Conference on Computer Vision, pp. 682-697, 2018.
doi: https://doi.org/10.1007/978-3-030-01267-0_40

[19] S. Xu, Y. Huang, J. Pan, Z. Ma, and J. Chai, “Inversion-Free Image Editing with Language-Guided Diffusion Models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Seattle, USA, pp.9452-9461, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00903

[20] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-Prompt Image Editing with Cross-Attention Control,” Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 2023.
doi: https://doi.org/10.48550/arXiv.2208.01626

[21] S. J. Lee and N. I. Cho, “The Semantic Imprinting Hypothesis: How Semantic Watermarks Survive Prompt-based Editing,” ICLR 2026 Workshop on Principled Design for Trustworthy AI-Interpretability, Robustness, and Safety across Modalities, 2026. url: https://openreview.net/forum?id=hNbkvSjEAD

[22] Y. Ni, Z. Yang, Z. Niu, E. Davis, and F. Carter, “On the Information-Theoretic Fragility of Robust Watermarking under Diffusion Editing,” arXiv preprint arXiv:2511.10933, Nov. 2025.
doi: https://doi.org/10.48550/arXiv.2511.10933

[23] W. Feng, W. Zhou, J. He, J. Zhang, T. Wei, G. Li, T. Zhang, W. Zhang, and N. Yu, “AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA,” Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, Vol.235, pp.13423-13444, 2024.
doi: https://doi.org/10.48550/arXiv.2405.11135

[24] Z. Wang, J. Guo, J. Zhu, Y. Li, H. Huang, M. Chen, and Z. Tu, “SleeperMark: Towards Robust Watermark against Fine-Tuning Text-to-image Diffusion Models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, pp.8213-8224, 2025.
doi: https://doi.org/10.1109/CVPR52734.2025.00769

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” Proceedings of the 10th International Conference on Learning Representations, 2022, arXiv:2106.09685
doi: https://doi.org/10.48550/arXiv.2106.09685

[26] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.22500-22510, 2023.
doi: https://doi.org/10.48550/arXiv.2208.12242

[27] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion,” Proceedings of the 11th International Conference on Learning Representations, 2023, arXiv:2208.01618
doi: https://doi.org/10.48550/arXiv.2208.01618

[28] R. Hu, J. Zhang, Y. Li, J. Li, Q. Guo, H. Qiu, and T. Zhang, “VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking,” Proceedings of the 13th International Conference on Learning Representations, Singapore, 2025, arXiv:2501.14195.
doi: https://doi.org/10.48550/arXiv.2501.14195

[29] X. Hu, H. Li, J. Li, Y. Huang, S. Liu, Q. Zheng, J. Chen, and A. Liu, “VideoMark: A Distortion-Free Robust Watermarking Framework for Video Diffusion Models,” arXiv preprint arXiv:2504.16359, Apr. 2025.
doi: https://doi.org/10.48550/arXiv.2504.16359

저 자 소 개

이 성 주

- 현재 : 서울대학교 전기·정보공학부 석박통합과정
- ORCID : https://orcid.org/0009-0003-0269-5483
- 주관심분야 : 생성형 확산 모델, 워터마킹 & 콘텐츠 출처 증명, 컴퓨터 비전

조 남 익

- 현재 : 서울대학교 전기·정보공학부 교수
- ORCID : https://orcid.org/0000-0001-5297-4649
- 주관심분야 : 디지털 신호처리, 영상처리, 컴퓨터 비전