참조 용어 | AI 워터마킹 p.106, 임베딩 p.64

# 112 비가시적 워터마킹

Invisible Watermarking

**보이지 않는 디지털 표식을 삽입해 출처 추적 및 진위를 검증하는 기술**

- AI 생성물의 식별과 저작권 보호, 위변조 방지를 위한 보이지 않는 인증 방식

## 비가시적 워터마킹 개요

비가시적 워터마킹은 이미지, 영상, 음성, 텍스트 등 디지털 콘텐츠 내부에 육안으로 인식되지 않는 표식을 삽입해 저작권과 출처 정보를 기록하는 기술입니다. 겉으로는 원본과 동일하게 보이지만, 파일의 픽셀 배열이나 주파수 영역, 데이터 비트 속에 특정 신호를 암호화하여 저장합니다. 이 표식은 사람이 볼 수 없지만 전용 검출 프로그램을 통해 인식할 수 있어, 콘텐츠의 진위 판별과 위변조 추적에 활용됩니다. 특히 AI 생성 콘텐츠가 확산되면서, 사람이 만든 것인지 AI가 만든 것인지를 구분하기 위한 핵심 기술로 주목받고 있습니다.

## 비가시적 워터마킹의 적용

비가시적 워터마킹은 콘텐츠의 구조적 특성을 이용해 신호를 숨기는 방식으로 구현합니다. 대표적인 방법으로는 주파수 변조는 이미지나 음성의 특정 주파수 대역에 미세한 진폭 변화를 주어 정보를 삽입하는 주파수 변조, 작은 신호를 데이터 전반에 넓게 분산시켜 삽입하여 일부 손상이나 압축이 발생하도록 하는 확산 스펙트럼, AI 모델이 스스로 워터마크를 삽입·검출하는 패턴을 학습하도록 한 딥러닝 기반 임베딩 등이 있습니다.

## 비가시적 워터마킹의 활용과 한계

비가시적 워터마킹은 콘텐츠의 진위 검증, 저작권 보호, 데이터 추적 등 다양한 목적으로 활용됩니다. 특히 AI 생성물의 무단 사용이나 허위정보 유통을 방지하기 위한 기술적 장치로 각국의 AI 정책에서 중요한 역할을 하고 있습니다. 하지만 이미지 편집·압축·확대 등의 후처리 과정에서 워터마크 신호가 손상될 수 있습니다. 또한 과도한 삽입은 콘텐츠 품질을 저하시킬 우려가 있어, 식별 강도와 품질 유지 간의 균형이 여전히 과제로 남아 있습니다.

## 비가시적 워터마킹의 의의

비가시적 워터마킹 또한 AI 신뢰성을 위한 기술로, 가시적 워터마크가 사람이 즉시 인식할 수 있는 '표시'라면, 비가시적 워터마크는 AI 시스템과 플랫폼이 자동으로 인식하고 검증할 수 있는 '인프라 수준의 식별 코드'입니다. 이는 콘텐츠를 시각적으로 구분하기보다는, 디지털 생태계 전체에서 데이터의 출처·생성 과정·진위 여부를 추적하고 기록하는 기술적 신뢰 구조를 형성합니다. 특히 AI 모델 간의 정보 교환, 검색엔진의 콘텐츠 인증, 정부·언론기관의 데이터 검증 체계 등에 적용되면서, AI 거버넌스의 기술적 토대로 확장되고 있습니다.