Biweekly Report
통권 제88호(2026. 8-2)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제88호(2026. 8-2)

AI 생성물 워터마크의 포렌식 능력과 한계

뉴스 브리프

콘텐츠의 출처와 진위를 표시하는 기술이 산업 전반으로 빠르게 번지고 있다. 2026년 7월 영상 공유 기반 SNS인 틱톡이 콘텐츠의 출처와 제작 이력을 인증하는 국제 표준화 기구인 C2PA의 운영위원회에 합류했고, AI가 만든 콘텐츠에 워터마크를 새기는 흐름이 여러 기업으로 확산됐다. 워터마크의 확산은 그것이 실제로 무엇을 증명하고 어디까지 추적할 수 있는지에 관한 물음을 키운다. 최근 연구는 워터마크가 기계가 만든 콘텐츠인지 가려내는 데 그치지 않고, 누가 만들었는지 밝히고 숨은 정보를 꺼내며 워터마크가 있는 부분을 짚어내는 네 가지 기능을 정보이론으로 분석했다. 이 보고서는 그 분석을 쉽게 풀어, 워터마크가 지닌 능력과 한계, 그리고 출처표시 의무화 국면에서의 전망을 살펴본다.

뉴스 플러스

Ⅰ. 서론: 콘텐츠 진위 판별을 둘러싼 기대와 현실

- **대형 플랫폼의 출처 표시 도입과 그 이면**

2026년 7월, 영상 공유 기반 SNS인 틱톡(TikTok)이 C2PA(Coalition for Content Provenance and Authenticity)*의 운영위원회에 합류했다. 틱톡은 디지털 표준 기술인 콘텐츠 크리덴셜(Content Credentials)과 비가시성 워터마크(invisible watermark), 라벨링 도구를 결합해 30억 건이 넘는 콘텐츠를 AI 생성물로 표시해 왔다. 이처럼 콘텐츠에 출처와 제작 이력을 새기고 이를 다시 읽어 콘텐츠의 내력을 밝히려는 시도가 산업의 표준으로 자리 잡아 가고 있다.

* C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 제작 이력을 확인할 수 있는 공통 규격을 개발하기 위해 여러 기업과 기관이 결성한 연합체

이러한 흐름은 틱톡 한 곳에 그치지 않는다. 여러 기업들이 AI가 생성한 콘텐츠에 AI 생성물임을 알리는 워터마크를 적용하기 시작했고, 규제 역시 이 방향을 재촉한다. 그런데 워터마크의 확산은 기술이 감당할 수 있는 범위에 관한 물음을 함께 키운다. 워터마크가 있다는 사실만으로 해당 콘텐츠의 권리가 누구에게 귀속되는지, 어느 부분이 조작됐는지까지 밝혀지는 것은 아니기 때문이다. 실제로 플랫폼이 콘텐츠를 처리하는 과정에서 워터마크가 지워지기도 하고, 하나의 기술이 요구되는 모든 기능을 충족하지 못하기도 한다. 결국 워터마크를 심는 일과 그 워터마크로 무엇을 증명할 수 있는가는 서로 다른 문제이며, 후자에 대한 답은 아직 충분히 정리되지 않았다.

• 워터마크 포렌식 연구가 등장한 배경과 의의

워터마크를 심는 기술이 빠르게 퍼지는 동안, 그 워터마크가 실제로 무엇을 증명할 수 있는지 따져 묻는 연구가 나왔다. 2026년 발표된 한 연구는 생성형 모델이 만든 콘텐츠에 담긴 워터마크를 포렌식(forensics)*의 관점에서 분석한다. 연구진은 하나의 워터마크가 답할 수 있는 물음을 네 가지로 나눈다. AI 생성물인지 가려내는 탐지(detection), 여러 사용자 가운데 누가 만들었는지 밝히는 귀속(attribution), 워터마크에 숨겨진 정보를 복원하는 추출(extraction), 워터마크가 있는 부분이 어디인지 짚어내는 국소화(localization)가 그것이다.

이 물음이 중요한 이유는 규제와 산업이 워터마크에 거는 기대가 실제 기술의 능력과 어긋날 수 있기 때문이다. 최근 여러 규정이 AI 생성물에 기계가 읽을 수 있는 워터마크를 요구하기 시작했고, 이러한 규정은 워터마크가 기계적으로 추출 및 판독할 수 있는 정보를 담고 있다는 전제 위에 서 있다. 하지만 워터마크가 남기는 정보의 양과 그 정보가 콘텐츠에 어떻게 퍼져 있는지에 따라, 같은 워터마크라도 어떤 물음에는 답하고 어떤 물음에는 답하지 못한다. 그래서 연구진의 분석은 이 경계를 정보이론(information theory)의 언어로 그려낸다. 이를 통해 워터마크의 능력과 한계가 워터마크를 심는 방식이 아니라 워터마크가 담은 정보의 구조에 따라 정해진다는 점이 드러난다.

* 포렌식(forensics): 남겨진 흔적을 거꾸로 읽어 그것을 만든 사건이나 원인에 도달하는 분석 작업-

II. 본론: 워터마크 포렌식의 정보이론 분석

• 워터마크가 남기는 정보량과 정보 프로파일

워터마크의 포렌식 능력을 따지려면 먼저 하나의 물음으로 되돌아가야 한다. 워터마크를 읽는 쪽이 콘텐츠에서 그 비밀 정보를 얼마나 되찾을 수 있는가. 여기서 비밀 정보(secret)란 워터마크가 실어 나르는 정보로, 사용자의 신원이나 숨겨진 짧은 메시지를 가리키며, 워터마크를 읽는 쪽이 되찾을 수 있는 정보의 양을 상호정보량(mutual information)으로 잡는다. 예를 들어 어떤 콘텐츠를 누가 만들었는지 모르는 상태에서는 많은 제작자가 후보로 남아 있다. 이때 콘텐츠를 관찰하면 후보의 범위를 좁힐 수 있기 때문에

제작자에 대한 불확실성도 줄어드는 것이다. 상호정보량이란 이렇게 콘텐츠를 살펴봄으로써 비밀 정보에 대한 추측이 얼마나 더 뚜렷해지는지를 수치로 나타낸 값이다. 한편, 콘텐츠를 정교하게 분석한다 하더라도 워터마크를 심을 때 담긴 정보보다 더 많은 것을 꺼낼 수는 없다. 따라서 되찾을 수 있는 정보의 총량은 이 상호정보량으로 이미 정해져 있다.

이 연구의 중심에는 정보 프로파일(information profile)이라는 개념이 있다. 콘텐츠는 텍스트의 단어나 문자 조각처럼 생성의 기본 단위가 되는 토큰(token)들이 순서대로 이어져 구성된다. 정보 프로파일은 각 위치의 토큰이 비밀 정보에 관해 새로 드러내는 정보의 양을, 그 앞의 토큰들을 이미 안다는 전제 아래 위치마다 기록한 것이다. 즉 워터마크가 콘텐츠의 어느 지점에서 얼마만큼의 정보를 흘리는지를 위치별로 나열한 값이다. 이 프로파일을 위치 전체에 걸쳐 더하면 앞서 말한 상호정보량이 된다.

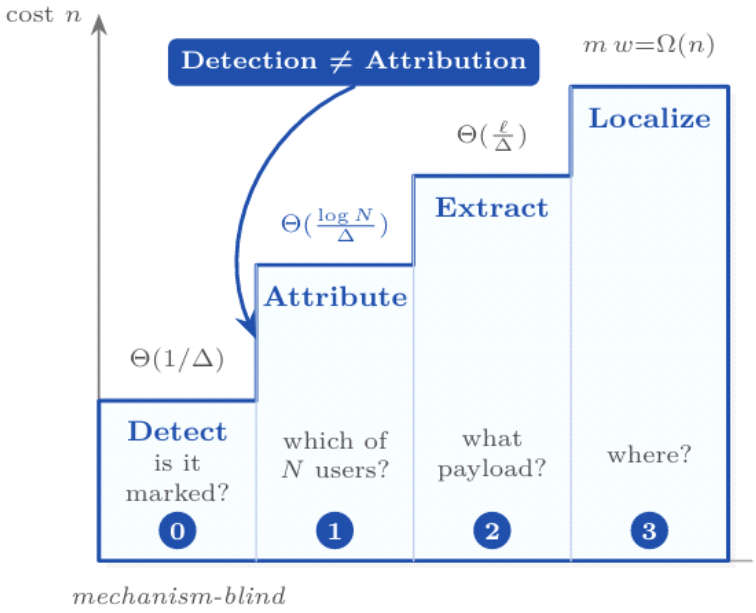
정보 프로파일이 중요한 까닭은 서로 달라 보이던 여러 포렌식 물음이 이 하나의 값을 읽는 방식의 차이로 정리되기 때문이다. 우선 누가 만들었는지 밝히는 일과 숨은 정보를 꺼내는 일은 프로파일의 총량, 곧 정보가 얼마나 많은지에 달려 있다. 반면 워터마크가 있는 부분을 짚어내는 일은 그 정보가 어느 위치에 퍼져 있는지, 곧 프로파일의 모양에 달려 있다. 탐지만은 예외로, 정보의 양이 아니라 워터마크가 자연스러운 콘텐츠와 얼마나 떨어져 있는가에 따라 결정된다. 이렇게 하나의 프로파일을 총량으로 읽느냐 모양으로 읽느냐에 따라, 워터마크가 무엇을 할 수 있는지가 갈린다.

• 탐지와 귀속을 가르는 네 단계 포렌식 사다리

워터마크의 포렌식 기능은 네 단계의 사다리로 정리된다. 가장 아래 단계는 탐지로, 콘텐츠가 AI 생성물인지 아닌지만 가려낸다. 그 위가 귀속으로, 여러 사용자 가운데 누가 만들었는지 지목한다. 다시 그 위가 추출로, 워터마크에 담긴 비트열 형태의 정보를 꺼낸다. 그리고 가장 위가 국소화로, 워터마크가 어느 부분에 있는지 짚어낸다. 각 단계는 목표한 오류 수준에 이르기까지 얼마나 긴 콘텐츠가 필요한지, 곧 표본 복잡도(sample complexity)로 그 비용이 매겨지는데, 표본 복잡도란 원하는 정확도에 도달하는 데 요구되는 데이터의 양을 뜻한다.

이 사다리에서 가장 중요한 지점은 탐지와 귀속이 근본적으로 다른 물음이라는 사실이다. 탐지는 워터마크가 있는지 없는지를 묻는 판정이기 때문에, 탐지에 필요한 비용은 사용자가 몇 명이든 늘어나지 않는다. 반면에 귀속은 여러 후보 가운데 하나의 신원을 집어내야 하므로 사용자 수에 따라 필요한 정보가 늘어난다. 그런데 이 정보는 토큰당 일정한 속도로만 쌓이기 때문에, 탐지가 끝난 뒤에도 귀속이 시작되지 못하는 구간이 생긴다. 이는 워터마크가 있음은 분명하나 누구의 것인지는 밝힐 수 없는 상태로, 워터마크를 심는 기술의 완성도가 낮아서가 아니라 되찾을 수 있는 정보의 총량이라는 원리에서 비롯한다. 결국 어떤 콘텐츠가 AI 생성물이라는 사실을 확인하는 일과 그것을 특정 사용자의 것으로 연결 짓는 일은 기술적으로 서로 다른 비용을 치르며, 워터마크가 있다는 확인만으로 곧바로 그 콘텐츠의 출처가 특정 개인에게 연결되는 것은 아니다.

[그림 1] 탐지에서 국소화로 이어지는 포렌식 4단계와 단계별 비용



출처: Xiaoyu Li 외 5인, "Watermark Forensics for Generative Models", arXiv, 2026.07.14., https://arxiv.org/pdf/2607.13003

• **바이어싱과 임베딩, 두 워터마크 메커니즘**

워터마크를 심는 방식은 두 갈래로 나뉜다. 하나는 바이어싱(biasing)으로, 워터마크가 콘텐츠 전체에 옅게 퍼져 모든 위치를 조금씩 변형하는 방식이다. 다른 하나는 임베딩(embedding)으로, 워터마크가 짧은 몇몇 위치에 뚜렷하게 새겨지는 방식이다. 이 두 방식은 같은 정보 프로파일을 서로 다른 모양으로 담는다. 바이어싱은 정보를 얇게 펴서 모든 토큰에 나누어 싣고, 임베딩은 정보를 좁은 구간에 몰아 싣는다. 담긴 정보의 총량이 같더라도 그 분포의 모양이 다른 것이다.

이 모양의 차이가 귀속 비용을 가른다. 바이어싱 방식에서는 워터마크가 토큰마다 일정 한도 안에서만 정보를 실을 수 있어, 신원을 특정할 만큼의 정보가 쌓이기까지 콘텐츠가 그만큼 길어야 한다. 게다가 사용자 수가 늘수록, 그리고 워터마크를 자연스러운 콘텐츠에 가깝게 감출수록 필요한 콘텐츠의 길이가 늘어난다. 반면 임베딩 방식은 사용자의 신원을 몇 개의 토큰에 직접 새기므로, 짧은 구간만으로 사용자 식별 정보를 복원할 수 있다. 이 경우 귀속 비용은 워터마크를 얼마나 옅게 감추느냐와 무관해진다. 곧 임베딩 방식은 귀속 비용을 워터마크의 세기에 직접 좌우되지 않도록 한다.

이러한 차이는 침해 추적과 국소화의 관점에서 중요하게 작용한다. 워터마크가 있는 부분을 짚어내는 국소화가 가능하려면 워터마크를 몇몇 위치로 좁힐 수 있어야 하는데, 바이어싱은 콘텐츠 전체에 퍼져 있어 그러한 위치를 특정할 수 없다. 따라서 콘텐츠 전체에 옅게 퍼지는 방식은 품질을 지키면서도 신원을 밝히려 할 때 더 긴 콘텐츠라는 비용을 치르며, 워터마크를 감출수록 그 비용은 커진다. 다만 몇몇 위치에

정보를 몰아 싣는 임베딩 방식의 이론적 사례는 단 한 번의 편집만으로도 지워질 수 있다. 곧 귀속 비용을 낮추는 구조와 편집에 견디는 견고함은 서로 상충하며, 이는 워터마크를 어떻게 심느냐에 따라 추적의 비용과 견고함이 함께 정해진다는 것을 의미한다.

• 국소화의 한계와 자취-해상도 불확정성

사다리의 가장 위에 놓인 국소화는 워터마크가 콘텐츠의 어느 부분에 있는지 짚어내는 기능이다. 특히 까다로운 형태는 콘텐츠의 일부만 잘려 나가도 워터마크를 짚어낼 수 있어야 하는 경우다. 콘텐츠에서 특정 구간만 남기고 나머지를 잘라내는 상황을 떠올려 보면, 만약 남겨진 구간이 워터마크가 놓인 자리를 완전히 비껴간다면 그 구간은 자연스러운 콘텐츠와 구별되지 않아 워터마크를 읽을 수 없다. 따라서 어떤 구간을 잘라내도 워터마크를 짚어내려면, 워터마크의 자취가 콘텐츠 전반에 촘촘히 놓여 있어야 한다.

여기서 워터마크 자취의 크기와 짚어내는 해상도(resolution)는 서로 상충 관계에 있다. 해상도란 워터마크를 얼마나 잘게 나누어 짚어낼 수 있는지를 가리킨다. 워터마크가 놓인 자리가 좁으면 남겨진 구간이 그 자리를 비껴가기 쉬워, 높은 해상도로 짚어내는 일이 어려워진다. 반대로 콘텐츠를 작은 조각으로 잘라도 그 조각마다 워터마크의 위치를 짚어내려면 워터마크가 콘텐츠 거의 전체에 퍼져 있어야 한다. 다시 말해 자취의 크기와 해상도를 곱한 값이 콘텐츠 전체 길이에 비례해 일정 수준 이상이어야 하며, 이러한 상충 관계가 자취-해상도 불확정성(footprint-resolution uncertainty)이다.

이 관계는 국소화를 잘 해내는 워터마크가 왜 콘텐츠 전체에 퍼진 형태를 띠는지 설명한다. 몇몇 위치에 정보를 몰아 심은 워터마크는 워터마크가 집중된 부분만 잘라내면 해당 신호가 사라질 수 있으므로, 잘림에 견디며 세밀하게 짚어내는 일과는 양립하기 어렵다. 실제로 콘텐츠를 세밀하게 짚어내는 워터마크들은 콘텐츠 전반에 퍼진 형태를 취하되, 그 워터마크를 구역별로 나누어 읽는 방식으로 국소화를 해낸다. 또한 콘텐츠가 잘리거나 편집된 뒤에도 조작된 부분을 짚어내려는 기능은 워터마크가 콘텐츠 전반에 퍼져 있을 것을 요구한다. 이러한 점에서 침해 지점을 세밀하게 국소화하려는 목표와 워터마크를 좁은 구간에 감추려는 목표는 기술적으로 함께 달성되기 어렵다.

III. 결론: 출처표시 의무화 시대의 기술적 과제와 전망

• 다층 표시 체계와 기술적 보완 과제

결국 워터마크의 포렌식 능력은 워터마크를 어떻게 심느냐가 아니라 워터마크가 담은 정보의 구조에서 정해진다. 탐지, 귀속, 추출, 국소화라는 네 가지 물음은 하나의 정보 프로파일을 총량으로 읽느냐 모양으로 읽느냐의 차이일 뿐이며, 각 물음은 서로 다른 비용을 치른다. 이를 통해 워터마크가 있다는 확인과 그것을 특정 사용자에게 연결 짓는 일이 서로 다른 문제라는 점이 드러난다. 또한 콘텐츠 전반에 퍼진 워터마크와

좁은 구간에 몰린 워터마크는 귀속 비용과 견고함 측면에서 서로 다른 균형을 이루며, 조작된 지점을 세밀하게 짚어내려면 워터마크가 콘텐츠 전반에 퍼져 있어야 한다. 이는 출처표시 기술에 거는 기대가 기술의 실제 능력과 어긋날 수 있음을 의미한다.

이러한 점에서 출처표시 의무화가 본격화되는 국면에서 제도와 기술이 함께 풀어야 할 과제가 뚜렷해진다. 최근의 규정들은 하나의 기술만으로는 요구를 충족하기 어렵다는 판단 아래, 메타데이터와 비가시성 워터마크, 핑거프린팅(fingerprinting)을 결합한 다층 방식을 규정한다. 여기서 핑거프린팅이란 콘텐츠의 고유한 특징을 기록해 두었다가 대조하는 방식이다. 앞서 살펴본 대로 워터마크의 능력에는 정보의 구조에서 비롯하는 한계가 있으므로, 제도가 요구하는 기능과 기술을 실제로 제공할 수 있는 성능 범위와 정합시키는 일이 필요하다. 따라서 앞으로는 편집과 부분 절단(cropping)에 견디면서도 신원을 밝힐 수 있는 워터마크, 그리고 여러 방식을 결합해 각 기술의 한계를 서로 보완하는 설계가 과제로 남는다. 워터마크를 심는 일과 그 워터마크로 무엇을 증명할 수 있는가를 함께 고려하는 접근이, 출처표시가 신뢰의 근거로 자리 잡기 위한 조건이 될 것으로 보인다.

참고문헌

- Xiaoyu Li 외 5인, “Watermark Forensics for Generative Models”, arXiv, 2026.07.14., https://arxiv.org/pdf/2607.13003
- The Linux Foundation, “C2PA Welcomes TikTok to Steering Committee, Advancing the Adoption of Content Credentials at a Global Scale”, PR Newswire, 2026.07.28., https://www.prnewswire.com/news-releases/c2pa-welcomes-tiktok-to-steering-committee-advancing-the-adoption-of-content-credentials-at-a-global-scale-302836730.html

기술용어

순번	용어	설명
1	C2PA (Coalition for Content Provenance and Authenticity)	콘텐츠의 출처와 제작 이력을 확인할 수 있는 공통 규격을 개발하기 위해 여러 기업과 기관이 결성한 연합체
2	포렌식 (forensics)	남겨진 흔적을 거꾸로 읽어 그것을 만든 사건이나 원인에 도달하는 분석 작업

타이달의 AI 생성 음원 저작권료 중단으로 본 탐지 기술의 한계

뉴스 브리프

음악 스트리밍 플랫폼 타이달은 2026년 7월 15일부터 자체 탐지 시스템이 100% AI 생성으로 식별한 음원에 저작권료를 지급하지 않기로 했다. 해당 음원에는 AI 표시가 적용되며, 플랫폼의 직접 판매 대상에서도 제외된다. 테크 타임스는 이를 주요 스트리밍 서비스가 AI 생성 음원 대응을 표시에서 저작권료 지급 중단으로 확대한 첫 사례라고 설명했다. AI 생성 여부가 수익 배분 기준으로 활용되면서 이를 뒷받침하는 탐지 기술의 정확성과 강건성도 중요해지고 있다. 최근 연구에서는 일부 탐지 기술이 일반적인 조건에서 높은 정확도를 보였지만, 음원 변형을 적용하면 탐지를 우회할 수 있는 것으로 나타났다. 본 보고서는 타이달의 저작권료 지급 중단 사례를 중심으로 AI 생성 음원의 유통 과정과 탐지 기술의 성능을 살펴보고, 음원 변형에 따른 탐지 한계와 기술적 과제를 분석한다.

뉴스 플러스

I. 서론: AI 생성 음원 판정과 저작권료 지급의 연계

• 타이달, 100% AI 생성 음원의 저작권료 지급 중단

음악 스트리밍 플랫폼 타이달(TIDAL)은 2026년 7월 15일부터 자체 탐지 시스템이 100% AI 생성으로 식별한 음원에 스트리밍 저작권료를 지급하지 않는다. 해당 음원에는 AI 표시를 적용하고 플랫폼의 직접 판매 대상에서도 제외한다. 실제 아티스트를 사칭한 AI 생성 음원은 자동화된 도구를 이용해 삭제할 계획이다.

미국 기술 전문 매체 테크 타임스(Tech Times)는 이를 주요 스트리밍 서비스가 AI 생성 음원 대응을 단순한 표시에서 저작권료 지급 중단으로 확대한 첫 사례라고 설명한다.[1] 타이달은 AI 생성 여부에 대한 자체 탐지 결과를 저작권료 지급 여부와 직접 연결했다.

1) Mike Gonzales, "AI Music Labels Land Tomorrow, but Industry Rules Remain Voluntary and Incomplete", Tech Times, 2026.07.14., https://www.techtimes.com/articles/320432/20260714/ai-music-labels-land-tomorrow-industry-rules-remain-voluntary-incomplete.htm

• **저작권료 지급 기준이 된 AI 생성 음원 탐지**

타이달의 정책은 자체 시스템이 전적으로 AI가 생성한 것으로 식별한 음원을 대상으로 한다. 따라서 AI 생성 음원을 얼마나 정확하게 구분하는지가 정책 집행의 핵심 조건이 된다.

AI 생성 음원 탐지 결과가 저작권료 지급 여부와 연결되면 탐지 기술의 신뢰성은 더욱 중요해진다. 일반적인 상태에서 높은 정확도를 확보하는 것뿐 아니라 음원에 변형이 가해진 이후에도 같은 판정 결과를 유지할 수 있어야 하기 때문이다.

2026년 6월 공개된 「음악 스트리밍 내 AI 생성 저품질 콘텐츠에 대한 실증 분석(An Empirical Analysis of AI Slop in Music Streaming)」은 이러한 문제를 AI 생성 음원의 유통과 탐지 과정에서 분석했다. 연구진은 직접 생성한 AI 음원을 11개 독립 음원 유통사를 통해 스트리밍 플랫폼에 배포하고, 자동 탐지 기술 4종의 정확도와 강건성(robustness)*을 평가했다.[2)]

* 강건성(robustness): 입력 데이터에 일부 변형이 발생해도 기존 성능을 안정적으로 유지하는 특성

II. 본론: AI 생성 음원 유통과 탐지에서 확인된 기술적 한계

• **AI 생성 음원 정책과 실제 유통 결과의 차이**

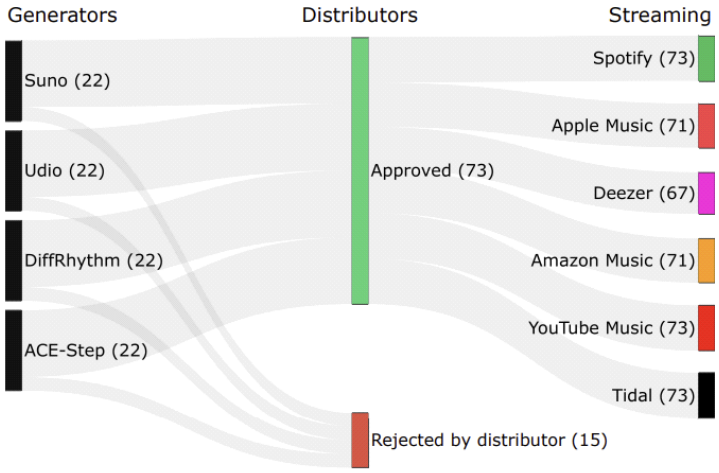
연구진은 AI 생성 음원이 실제 스트리밍 환경에 유통되는 과정을 확인하기 위해 11개 독립 음원 유통사를 조사했다. 조사 대상 모두 AI 사용을 정책에서 언급했지만, AI 생성 음원을 허용하는 범위와 기준에는 차이가 있었다. 이 가운데 5개 유통사는 AI 생성 음원을 배포하지 않는다고 명시했다.

연구진은 각 유통사에 8곡씩 총 88개의 AI 생성 음원을 업로드했다. 이 가운데 15곡이 2개 유통사에서 거절됐다. 주목할 점은 AI 생성 음원의 배포를 허용하지 않는다고 명시한 5개 유통사 가운데 4개도 연구진이 제출한 AI 생성 음원을 모두 승인했다는 것이다. 거절된 음원도 AI 생성 여부가 사유로 명시되지는 않았다.

유통사의 검토를 통과한 AI 생성 음원은 이후 스트리밍 플랫폼에서도 거의 거절되지 않았다. 연구 결과는 AI 생성 음원 관련 정책을 마련하는 것과 실제 유통 단계에서 AI 생성 여부를 식별해 적용하는 것 사이에 차이가 있음을 보여준다.

2) Stanley Wu 외 5명, “An Empirical Analysis of AI Slop in Music Streaming”, arXiv, 2026.06.16., https://arxiv.org/abs/2606.18052

[그림 1] AI 생성 음원의 유통사와 스트리밍 플랫폼 업로드 결과



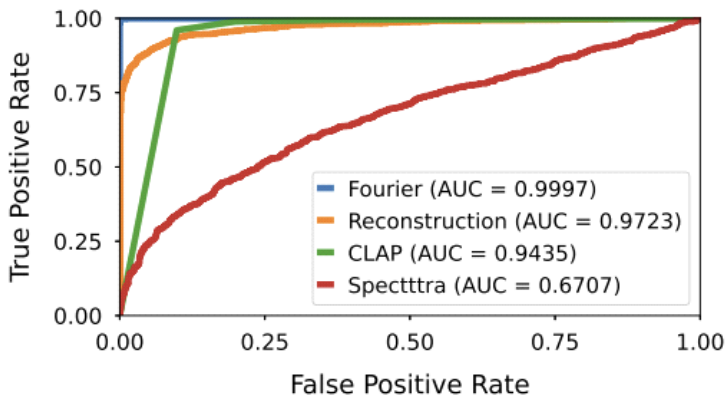
출처: Stanley Wu 외 5명, "An Empirical Analysis of AI Slop in Music Streaming", arXiv, 2026.06.16., https://arxiv.org/abs/2606.18052

• **별도 변형을 적용하지 않은 조건에서의 AI 생성 음원 탐지 성능**

연구진은 자동 탐지 기술의 성능을 확인하기 위해 AI 생성 음원 1,000곡과 사람이 제작한 음원 1,000곡을 평가했다. 평가에는 수노(Suno)와 우디오(Udio)가 생성한 AI 음원을 사용했으며, 서로 다른 방식으로 작동하는 탐지 기술 4종을 비교했다.

평가 결과, 음원 스펙트로그램(spectrogram)에 나타나는 인공적 주파수 피크를 식별하는 푸리에(Fourier) 기반 탐지 기술이 99.6%의 정확도를 기록했다. 또 다른 푸리에 기반 탐지기인 퀵실버(Quicksilver)도 99.75%의 정확도를 보여 높은 탐지 성능을 나타냈다.

[그림 2] AI 생성 음원 탐지 기술 4종의 성능 비교



출처: Stanley Wu 외 5명, "An Empirical Analysis of AI Slop in Music Streaming", arXiv, 2026.06.16., https://arxiv.org/abs/2606.18052

이는 별도의 변형을 적용하지 않은 일반적인 조건에서 일부 탐지 기술이 AI 생성 음원과 사람이 제작한 음원을 높은 정확도로 구분할 수 있음을 보여준다. 그러나 이러한 탐지는 음원에 간단한 처리를 가하는 방식으로 의도적으로 우회될 수 있어, 연구진은 변형된 조건에서 탐지 성능을 추가로 평가했다.

• 음원 변형으로 달라진 AI 생성 음원 탐지 결과

연구진은 탐지 기술이 음원 변형 이후에도 같은 성능을 유지하는지 확인했다. 수노로 생성한 AI 음원 20곡에 MP3 손실 압축(MP3 Compression), 피치 변경(Pitch Shift), 리버브(Reverb), 오디오 재구성(Reconstruction) 등 네 가지 변형을 적용한 뒤 프랑스 음악 스트리밍 플랫폼 디저(Deezer)의 AI 생성 음원 탐지 결과를 확인했다.

각 변형 방식에는 5곡씩 사용됐다. 실험 결과 네 가지 방식 모두 5곡 전체가 디저의 AI 생성 음원 탐지를 우회했다. 피치 변경을 제외한 나머지 변형은 사람이 들었을 때 뚜렷한 차이를 구분하기 어려운 수준이었다.

[표 1] 음원 변형에 따른 AI 생성 음원 탐지 우회 결과

변형 방식	탐지 우회 결과
MP3 손실 압축	5/5곡
피치 변경	5/5곡
리버브	5/5곡
오디오 재구성	5/5곡

출처: Stanley Wu 외 5명, “An Empirical Analysis of AI Slop in Music Streaming”, arXiv, 2026.06.16., https://arxiv.org/abs/2606.18052

연구진은 음원 변형에 대한 탐지 성능을 높이기 위해 푸리에 기반 탐지기에 변형된 음원을 추가로 학습시키는 적대적 학습(adversarial training)*을 적용했다. 그 결과 음원 변형 조건에서 정확도가 95%까지 회복됐으며, 별도 변형을 적용하지 않은 조건에서는 정확도가 99%에서 98%로 소폭 낮아졌다.

다만 특정 변형을 학습에 포함하더라도 다른 유형의 변형에서도 같은 성능이 유지되는 것은 아니었다. 연구진은 학습에 포함되지 않은 새로운 우회 방식에는 여전히 취약할 수 있다고 설명했다. 따라서 AI 생성 음원 탐지에서는 일반적인 조건의 정확도와 함께 다양한 음원 변형에 대응하는 강건성을 지속적으로 확인할 필요가 있다.

* 적대적 학습(adversarial training): 변형된 입력 데이터를 학습 과정에 포함해 이러한 변형에 대한 모델의 대응 성능을 높이는 방법

III. 결론: AI 생성 음원 탐지의 정확도와 강건성 확보

• 수익 관리에 활용되는 AI 생성 음원 탐지의 기술적 과제

타이달의 사례는 AI 생성 음원 탐지가 콘텐츠 표시뿐 아니라 저작권료 지급 여부를 판단하는 데 활용되기 시작했음을 보여준다. AI 생성 여부에 대한 기술적 판정이 플랫폼의 수익 관리와 연결되면서 탐지 기술의 성능도 중요해지고 있다. 그러나 연구 결과에서는 AI 생성 음원 관련 정책을 마련한 유통사에서도 AI 생성 음원이 승인됐으며, 자동 탐지 기술도 음원 변형에 따라 결과가 달라졌다. 별도 변형을 적용하지 않은 조건에서는 높은 정확도를 보인 탐지 방식이 존재했지만, 손실 압축이나 피치 변경, 리버브, 오디오 재구성 등의 변형을 통해 탐지를 우회할 수 있는 것으로 나타났다.

따라서 AI 생성 음원 탐지를 수익 관리에 활용하기 위해서는 일반적인 환경에서의 정확도뿐 아니라 음원 변형 이후에도 성능을 유지하는 강건성을 함께 고려할 필요가 있다. 유통 단계의 AI 생성 음원 식별과 스트리밍 플랫폼의 자동 탐지를 함께 개선하는 것도 AI 생성 음원 관리의 주요 기술적 과제가 될 수 있다.

참고문헌

• Mike Gonzales, “AI Music Labels Land Tomorrow, but Industry Rules Remain Voluntary and Incomplete”, Tech Times, 2026.07.14., https://www.techtimes.com/articles/320432/20260714/ai-music-labels-land-tomorrow-industry-rules-remain-voluntary-incomplete.htm
• Stanley Wu 외 5명, “An Empirical Analysis of AI Slop in Music Streaming”, arXiv, 2026.06.16., https://arxiv.org/abs/2606.18052

기술용어

순번	용어	설명
1	강건성 (robustness)	입력 데이터에 일부 변형이 발생해도 기존 성능을 안정적으로 유지하는 특성
2	적대적 학습 (adversarial training)	변형된 입력 데이터를 학습 과정에 포함해 이러한 변형에 대한 모델의 대응 성능을 높이는 방법

시간적으로 일관된 범용 적대적 교란을 활용한 영상 보호 기술

뉴스 브리프

확산 모델을 활용한 영상 생성 기술이 정교해지고 이용 비용도 낮아지면서, 창작자가 온라인에 올린 영상이 손쉽게 수집되고 다시 쓰이는 환경이 만들어지고 있다. 공개된 영상 한 편은 모델을 미세 조정하는 재료가 되기도 하고, 한 장면이 참조 이미지로 뽑혀 새로운 영상을 만드는 데 쓰이기도 한다. 이 과정에서 영상 창작물이 창작자의 의사와 무관하게 이용될 수 있다는 우려가 커진다. 그러나 지금까지의 보호 기술은 대체로 이미지 한 장을 지키는 데 머물러, 영상 자체를 보호하기는 어려웠다. 본 보고서는 사람이 알아채기 어려운 미세한 교란을 영상에 심어 두 가지 무단 영상 제작을 무력화하는 시간적으로 일관된 범용 적대적 교란 기술의 작동 원리를 살펴본다.

뉴스 플러스

Ⅰ. 서론: 영상 생성의 확산과 창작물 보호의 과제

- **영상 생성 모델의 확산과 무단 이용 우려**

확산 모델(diffusion model)을 활용한 영상 생성 기술이 빠르게 발전하면서, 이제는 사람이 직접 촬영한 장면과 구분하기 어려운 영상이 짧은 지시만으로도 손쉽게 만들어지고 있다. 공개된 모델과 상용 서비스가 모두 정교한 결과를 내놓는 가운데, 이용 비용까지 낮아지면서 영상 생성 기술은 누구나 다룰 수 있는 도구로 자리 잡았다. 이용자가 원하는 대상이나 장면을 직접 지정해 영상을 만들어내는 기능이 널리 쓰이는데, 이런 방식을 영상 커스터마이징(video customization)이라 부른다.

영상 커스터마이징은 미세 조정(fine-tuning) 방식과 참조 이미지 방식(image-to-video)이라는 두 갈래로 작동한다. 미세 조정 방식은 일반적으로 대상의 여러 학습 샘플을 활용하는 반면, 참조 이미지

방식은 한 장면만 참조로 삼아도 그 대상을 담은 영상을 만들어낸다. 문제는 두 방식이 쓰는 재료가 대부분 온라인에 이미 공개된 영상이라는 점이며, 창작자가 스스로 올린 영상이 본인의 뜻과 무관하게 새 영상의 재료로 쓰일 수 있다는 데 있다. 이 지점에서 이미 공개된 영상을 창작자의 뜻에 반해 전혀 다른 영상의 재료로 쓰는 흐름을 어떻게 막을지가 새로운 문제로 부상한다.

• 이미지 보호를 넘어선 영상 보호의 필요성

이러한 무단 이용에 맞선 보호 기술은 이전부터 연구됐지만, 그동안의 노력은 대부분 이미지 한 장을 지키는 데 초점을 맞춰 왔다. 대표적인 방식은 사람이 알아채기 어려운 미세한 교란(perturbation)을 이미지에 더해, 그 이미지를 재료로 삼은 모델을 미리 방해해 두는 것이다. 미세 조정 방식에는 교란이 담긴 이미지로 학습한 모델이 대상을 제대로 담아내지 못하게 하고, 참조 이미지 방식에는 참조 이미지에 교란을 넣어 결과가 일그러지도록 유도한다.

그러나 이미지 보호만으로는 영상 자체의 권리를 지켜내기 어려운 상황이 나타나기 시작했다. 공개된 영상이 참조 이미지를 뽑아내는 재료로도 쓰이고, 모델을 학습시키는 재료로도 함께 쓰이기 때문이다. 이는 이미지 한 장이 아니라 시간의 흐름을 지닌 영상 전체를 온전히 지켜야 두 방식의 무단 이용을 함께 막을 수 있음을 뜻한다. 이 빈자리를 겨냥해 등장한 기술이 시간적으로 일관된 범용 적대적 교란(Temporally Consistent Universal Adversarial Perturbations, 이하 TC-UAP) 기술이다.

II. 본론: 시간적 일관성을 갖춘 영상 보호 기술

• 영상 VAE라는 공통 취약점과 보호 원리

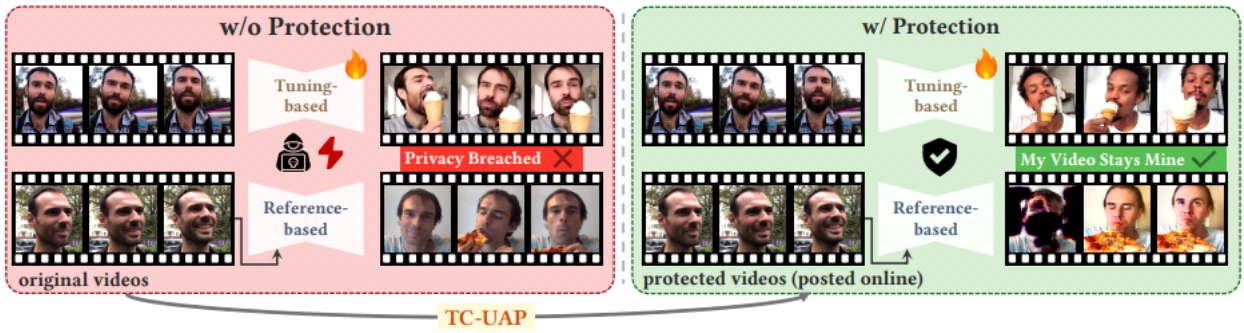
미세 조정 방식과 참조 이미지 방식은 겉으로는 서로 완전히 다른 방향으로 작동하는 것처럼 보이지만, 정작 영상이나 이미지를 그대로 다루지 않고 먼저 압축된 형태로 바꾼 뒤에야 각자의 작업을 시작한다는 중요한 공통점을 지니고 있다. 이 변환을 맡는 부분이 영상 변분 오토인코더(video Variational Auto-encoder, 이하 영상 VAE)로, 이는 영상을 압축된 잠재 표현(latent)으로 바꾸었다가 나중에 필요할 때 다시 원래의 영상으로 되돌려 놓는 역할을 하는 장치다. 이렇게 압축하는 까닭은 영상 한 편이 담고 있는 정보의 양이 너무 많아서, 원본을 압축하지 않고 그대로 계산에 쓰면 시간과 자원의 부담이 지나치게 커지기 때문이다.

참조 이미지 방식은 참조로 쓸 이미지 한 장을 영상 VAE로 먼저 압축하고, 미세 조정 방식은 학습에 쓸 여러 편의 영상을 역시 같은 영상 VAE로 압축한 다음에야, 비로소 서로 다른 각자의 작업으로 나아가게 된다. 두 방식이 서로 다른 결과를 만들어내지만, 영상이나 이미지를 잠재 표현으로 바꾸는 이 첫 단계는 어느 쪽도 건너뛰지 못하고 똑같이 거친다. 바로 이 단계가, 서로 다른 두 방식이 예외 없이 나란히 지나갈

수밖에 없는 하나의 공통된 취약점으로 작용하게 되는 셈이다. 잠재 표현이 두 방식 모두의 출발점인 만큼, 이 표현에 담긴 대상의 정보를 처음부터 미리 흐트러뜨려 두면 그 뒤로 이어지는 작업 전체가 함께 흔들리게 되기 때문이다.

TC-UAP는 바로 이 취약점을 공략하여, 영상에 심어 두는 교란을 그 영상의 잠재 표현이 원본과 최대한 차이를 유발하도록 다듬는 기술이다. 구체적으로, 원본 영상과 교란이 담긴 영상을 각각 따로 압축했을 때, 그렇게 얻어진 두 잠재 표현 사이의 거리가 가능한 한 크게 벌어지도록 교란의 값 자체를 아주 조금씩 여러 차례에 걸쳐 반복해서 조정해 나간다. 그 결과 대상의 핵심 정보는 잠재 표현으로 압축되는 그 단계에서 이미 훼손되고, 두 방식은 온전한 잠재 표현을 끝내 넘겨받지 못한 채로 이후의 작업을 이어가게 된다. 이때 교란은 육안에는 거의 드러나지 않을 만큼 미세하게 유지되므로, 창작자가 올린 원본 영상의 겉모습은 거의 그대로 남는다.

[그림 1] TC-UAP를 통한 두 방식의 무단 영상 제작 차단



출처: Yuxin Huang 외 5명, “Delving into the Temporal Challenges of Unified Video Protection Against Image-to-Video and Fine-Tuning-based Customization”, arXiv, 2026.07.14., https://arxiv.org/pdf/2607.13336

• 영상 보호를 가로막는 세 가지 시간적 난제

이미지를 지키던 보호 기술을 시간의 흐름을 지닌 영상에 그대로 옮기려 할 때는, 크게 세 가지 시간적 난제가 앞을 가로막는다. 첫째 난제는 영상 VAE의 시간 압축(temporal compression)으로, 프레임마다 따로 심은 교란이 압축 과정에서 힘을 잃게 만드는 문제다. 시간 압축은 이웃한 여러 프레임의 정보를 하나의 잠재 표현에 함께 눌러 담는 과정이다. 이미지 보호를 프레임마다 하나씩 따로 적용하면, 각 교란은 그 프레임 하나만 놓고 보면 잠재 표현을 흐트러뜨리지만, 여러 프레임이 함께 압축되는 과정에서 이웃 프레임의 정보와 뒤섞이면서 그 효과가 크게 약해지고 만다. 게다가 영상 VAE는 앞선 프레임의 정보까지 끌어모아 뒤쪽 프레임을 압축하므로, 영상이 뒤로 갈수록 교란은 더 크게 희석된다.

둘째 난제는 교란이 특정 영상 한 편에만 맞춰지는 시간 과적합(temporal overfitting)으로, 한 번 만든 교란을 다시 쓰기 어렵게 만드는 문제다. 프레임을 하나씩 떼는 대신 영상 한 편을 통째로 놓고 교란을

만들면 그 영상 안에서는 보호가 강하게 걸리지만, 이렇게 얻은 교란은 그 영상의 특정한 프레임 순서에만 맞춰져 순서가 조금만 달라져도 보호 효과가 크게 떨어진다. 실제로 보호된 영상을 뒷부분 몇 프레임만 잘라내어 전체 길이를 줄이는 것만으로도, 그 영상에 걸려 있던 보호는 대부분 사라진다. 또한 같은 교란을 학습에 사용되지 않은 다른 영상에 그대로 붙이면 보호가 걸리지 않아, 결국 영상마다 교란을 새로 만들어야 하는 부담이 남는다.

셋째 난제는 교란을 겨냥한 별도의 조작에서 비롯되는 시간 비일관성(temporal inconsistency)으로, 교란이 공격에 쉽게 지워지게 만드는 문제다. 공격자는 영상을 이용하기 전에 프레임 사이를 메우거나 이웃한 프레임을 평균 내는 시간 공격(temporal attack)을 가해 교란을 미리 약화시킬 수 있다. 각 프레임의 교란을 서로 따로 최적화하면 그 교란은 프레임과 프레임 사이에 아무런 연관이 없는 상태가 되는데, 이렇게 서로 어긋나 있는 교란은 이웃한 프레임을 평균 내는 조작에 그대로 상쇄되어 픽셀 단계에서 대부분 지워진다. 그 결과 이런 교란은 공격을 한 번 거치고 나면 거의 남아나지 못하며, 이는 교란이 프레임 사이에 일정한 일관성을 갖춰야 공격을 견딜 수 있음을 보여준다.

• 세 난제를 풀어내는 TC-UAP의 설계

TC-UAP는 영상마다 교란을 매번 새로 만들어야 했던 기존 방식과 달리, 같은 인물이나 대상을 하나의 기준으로 삼아 교란을 단 한 번만 만들어 두고, 그 대상이 나오는 여러 편의 영상에 공통 적용하는 방식을 사용한다. 이때 교란은 한 장이 아니라 여러 프레임으로 이뤄진 하나의 묶음이며, 이 묶음을 영상 길이에 맞춰 처음부터 끝까지 반복해서 덧입힌다. 이렇게 정해진 묶음을 영상 위에 반복해서 덧붙이는 구조 덕분에, 길이가 짧은 영상이든 긴 영상이든 가리지 않고 언제나 같은 하나의 교란을 그대로 가져다 적용할 수 있다. 그 결과 같은 대상이 나오는 처음 보는 영상까지도 이 하나의 교란만으로 지킬 수 있어, 앞서 본 시간 과적합의 부담이 사라진다.

남은 문제는 이 여러 프레임짜리 교란 묶음을 과연 어떤 방식으로 최적화하느냐이다. 영상 한 편 전체를 한 번에 최적화할 경우 과도한 메모리 오버헤드가 발생할 뿐만 아니라, 해당 영상의 순서에 다시 과적합될 위험까지 함께 생긴다. TC-UAP는 이를 피하려고 슬라이딩 윈도우(sliding window)를 쓰는데, 이는 영상 전체가 아니라 일정 길이의 구간만 잘라 그 안에서 교란을 다듬는 방법이다. 쉽게 말하면 영상을 짧은 토막으로 나눈 뒤, 그 토막을 조금씩 옆으로 옮겨 가며 교란을 여러 차례 반복해서 최적화하는 것이다. 이 구간의 길이는 압축된 한 값이 실제로 앞뒤 프레임에서 영향을 받는 범위를 포괄하도록 설정하여, 시간 압축의 효과를 그 안에서 충분히 반영한다. 게다가 여러 영상의 여러 위치에서 구간을 바꿔 가며 뽑아 쓰므로, 교란이 특정한 순서에 매이지 않고 다양한 맥락을 두루 견디게 된다.

다중 프레임 묶음이라고 해서 프레임 사이의 교란이 저절로 이어지는 것은 아니어서, 그대로 두면 시간 공격에 약한 상태로 남는다. TC-UAP는 이를 두 갈래로 보완하는데, 첫째는 교란을 최적화하는 동안

이웃한 프레임을 서로 평균 내는 공격을 미리 흉내 내어, 그 공격을 거친 뒤에도 교란이 제 효과를 그대로 유지하도록 미리 함께 훈련해 두는 것이다. 둘째는 교란을 만드는 방식 자체에 연결성을 심는 것으로, 각 프레임의 교란이 앞선 프레임들의 교란을 이어받도록 하면서도 프레임마다 고유한 값을 함께 두어 일관성과 표현력을 동시에 확보한다. 이렇게 이중으로 대비한 덕분에, 미리 흉내 낸 공격은 물론 처음 보는 시간 공격에도 교란이 무너지지 않고 견딜 수 있게 된다.

III. 결론: 선제적 영상 보호의 의의와 과제

• 선제적 영상 보호 기술의 의의와 남은 과제

TC-UAP는 미세 조정 방식과 참조 이미지 방식이 나란히 지나가는 영상 VAE라는 공통 취약점을 노려, 서로 달라 보이는 두 갈래의 무단 이용을 단 하나의 교란만으로 함께 막아낸다. 결국 이는 영상마다 교란을 매번 새로 만들어 붙여야 했던 부담 없이도, 같은 대상이 담긴 여러 편의 영상은 물론 아직 세상에 나오지 않은 처음 보는 영상까지도 하나의 교란만으로 함께 지켜낼 수 있게 되었음을 의미한다. 이를 통해 교란을 지우려고 프레임을 뒤섞거나 평균 내는 시간 공격 앞에서도, 보호가 쉽게 무너지지 않고 버틸 수 있게 되었다. 따라서 그동안 이미지 한 장을 지키는 데 머물러 있던 보호의 범위가, 시간의 흐름을 지닌 영상 전체로 넓어지는 하나의 전환점이 마련된 셈이다.

다만 이러한 보호가 어떤 상황에서도 빠짐없이 통하는 완결된 해법이라고 보기에는 아직 이른 감이 있으며, 몇 가지 조건과 과제가 함께 남는다. 교란은 같은 대상을 하나의 기준으로 삼아 미리 만들어 두어야 하므로, 그 대상이 담긴 영상이 어느 정도는 확보되어 있어야 비로소 제 성능을 온전히 낸다는 조건이 따른다. 또한 앞으로 새로운 형태의 영상 VAE가 나오거나 지금까지 없던 방식의 시간 공격이 새로 등장하면, 그때마다 교란을 만들고 다듬는 방법 역시 그 변화에 맞추어 새로 발전시켜야 하는 과제가 남는다.

참고문헌

• Yuxin Huang 외 5명, “Delving into the Temporal Challenges of Unified Video Protection Against Image-to-Video and Fine-Tuning-based Customization”, arXiv, 2026.07.14., https://arxiv.org/pdf/2607.13336