Biweekly Report
통권 제84호(2026. 6-2)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제84호(2026. 6-2)

AI 산출물 추적을 위한 능동적 워터마크 기술 동향

뉴스 브리프

AI가 글과 그림, 음악을 산출하는 시대가 자리 잡으면서, 창작물의 소유와 출처를 둘러싼 문제도 제기되고 있다. 특히 특정 작가의 화풍이나 작품이 동의 없이 학습에 쓰이고, 그 결과물이 원작과 경쟁하는 상황이 반복되면서 분쟁이 늘고 있다. 핵심 쟁점은 어떤 원작이 어떤 산출물에 기여했는지를 입증하기 어렵다는 점에 있다. 사후에 유사성을 비교하는 방식은 압축이나 편집 같은 변형에 쉽게 무력화되고, 한 이미지에 여러 개념이 겹쳐 있을 때 이를 분리해 추적하기도 어렵다. 이러한 한계를 기술적으로 풀기 위해, 산출 과정 자체에 보이지 않는 서명을 미리 심어 출처를 추적하는 능동적 워터마크 방식이 주목받고 있다. 본 보고서는 객체와 화풍 같은 여러 개념을 분리해 추적하는 새로운 워터마크 기술의 작동 원리를 살펴보고, 나아가 기술적 대응이 가질 수 있는 의의와 한계를 함께 짚어본다.

뉴스 플러스

Ⅰ. 서론 : AI 산출물과 저작권 보호의 새 국면

- **AI 산출물을 둘러싼 출처 입증의 어려움**

AI가 글과 이미지, 음악을 산출하는 도구로 자리 잡으면서, 그 결과물의 권리 귀속을 둘러싼 논의가 본격화되고 있다. 산출물이 누구의 작품에서 비롯되었는지, 어떤 원작이 그 결과에 기여했는지를 가려내는 일이 새로운 과제로 떠올랐다. 특정 작가의 화풍이 동의 없이 학습에 쓰이고, 그렇게 만들어진 결과물이 원작과 시장에서 경쟁하는 문제가 이미지·음악·출판 전반에서 제기된다. 이때 창작자가 자신의 기여를 주장하려면 원작과 산출물 사이의 연관성을 입증 할 수 있어야 한다.

문제는 그 연결을 객관적으로 입증하기 어렵다는 데 있다. 산출물에는 어떤 원작이 쓰였는지를 가리키는 단서가 남지 않는 경우가 많기 때문이다. 화풍처럼 추상적인 요소가 결과물에 녹아 있을 때는 그 출처를 짚어 내기가 한층 어렵다.

• **능동적 워터마크 기술의 등장 배경과 의의**

이러한 입증의 공백을 메우는 수단으로 워터마크 기술이 주목받고 있다. 다만 산출이 끝난 뒤 결과물에 워터마크를 삽입하는 수동적 워터마크(passive watermarking)는 압축이나 잘라내기 같은 일상적 변형에 쉽게 지워져, 안정적인 추적 수단으로는 한계가 있다. 이에 따라 산출 과정 자체에서 워터마크를 삽입하는 능동적 워터마크(proactive watermarking)가 대안으로 제시되고 있다. 워터마크를 결과물에 깊이 넣을수록 변형 이후에도 출처를 추적할 가능성이 높아지기 때문이다.

그러나 기존 능동적 방식에도 풀지 못한 과제가 남아 있다. 확산 모델(diffusion model)은 하나의 이미지 안에 특정 객체와 특정 화풍처럼 서로 다른 개념을 동시에 결합해 산출하는데, 단일한 워터마크만 삽입하는 방식으로는 이렇게 겹쳐 있는 개념들을 분리해 각각 추적하기 어렵다. 미국의 소프트웨어 기업 어도비(Adobe)의 연구 조직인 어도비 리서치(Adobe Research) 등이 제안한 토큰트레이스(Token Trace)는 바로 이 다중 개념 귀속(multi-concept attribution)* 문제를 겨냥한 기술이다. 이 기술은 개념마다 고유한 워터마크를 의미와 연결해, 한 장의 이미지에서 객체와 화풍을 따로 떼어 검증한다.

* 다중 개념 귀속(multi-concept attribution): 한 이미지에 객체와 화풍처럼 여러 개념이 섞여 있을 때, 각 개념이 어디서 비롯되었는지를 따로 가려내는 작업

Ⅱ. 본론: 능동적 워터마크 기술의 작동 원리

• **능동적 워터마크와 다중 개념 귀속의 과제**

워터마크는 삽입하는 시점에 따라 두 갈래로 나뉜다. 산출이 끝난 뒤 결과물에 워터마크를 삽입하는 수동적 워터마크는 구현이 단순하지만, 압축이나 잘라내기 같은 단순한 변형에 쉽게 지워지는 약점이 있다. 이에 비해 능동적 워터마크는 산출이 이루어지는 과정에서 워터마크를 삽입해, 변형 이후에도 신호가 살아남도록 한다. 기존 능동적 기법은 워터마크를 주로 픽셀 영역(pixel domain)이나 잠재 영역(latent domain)에 삽입해 왔는데, 픽셀 영역 방식은 변형에 약하고 잠재 영역 방식은 견고하지만 생성물의 내용과는 무관한 단일 신호에 머무는 한계가 있었다.

이 단일 신호 방식의 문제는 확산 모델의 산출 특성과 맞물려 드러난다. 확산 모델은 하나의 이미지 안에 특정 객체와 화풍처럼 서로 다른 개념을 동시에 결합해 산출하는데, 이미지 전체에 워터마크 하나만 삽입하는 방식으로는 이렇게 겹쳐 있는 개념들을 분리해 각각의 출처를 밝히기 어렵다. 여기서 다중 개념 귀속이라는 과제가 제기된다. 이는 한 장의 그림에서 객체의 출처와 화풍의 출처를 별개로 가려내는 일이다.

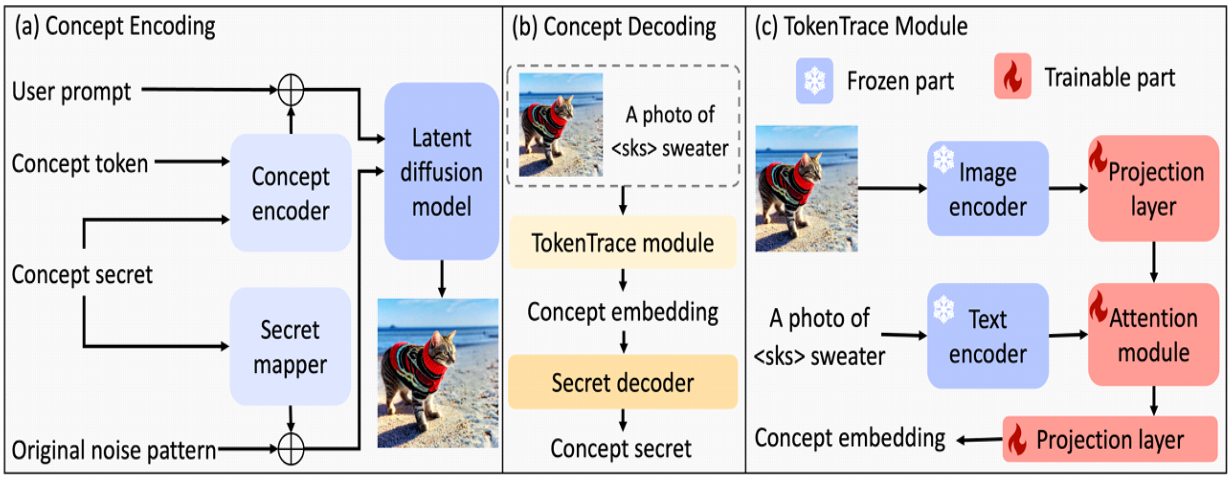
• **의미 영역과 잠재 영역의 이중 조건화 방식**

토큰트레이스가 제시하는 해법의 출발점은 이중 조건화(dual-conditioning)다. 이는 워터마크를 한 곳이 아니라 두 곳에 동시에 삽입해 산출 과정을 양쪽에서 조건 짓는 전략으로, 개념의 의미를 담는 토큰 임베딩과 이미지의 형상을 결정하는 초기 노이즈 양쪽에 워터마크를 삽입하는 방식이다. 핵심은 워터마크를 결과물의 의미와 결합하여, 견고성과 개념별 분리를 함께 확보하는 데 있다. 기존 방식이 워터마크를 결과물의 표면에 삽입하는 데 가까웠다면, 이중 조건화는 수정 초기 단계에서부터 워터마크를 삽입한다.

작동 원리는 두 개의 병렬 네트워크로 나뉜다. 첫 번째 네트워크인 개념 인코더(concept encoder)는 워터마크의 대상이 되는 특정 개념의 토큰 임베딩에만 변형을 가해, 그 개념을 가리키는 의미 표현 안에 워터마크를 삽입한다. 두 번째 네트워크인 비밀 맵퍼(secret mapper)는 같은 비밀키를 받아 산출의 출발점이 되는 초기 잡음 패턴에 변형을 더한다. 이렇게 의미 영역의 토큰과 잠재 영역의 잡음 양쪽에 워터마크가 삽입된 상태로 확산 모델이 이미지를 산출하면, 워터마크는 결과물에 깊이 삽입된다.

이 과정이 갖는 의미는 워터마크가 이미지 표면에 덧붙은 정보가 아니라는 점이다. 워터마크가 개념의 의미와 함께 삽입되기 때문에, 어떤 화풍이나 객체가 쓰였는지를 이미지를 만드는 순간부터 워터마크가 함께 담아낸다. 이는 산출물의 출처를 추적할 단서가 픽셀 표면이 아니라 의미 영역(semantic domain)과 결합한다는 것을 뜻한다. 이러한 점에서 이중 조건화는 변형 이후에도 어떤 개념이 쓰였는지를 확인할 가능성을 높인다.

[그림] 토큰트레이스의 워터마크 삽입과 추출 구조



출처: Li Zhang 외 4인, "TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery", arXiv, 2026.04.24., https://arxiv.org/pdf/2602.19019

• **질의 기반 모듈을 통한 개념 분리와 추적**

워터마크를 삽입하는 것만큼 중요한 것이 나중에 그것을 정확히 읽어 내는 일이다. 토큰트레이스는 이를 위해 질의 기반 모듈(query-based module)을 둔다. 이는 산출물과 함께 어떤 개념을 확인할지 지정하는 짧은 질의를 입력으로 받아, 해당 개념의 워터마크만 골라 읽어 내는 장치다. 확인하려는 개념을 질의문으로 입력하면, 모듈이 그 개념에 해당하는 워터마크를 되짚어 원래의 비밀키로 복원한다.

워터마크를 읽는 과정은 단계적으로 이루어진다. 먼저 워터마크가 삽입된 이미지와 질의문이 각각 고정된 이미지 인코더와 텍스트 인코더를 거쳐 특징으로 바뀐다. 이 두 특징은 주의 모듈(attention module)에서 하나로 합쳐진 뒤 개념 임베딩으로 예측되고, 이어 비밀 해독기가 이를 원래의 비밀 신호로 되돌려 확인을 마친다. 이때 두 인코더로는 널리 쓰이는 대조 언어·이미지 사전 학습(Contrastive Language-Image Pre-training, CLIP) 모델을 그대로 쓰되 일부 가벼운 층만 학습시켜 새로운 개념을 빠르게 추가할 수 있다.

이 질의 기반 방식이 다중 개념 문제를 푸는 열쇠다. 확인하려는 개념을 질의문으로 지정하면 모듈이 그 개념의 워터마크만 골라 읽어내므로, 한 이미지에 객체와 화풍이 겹쳐 있어도 각각을 따로 떼어 추적할 수 있다. 실제로 이 방식은 화풍 추적에서 약 91.67%, 객체 추적에서 약 90.43%의 정확도를 보여,[1] 기존 기법을 모두 앞섰다. 이는 한 장의 산출물에 여러 개념이 섞여 있어도 각각이 어디서 비롯되었는지를 따로 가려낼 수 있음을 뜻한다.

• **변형 내성과 확장성이 갖는 보호**

추적 기술이 실효성을 가지려면 실제 환경에서도 변형을 견뎌야 한다. AI 산출물은 인터넷에 유통되는 과정에서 압축되거나 잘리고, 색이 바뀌거나 회전되는 등 다양한 편집을 거치기 때문이다. 토큰트레이스는 의미와 잠재 양쪽 영역에 워터마크를 삽입하는 구조 덕분에, 실제 환경에서도 워터마크가 비교적 안정적으로 남는다. 또한 다루는 개념의 수가 크게 늘어도 정확도가 완만하게만 떨어지고, 이미 학습한 모델을 다시 훈련하지 않고도 새 개념을 더할 수 있어 창작자가 늘어나는 현실에 맞춰 체계를 넓혀 가기에 적합하다.

이러한 특성들이 모이면 추적 기술은 현실에서 쓸 수 있는 수단이 된다. 변형에 강하고 개념별로 따로 읽어낼 수 있는 워터마크는, 화풍을 무단 복제한 산출물이 편집을 거쳐 퍼지더라도 그 출처를 되짚을 단서를 남긴다. 다만 이 기술은 산출 과정에 워터마크를 미리 삽입해야 작동하므로, 워터마크가 없는 상태로 이미 만들어진 산출물까지 거슬러 추적하지는 못한다. 따라서 이 기술은 출처를 사후에 단정하는 도구가 아니라, 어떤 개념이 쓰였는지를 확인할 단서를 제공하는 수단으로 이해하는 것이 적절하다.

1) Li Zhang 외 4인, “TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery”, arXiv, 2026.04.24. https://arxiv.org/pdf/2602.19019

III. 결론 및 전망: 기술적 보호의 방향과 과제

• **기술적 의의와 한계**

토큰트레이스가 보여주는 바는, 산출물의 출처 추적이 사후 비교가 아니라 산출 과정 자체에 내장된 기술로 옮겨 가고 있다는 점이다. 이 기술은 한 장의 이미지에 객체와 화풍이 겹쳐 있어도 각각이 어디서 비롯되었는지를 따로 읽어낼 단서를 남긴다. 그동안 짚어 내기 어려웠던 기여 관계를 기술적 토대 위에서 확인할 수 있게 된다는 뜻이다. 이러한 점에서 워터마크는 산출물을 사후에 확인하는 장치를 넘어, 어떤 원작이 어떤 결과물에 쓰였는지를 짚어 내는 출처 확인의 수단으로 자리를 옮겨 가고 있다.

다만 이 기술이 모든 과제를 대신하지는 못한다. 워터마크는 산출 과정에 미리 삽입되어야 작동하므로, 이미 워터마크 없이 만들어진 산출물에는 적용하기 어렵다. 따라서 기술의 실효성을 높이려면 산출 단계에서 워터마크 삽입을 이끌어내는 장치와, 읽어낸 출처 정보를 창작자에게 연결해 보상으로 잇는 체계가 함께 마련될 필요가 있다. 기술과 제도가 보조를 맞출 때, 산출물의 출처를 둘러싼 불확실성은 점차 줄어들 것으로 전망된다.

참고문헌

• John Collomosse, “TokenTrace at CVPR 2026: Tracing Creative Influence in Generative AI”, Adobe, 2026.06.02., https://research.adobe.com/news/tokentrace-at-cvpr-2026-tracing-creative-influence-in-generative-ai/
• Li Zhang 외 4인, “TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery”, arXiv, 2026.04.24., https://arxiv.org/pdf/2602.19019

기술용어

순번	용어	설명
1	다중 개념 귀속 (multi-concept attribution)	한 이미지에 객체와 화풍처럼 여러 개념이 섞여 있을 때, 각 개념이 어디서 비롯되었는지를 따로 가려내는 작업

AI 음악 라이선스 분쟁과 블랙박스 멤버십 추론 기술

뉴스 브리프

2026년 6월, 미국 및 캐나다 음악가 연맹은 유니버설뮤직그룹과 워너뮤직그룹이 우디오 및 수노 관련 저작권 분쟁 합의와 라이선스 과정에서 연주자가 참여한 녹음물의 AI 학습 이용을 허용했으나, 연주자에게 합의금과 향후 라이선스 수익을 배분하지 않았다고 주장하며 소송을 제기했다. 이는 AI 음악 라이선스가 음반사와 기술기업 간 권리 처리 문제를 넘어, 실제 녹음에 참여한 연주자의 보상 권리와 학습 데이터 이용 범위에 대한 검증 문제로 확대되고 있음을 보여준다. 본 보고서는 상용 AI 음악 생성 모델의 학습 데이터가 공개되지 않는 블랙박스 환경에서 특정 음원이 학습에 포함됐는지 확인하기 위한 블랙박스 멤버십 추론 기술을 살펴보고, 후보 음원과 생성 음원 간 정렬 비교, 그림자 모델 기반 판별 기준 구축, 오디오 인코더별 성능 차이를 중심으로 학습 데이터 감사의 기술적 가능성과 상용 모델 적용을 위한 검증 과제를 분석한다.

뉴스 플러스

I. 서론: AI 음악 라이선스 확대와 학습 데이터 검증 문제

• 미국 및 캐나다 음악가 연맹이 제기한 녹음물 이용 문제

2026년 6월, 미국 및 캐나다 음악가 연맹(American Federation of Musicians of the United States and Canada, AFM)은 뉴욕 남부 연방지방법원에 유니버설뮤직그룹(Universal Music Group, UMG)과 워너뮤직그룹(Warner Music Group, WMG)을 상대로 소송을 제기했다.[2] 미국 및 캐나다 음악가 연맹은 두 음반사가 우디오(Udio) 및 수노(Suno) 관련 저작권 분쟁 합의와 라이선스 계약을

2) Blake Brittain, “Musicians union sues record labels over AI licensing”, Reuters, 2026.06.06., https://www.reuters.com/legal/litigation/musicians-union-sues-record-labels-over-ai-licensing-2026-06-05/

추진하는 과정에서, 연주자가 참여한 녹음물이 AI 학습 및 생성 모델 개발에 활용될 수 있도록 허용했다고 주장했다. 미국 및 캐나다 음악가 연맹은 연주자에게 합의금과 향후 라이선스 수익을 배분하지 않았다는 점을 문제 삼고 있다.

미국 및 캐나다 음악가 연맹은 음반사들이 체결한 음향 녹음 노동 협약(Sound Recording Labor Agreement, 이하 SRLA) 제21조에 따라, 녹음물의 새로운 형태의 이용이 발생할 경우 연주자에게 보상이 제공돼야 한다고 주장한다.[3] 소장에 따르면 음반사들이 앞서 수노와 우디오를 상대로 제기했던 저작권 침해 소송에서 AI 모델의 음악 학습 문제를 지적했던 점을 언급한다. 미국 및 캐나다 음악가 연맹은 음반사들이 과거 문제 삼았던 것과 동일한 유형의 AI 학습을 합의 및 라이선스 과정에서 허용한 것이라고 주장했다.[4]

이 사례는 AI 음악 라이선스가 음반사와 기술기업 간 권리 처리 문제에 그치지 않음을 보여준다. 실제 녹음에 참여한 연주자의 보상 권리와 수익 배분 문제도 함께 논의될 수 있다. 이에 따라 AI 음악 모델이 어떤 데이터를 학습했는지 확인하는 기술적 검증 수단의 필요성이 커지고 있다.

• 블랙박스 AI 음악 모델에서 학습 데이터 검증이 필요한 이유

수노와 우디오 같은 상용 AI 음악 시스템은 고품질의 수 분 길이의 음악 생성 능력을 보여주고 있으나, 내부 구조와 학습 데이터 구성은 공개되지 않은 경우가 많다. 생성형 음악 모델의 학습 데이터 구축 과정이 불투명할수록, 특정 음악이 모델 학습에 포함됐는지 외부에서 확인하기 어렵다.

최근 음악 생성 모델에 특화된 블랙박스 멤버십 추론(Black-Box Membership Inference) 기술이 제안되었다. 이 기술은 모델 내부 구조나 학습 데이터에 접근하지 않고, 입·출력만으로 후보 음원이 학습에 사용됐는지 추정한다. 핵심은 학습 데이터에 포함된 음원일수록 후보 음원과 해당 음원의 설명문을 바탕으로 생성된 결과물이 더 강한 의미적·구조적 정렬을 보일 수 있다는 점이다.

즉, 특정 음원이 학습 데이터에 포함된 경우 모델이 해당 음원의 음악적 특징을 더 잘 반영할 가능성이 있다. 블랙박스 멤버십 추론 기술은 후보 음원과 생성 결과물 사이의 정렬 차이를 측정해 학습 포함 가능성을 판단하는 방식이다.

3) Dylan Smith, “AFM Sues Universal Music, Warner Music for Failing ‘To Share in the Settlement Proceeds and Future Revenue’ Under Their Suno and Udio Deals”, Digital Music News, 2026.06.05., https://www.digitalmusicnews.com/2026/06/05/afm-universal-music-lawsuit/
4) Blake Brittain, “Musicians union sues record labels over AI licensing”, Reuters, 2026.06.06., https://www.reuters.com/legal/litigation/musicians-union-sues-record-labels-over-ai-licensing-2026-06-05/

II. 본론: 블랙박스 멤버십 추론 기술로 학습 음원을 판별하는 방식

• 후보 음원의 설명문을 활용한 블랙박스 질의 방식

음악 생성 모델에 블랙박스 멤버십 추론을 적용할 경우, 먼저 검사 대상이 되는 후보 음원의 음악적 특징을 설명하는 캡션(Caption)을 추출한다.

연구에서는 캡션 추출을 위해 음원을 입력받아 해당 음원의 음악적 특징을 자연어로 설명하는 오디오-언어 모델인 뮤직 플라밍고(Music Flamingo)를 사용했다. 이렇게 추출된 캡션을 대상 음악 생성 모델에 입력하면, 모델은 캡션에 맞는 새로운 음원을 생성한다.

이후 후보 음원과 생성 음원을 비교해 두 음원 사이의 음악적 정렬(semantic and structural alignment)* 정도를 측정한다. 학습에 사용된 음원은 비학습 음원보다 생성 결과와 더 강한 정렬을 보일 가능성이 있으며, 이러한 통계적 차이가 학습 포함 여부를 판단하는 주요 단서로 활용된다.

* 음악적 정렬(semantic and structural alignment): 후보 음원과 생성 음원이 장르, 분위기, 악기 구성, 멜로디·리듬 구조 등에서 얼마나 유사하게 맞물리는지를 의미

• 그림자 모델을 활용한 학습·비학습 음원 비교

후보 음원과 생성 음원의 정렬 차이가 학습 여부와 어떤 관계를 갖는지 판단하려면, 먼저 학습 음원과 비학습 음원의 비교 기준을 만들어야 한다. 이를 위해 그림자 모델(shadow model)*이 활용됐다.

연구에서는 오디오엘디엠2(AudioLDM2), 스테이블 오디오 오픈(Stable Audio Open), 머스탱고(Mustango)를 실험 대상으로 활용했다. 세 모델은 학습 데이터셋 일부가 공개적으로 문서화되어 있어 학습 음원과 비학습 음원을 구분할 수 있다. 실험은 세 모델 중 하나를 타깃 모델로 두고, 나머지 두 모델을 그림자 모델로 사용하는 방식으로 반복됐다.

학습 음원은 실제 학습 데이터에 포함된 원본 음원으로 구성했다. 대조 대상인 비학습 음원은 보조 음악 생성 모델인 잼(Jam)을 이용해 만들었다. 학습 음원의 캡션을 잼에 입력해 의미적으로 유사하지만 학습 데이터에는 포함되지 않은 음원을 생성한 것이다.

이 과정을 통해 학습에 사용된 음원과 그에 대응하는 생성 음원, 학습에 사용되지 않은 음원과 그에 대응하는 생성 음원으로 구성된 두 종류의 데이터 쌍이 구축됐다. 이후 이 비교쌍을 이용해 음악 감사 모델(music auditor)**을 훈련했다. 음악 감사 모델은 두 음원 사이의 정렬 패턴을 학습하고, 새로운 후보 음원이 학습 데이터에 포함됐을 가능성을 점수로 산출한다.

* 그림자 모델(shadow model): 대상 모델을 직접 확인할 수 없을 때, 유사한 방식으로 작동하는 다른 모델을 이용해 판별 기준을 학습하는 보조 모델
** 음악 감사 모델(music auditor): 후보 음원과 생성 음원 사이의 정렬 패턴을 분석해 학습 포함 여부를 판별하는 모델

• **오디오 인코더를 활용한 음향 특징 비교**

음악 감사 모델은 후보 음원과 생성 음원 사이의 유사성을 직접 듣고 판단하지 않는다. 대신 오디오 인코더(audio encoder)*를 사용해 음원을 숫자 형태의 특징값으로 바꾼 뒤 비교한다.

연구에서는 여러 오디오 인코더를 비교해 어떤 방식이 학습 여부 판별에 효과적인지 확인했다. 후보 음원과 생성 음원을 각각 오디오 인코더에 입력하고, 두 음원의 특징값을 결합해 음악 감사 모델에 넣었다. 음악 감사 모델은 이 정보를 바탕으로 후보 음원이 학습 데이터에 포함됐을 가능성을 점수로 산출한다.

실험 결과, 오디오 인코더의 선택은 판별 성능에 큰 영향을 미쳤다. 특히 DAC 인코더를 사용했을 때 평균 정확도는 97.7%로 나타났다. 이는 후보 음원과 생성 음원 사이의 세부 음향 패턴이 학습 데이터 포함 여부를 판단하는 데 중요한 단서가 될 수 있음을 보여준다. 특히 저수준 음향 특징을 잘 포착하는 인코더가 학습 여부 판별에 유리하게 작용한 것으로 해석된다.

* 오디오 인코더(audio encoder): 음원을 AI가 비교할 수 있는 숫자 형태의 특징값으로 변환하는 모델 또는 구성요소

• **오디오 인코더별 판별 성능과 모델 간 차이**

연구에서는 MERT, Music2Vec, DAC, Fx-Encoder++, CLAP 등 5종의 오디오 인코더를 비교했다. 이 중 DAC는 세 모델 전반에서 가장 안정적인 성능을 기록했다. DAC 인코더의 정확도는 오디오엘디엠2 대상 97.5%, 머스탱고 대상 97.1%, 스테이블 오디오 오픈 대상 98.6%로 나타났다.

반면 Music2Vec는 평균 정확도 78.3%로 상대적으로 낮은 성능을 보였다. Fx-Encoder++는 스테이블 오디오 오픈에서는 94.5%의 정확도를 기록했으나, 오디오엘디엠2에서는 69.4%에 그쳤다. 이는 인코더가 포착하는 특징의 종류에 따라 학습 여부 판별 성능이 달라질 수 있음을 보여준다.

생성 모델에 따른 성능 차이도 확인됐다. 스테이블 오디오 오픈은 인코더 평균 기준 가장 높은 판별 성능을 보였으나, 머스탱고와 오디오엘디엠2에서는 일부 인코더의 성능이 상대적으로 낮게 나타났다. 이는 출력 길이, 데이터 분포, 생성 모델별 제약 차이가 판별 성능에 영향을 줄 수 있음을 시사한다.

III. 결론: 블랙박스 학습 데이터 감사의 가능성과 과제

• **블랙박스 환경에서의 학습 데이터 감사 가능성**

블랙박스 멤버십 추론은 음악 생성 모델의 내부 구조나 학습 데이터에 접근하지 않고도 특정 음원의 학습 포함 가능성을 점검할 수 있는 기술이다. 감사 과정은 사후적으로 이루어지며, 대상 모델의 훈련이나 작동에 직접 영향을 주지 않는다. 이번 연구는 학습 데이터 감사가 블랙박스 AI 음악 모델 환경에서도 활용될 가능성을 보여준다. 특히 DAC와 같은 오디오 인코더를 활용할 경우 높은 판별 정확도를 보였다는 점에서, 음악 생성 모델의 데이터 출처 검증에 참고할 수 있는 기술적 접근으로 평가된다.

미국 및 캐나다 음악가 연맹의 소송에서 드러난 것처럼 AI 음악 라이선스가 확대될수록 연주자 보상과 데이터 이용 범위에 대한 논의도 커질 수 있다. 이때 블랙박스 학습 데이터 감사 기술은 특정 음원이 학습에 사용됐는지 확인하는 보조적 검증 수단으로 활용될 수 있다.

• 상용 모델 적용을 위한 검증 과제

다만 이번 연구는 공개 음악 생성 모델을 대상으로 수행됐다. 수노나 우디오와 같은 실제 상용 폐쇄형 모델에 대해서는 직접 검증이 이루어지지 않았다. 따라서 연구 결과를 상용 모델에 그대로 적용하기에는 한계가 있다. 또한 현재 평가는 데이터 규모와 다양성이 제한적이다. 특정 캡션 추출 방식과 보조 생성 모델에 의존하고 있어 분포 편향이 발생할 가능성도 있다. 캡션 작성 방식이 달라지거나 보조 생성 모델이 바뀌면 후보 음원과 생성 음원의 비교 결과도 달라질 수 있다. 상용 모델에 적용하기 위해서는 다양한 장르, 길이, 음질, 제작 방식의 음원을 활용한 추가 검증이 필요하다. 또한 모델 업데이트, 출력 제한, 프롬프트 응답 방식 등 실제 서비스 환경의 변수도 고려해야 한다.

참고문헌

- Blake Brittain, “Musicians union sues record labels over AI licensing”, Reuters, 2026.06.06., https://www.reuters.com/legal/litigation/musicians-union-sues-record-labels-over-ai-licensing-2026-06-05/
- Dylan Smith, “AFM Sues Universal Music, Warner Music for Failing ‘To Share in the Settlement Proceeds and Future Revenue’ Under Their Suno and Udio Deals”, Digital Music News, 2026.06.05., https://www.digitalmusicnews.com/2026/06/05/afm-universal-music-lawsuit/
- Yi Chen Liu 외 5명, “Auditing Training Data in Generative Music Models via Black-Box Membership Inference”, arXiv, 2026.05.28., https://arxiv.org/abs/2605.29202

기술용어

순번	용어	설명
1	음악적 정렬 (semantic and structural alignment)	후보 음원과 생성 음원이 장르, 분위기, 악기 구성, 멜로디·리듬 구조 등에서 얼마나 유사하게 맞물리는지를 의미
2	그림자 모델 (shadow model)	대상 모델을 직접 확인할 수 없을 때, 유사한 방식으로 작동하는 다른 모델을 이용해 판별 기준을 학습하는 보조 모델
3	음악 감사 모델 (music auditor)	후보 음원과 생성 음원 사이의 정렬 패턴을 분석해 학습 포함 여부를 판별하는 모델
4	오디오 인코더 (audio encoder)	음원을 AI가 비교할 수 있는 숫자 형태의 특징값으로 변환하는 모델 또는 구성요소

진위 판별을 넘어선 합성 영상 출처 식별 기술의 발전

뉴스 브리프

영상을 만들어 내는 AI 기술이 빠르게 발전하면서, 실제 촬영물과 구분하기 어려운 합성 영상이 온라인에 늘고 있다. 유튜브는 합성 영상에 라벨을 자동으로 붙이는 체계로 전환하고, 특정 인물의 얼굴과 외형이 무단 복제된 영상을 찾아 신고하는 도구를 일반에 개방했다. 그동안의 대응은 영상의 진위를 가리는 데 머물렀다. 그러나 합성 모델이 수없이 늘어나면서, 이제 핵심 질문은 영상이 진짜인가에서 어느 모델이 만들어 냈는가로 바뀌었다. 출처를 식별하는 기술은 합성 영상이 남긴 미세한 시간적 흔적을 읽어 모델별 고유한 지문을 가려낸다. 이는 초상과 저작물의 무단 활용을 기술적으로 추적하는 토대가 된다. 본 보고서는 이 기술의 작동 원리와 산업적 의미를 살피고, 영상 생태계의 책임성을 높이기 위한 과제를 함께 짚는다.

뉴스 플러스

I. 서론: 진위 판별을 넘어선 영상 출처 식별의 부상

• 자동 라벨링 전환과 초상 보호 도구 개방이 드러낸 문제

유튜브(YouTube)는 지난 5월 합성 영상 라벨링 방식을 변경했다. 그동안은 게시자가 스스로 합성 여부를 밝히는 자발적 방식에 의존했으나, 이제는 합성 영상을 자동 감지해 표시하는 체계로 전환한다. 자발적 공개는 게시자의 라벨링 단순 누락, 규정 숙지 미흡, 조회수 확보를 위한 의도적인 회피 사례가 잦아 신뢰하기 어려웠다. 해당 전환은 합성 영상을 자체 도구로 만든 경우나 C2PA (Coalition for Content Provenance and Authenticity)* 표준의 메타데이터가 붙은 경우에 표시를 영구적으로 유지한다는 점에서, 표시의 근거를 기술적 신호에 두려는 시도로 볼 수 있다.

동시에 유튜브는 인물 초상 보호 도구도 일반에 개방했다. 연예인, 운동선수, 공직자처럼 외형이 도용될 위험이 큰 사람은 채널 보유 여부와 무관하게 자신의 초상을 등록하고, 이를 복제한 영상을 찾아 삭제를 요청할 수 있다. 이 도구는 업로드된 저작물을 식별하는 기존 시스템인 콘텐츠 아이디(Content ID)의 발상을 인물의 얼굴과 외형으로 넓힌 것이다. 초상이 무단으로 도용되어도 당사자가 이를 우연히 발견하기 전에는 대응할 길이 없던 상황에서, 이 도구는 권리자에게 침해를 확인하고 대응할 수단을 마련해 준다는 점에서 의미가 있다.

* C2PA(Coalition for Content Provenance and Authenticity): 디지털 콘텐츠의 출처와 진본성을 인증하는 국제 기술 표준을 개발하는 연합체로, 어도비(Adobe)·마이크로소프트(Microsoft)·인텔(Intel) 등이 참여하고 있음

• 진위 판별을 넘어선 출처 식별 기술의 등장 배경과 중요성

합성 영상이 폭증하는 환경에서 라벨링과 초상 보호가 실효를 가지려면, 영상의 진위를 넘어 그 출처까지 식별하는 기술이 뒷받침되어야 한다. 그러나 지금까지의 대응은 대부분 영상이 진짜인지 가짜인지만 가르는 이진 판별에 그쳤다. 가짜라고 판정하더라도 그것이 어느 모델에서 비롯되었는지를 모르면, 무단 활용을 추적하고 권리를 보호하는 후속 대응으로 나아가기 어렵다.

캘리포니아대학교 리버사이드(University of California, Riverside)와 유튜브 연구진이 내놓은 SAGA(Source Attribution of Generative AI videos)는 진위 여부에서부터 정확한 모델명까지 5단계의 세분화된 식별 단계로 짚어 낸다. 이 기술은 합성 영상에 남은 미세한 시간적 흔적을 읽어 모델마다 다른 고유한 지문을 분간한다. 기술을 통해 콘텐츠의 출처를 식별할 수 있다면 무단으로 이용된 저작물을 라벨링하는 작업이 더 이상 권리자의 신고나 우연한 발견에만 의존하지 않아도 되며, 이러한 점에서 출처 식별 기술은 권리 보호 체계를 구축하는 토대가 된다.

II. 본론: 합성 영상 출처 식별 기술의 구조와 원리

• 이진 판별의 한계와 다단계 출처 식별의 구조

기존 합성 영상 대응은 대부분 영상이 진짜인지 가짜인지만 가르는 이진 판별에 머물렀다. 합성 모델이 몇 개에 불과하던 시절에는 이 방식으로도 충분했으나, 모델이 수없이 늘고 빠르게 진화하는 지금은 가짜라는 사실 하나만으로는 대응의 실마리를 얻기 어렵다. 진짜와 가짜라는 두 갈래로만 나누는 분류는 같은 가짜 안에 뒤섞인 수많은 모델을 구별하지 못하기 때문이다.

SAGA는 이 한계를 넘어, 특정 합성 영상이 구체적으로 어느 모델에서 비롯되었는지까지 여러 단계로 짚는다. SAGA는 이를 다섯 단계로 세분화하는데, 진위 여부, 생성 방식, 모델 계열, 개발팀, 그리고 정확한 모델명이다. 이렇게 단계를 나누면 가장 세밀한 모델명 식별에서 확신이 낮더라도, 개발팀이나 모델 계열 같은 더 넓은 범주의 단계에서는 단서를 건질 수 있다.

이 다단계 구조는 권리 보호의 관점에서 의미가 있다. 무단 활용이 의심되는 영상이 어디서 나왔는지를 단계별로 좁혀 가면, 침해의 경로를 기술적으로 추적할 길이 열린다. SAGA는 학습에 쓰이지 않은 모델에도 어느 정도 대응하는 강건성을 보이는데, 이는 새로운 합성 모델이 계속 등장하는 환경에서 의미가 있다. 다만 출처 식별이 곧바로 법적 침해 판단으로 이어지는 것은 아니며, 어디까지나 그 판단에 필요한 기술적 단서를 제공하는 단계로 보는 편이 정확하다.

• 영상의 시간적 흔적을 읽어내는 변환기의 작동 원리

SAGA의 출처 식별은 영상에 남은 시간적 흔적을 분석하는 변환기(Transformer)를 통해 이루어진다. 변환기는 입력된 데이터의 각 부분이 서로 어디에 주목해야 하는지를 따지며 관계를 학습하는 신경망 구조를 뜻한다. 합성 영상에는 사람 눈에 잘 띄지 않는 흔적이 남는데, 프레임이 이어지는 방식이나 움직임의 어색함이 그 예다. SAGA는 이를 단서로 삼아 어느 모델에서 비롯되었는지를 식별한다.

SAGA의 작동은 크게 두 단계로 나뉜다. 먼저 공간 인코더(Spatial Encoder)가 개별 프레임 내 공간적 연관성과 특징을 분석한다. 다음으로 시간 인코더(Temporal Encoder)가 여러 프레임을 시간 순서대로 늘어놓고, 프레임과 프레임 사이에 나타나는 변화의 양상을 살핀다. 이때 각 프레임에 순서 정보를 더해, 어느 장면이 먼저이고 나중인지를 모델이 인식하도록 한다. 이렇게 공간과 시간을 차례로 분석하면, 한 장면 안의 특징과 장면이 흐르며 드러나는 특징을 함께 포착할 수 있다.

이 작동 원리는 합성 영상 식별에서 의미가 있다. 정지된 이미지만 보는 방식은 영상 특유의 시간적 흔적을 놓치지만, 시간 축을 따라 프레임의 연결을 살피면 모델마다 다른 고유한 양상을 잡아낼 수 있기 때문이다. 무단 활용이 의심되는 영상에서 이러한 흔적을 읽어내면, 그것이 실제 촬영물인지 특정 모델의 산출물인지를 가려내는 기술적 근거가 된다.

• 소량의 데이터로 성능을 끌어올리는 데이터 효율적 학습 전략

SAGA는 적은 양의 라벨 데이터(label data)만으로 높은 식별 성능을 달성하는 2단계 학습 전략을 쓴다. 라벨 데이터란 어느 영상이 어느 모델에서 비롯되었는지를 사람이 미리 표시해 둔 학습 자료를 뜻한다. 출처를 단계별로 표시한 자료는 구하기 어렵고 절대적인 데이터량도 적은 반면, 진짜와 가짜만 구분한 자료는 비교적 풍부하다. SAGA는 이 차이를 활용해 두 단계로 나누어 학습한다.

1단계에서는 풍부한 진위 구분 자료로 진짜와 가짜를 가르는 기본기를 먼저 익힌다. 2단계에서는 이렇게 다진 토대 위에, 출처를 표시한 소량의 자료만으로 모델별 식별 능력을 더한다. 이 과정에서 서로 닮아 구분하기 어려운 모델들을 떼어 놓기 위해, 가장 헷갈리는 사례를 골라 집중적으로 학습하는 기법을 더한다. 그 결과 전체 자료의 0.5%만 써도 자료 전량을 학습한 경우에 맞먹는 성능에 이른다.

이 데이터 효율성은 산업적 관점에서 의미가 있다. 새로운 합성 모델이 나올 때마다 방대한 학습 자료를 모으는 일은 비용과 시간이 크게 드는데, 소량의 자료로 식별 능력을 갖출 수 있다면 플랫폼이 변화에 빠르게 대응할 여지가 생긴다. 무단 활용이 의심되는 영상의 출처를 가려내는 체계를 현실적인 비용으로 운영할 수 있다는 점에서, 데이터 효율성은 출처 식별이 권리 보호 수단으로 확산되는 데 필요한 조건이 된다.

- **시간적 어텐션 서명과 출처 식별의 가능성과 한계**

SAGA는 모델별 고유한 흔적을 시간적 어텐션 서명(Temporal Attention Signature, T-Sig)으로 시각화한다. 이는 영상의 프레임과 프레임 사이에서 모델이 어디에 주목하는지를 모아 하나의 패턴처럼 나타낸 것을 뜻한다. 같은 모델에서 비롯된 영상은 내용이 서로 달라도 일관된 패턴을 보이고, 다른 모델은 서로 구별되는 패턴을 보인다. 이를 통해 SAGA가 단지 결과만 내놓는 것이 아니라, 무엇을 근거로 출처를 가려냈는지를 사람이 들여다볼 수 있게 된다.

이 해석 가능성은 권리 보호의 관점에서 의미가 있다. 무단 활용이 의심되는 영상의 출처를 가려낼 때, 판단의 근거가 패턴 형태로 제시되면 그 결과를 검토하고 신뢰할 여지가 커진다. 학습에 쓰이지 않은 모델도 고유한 패턴을 남긴다는 점에서, 새로 등장하는 합성 모델에도 일정한 대응이 가능하다. 이는 출처 식별이 특정 모델 목록에 갇히지 않고 변화하는 환경에 적용될 수 있음을 뜻한다.

다만 출처 식별에는 분명한 한계가 있다. 가장 세밀한 모델명 단계에서는 서로 닮은 모델을 구분하기 어려워 정확도가 낮아지는 경우가 있으며, 일부 모델에서는 식별 정확도가 크게 떨어지는 사례도 확인된다. 탐지 자체가 완벽하지 않아 실제 촬영물을 합성물로 잘못 가려낼 가능성도 남는다. 이러한 점에서 출처 식별의 결과는 그 자체로 침해를 단정하는 근거가 아니라, 사람의 검토와 후속 판단을 보조하는 기술적 단서로 다루는 편이 정확하다.

III. 결론 및 전망: 출처 식별 기술의 전망과 과제

• 출처 식별 기술이 여는 책임 있는 영상 생태계

최근의 식별 기술은 진짜와 가짜를 가르는 데 머물던 기존 방식과 달리, 출처를 여러 단계로 짚고 그 근거를 패턴 형태로 확인할 수 있다는 점에서 차이가 있다. 이를 통해 무단으로 이용된 저작물과 초상을 추적하는 작업이 권리자의 신고나 우연한 발견에만 기대지 않아도 된다. 따라서 출처 식별은 라벨링과 초상 보호 같은 플랫폼의 대응을 받치는 기술적 토대가 되며, 합성 영상이 폭증하는 환경에서 책임의 소재를 좁히는 출발점이 된다.

다만 이러한 가능성이 곧 완전한 해법을 뜻하지는 않는다. 출처 식별의 정확도는 모델에 따라 고르지 않고, 실제 촬영물을 합성물로 잘못 판단할 여지도 남아 있어 기술적으로 보완할 과제가 적지 않다. 이러한 점에서 출처 식별의 결과는 침해를 단정하는 근거가 아니라 검토를 돕는 단서로 다루는 제도적 신중함이 함께 요구된다. 기술의 발전과 더불어 그 결과를 권리 보호 절차에 어떻게 연결할지에 관한 사회적 합의가 마련될 때, 책임 있는 영상 생태계가 자리 잡을 수 있을 것으로 전망된다.

참고문헌

- Alina Maria Stan "YouTube will now automatically label AI-generated videos, whether creators disclose them or not", TNW, 2026.05.27., https://thenextweb.com/news/youtube-will-now-automatically-label-ai-generated-videos-whether-creators-disclose-them-or-not
- Alex Weprin, "YouTube Opens Up AI Deepfake Detection Tool to All of Hollywood (Exclusive)", The Hollywood Reporter, 2026.04.21., https://www.hollywoodreporter.com/business/digital/youtube-ai-deepfake-detection-tool-1236569593/
- Rohit Kundu 외 5명, "SAGA: Source Attribution of Generative AI Videos", 2026.04.02., https://arxiv.org/pdf/2511.12834

기술용어

순번	용어	설명
1	C2PA (Coalition for Content Provenance and Authenticity)	디지털 콘텐츠의 출처와 진본성을 인증하는 국제 기술 표준을 개발하는 연합체로, 어도비(Adobe)·마이크로소프트(Microsoft)·인텔(Intel) 등이 참여하고 있음