저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 구글, 논문 작성 자동화하는 다중 AI 에이전트 페이퍼오케스트라 공개

▶ 구글이 연구 자료를 취합해 학술 논문으로 자동 변환하는 다중 에이전트 시스템인 페이퍼오케스트라를 공개했다. 이 시스템은 논문 구조 설계, 그림 생성, 문헌 조사, 본문 작성, 품질 정제 등 논문 작성 단계를 전문 에이전트별로 분배해 독립적으로 수행한다. 특히 문헌 조사 에이전트는 참조 논문의 존재 여부를 검증하여 허위 인용을 방지하고, 품질 정제 에이전트는 인간의 동료 심사 체계처럼 초안을 반복 검토한다. 평가 결과, 기존 자동화 시스템 대비 전체 완성도가 개선되었으며, 학회 심사 시뮬레이션에서도 인간 논문에 근접한 성능을 보였다. 다만 연구진은 이를 연구자를 대체하는 도구가 아닌 생산성 향상 보조 수단으로 규정하며, 학계 차원의 윤리 지침과 품질 검증 체계 마련이 필요하다고 강조했다.

산업 라이브 스트리밍 불법 유통 탐지를 위한 워터마킹 기반 플랫폼 출시

▶ 라이브 스트리밍 콘텐츠의 불법 유통이 확산되면서, 불법 스트리밍을 신고받은 후 차단하는 기존 방식에 한계가 있다는 지적이 이어져 왔다. 불법 스트리밍이 이미 확산된 뒤에야 조치가 이루어지는 구조여서, 실시간 대응이 어렵다는 문제가 있었다. 이러한 상황에서 아일랜드의 디지털 솔루션 기업 스테가웨이브는 불법 스트리밍을 실시간으로 탐지하고 차단하는 워터마킹 기반 플랫폼을 출시하였다. 스테가웨이브 플랫폼은 구독자별로 고유한 토큰을 발급하고, 해당 토큰을 송출되는 각 스트리밍을 구분해 서로 다른 워터마크를 삽입하는 방식으로 작동한다. 불법 스트리밍이 감지되면, AI가 해당 스트리밍에 삽입되어 있는 워터마크를 분석해 유출 계정을 특정하고, 해당 구독자의 접속을 즉시 차단한다.

산업 AI 기반 영상 편집 기술과 동일성 판단 과제

▶ 영상 후반 작업에서 불필요한 객체를 삭제하는 인페인팅 기술은 오랜 기간 활용됐으나, 삭제된 객체가 주변에 미치던 물리적 영향까지는 반영하지 못하는 한계가 있었다. 2026년 4월 넷플릭스 연구팀이 공개한 AI 기반 영상 편집 기술 'VOID'는 객체 삭제 이후 나머지 장면의 물리적 인과관계까지 재구성하는 방식으로, 기존 도구와의 비교 선호도 조사에서 64.8%를 기록하며 기술적 차별성을 보였다. 해당 모델이 누구나 활용할 수 있는 공유 플랫폼에 공개된 점과, 같은 시기 에이비드가 구글의 AI 모델을 자사 편집 소프트웨어에 통합한 사례는 영상 편집 작업에 AI 기능이 확산되는 흐름을 함께 보여준다. 이러한 기술이 확산될 경우, 촬영된 영상의 사후 변형 범위가 넓어지면서 원본과 편집본 간 경계가 모호해질 가능성이 있다. 또한, 동일성 판단 기준 등 기존 저작권 체계에 대한 검토와 함께 AI 산출물의 창작 기여도 및 저작권 귀속에 관한 기준 정립이 향후 과제로 남아있다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 구글, 개인 사진첩·앱 데이터 활용한 AI 이미지 생성 기능 공개

▶ 구글은 2026년 4월 16일, 맞춤형 응답 기능인 퍼스널 인텔리전스와 이미지 생성 AI인 나노 바나나 2를 결합해, 구글 포토 사진첩의 사진을 불러와 이미지 생성에 반영하는 기능을 공개하였다. 이용자는 긴 프롬프트를 작성하거나 참고 이미지를 직접 업로드하지 않아도, 짧은 지시만으로 자신의 취향과 일상은 물론 가족, 반려동물까지 반영된 결과물을 얻을 수 있다. 구글은 비공개 사진첩 자체를 모델 학습에 직접 활용하지 않으며, 사전 동의와 출처 확인 절차를 통해 이용자의 통제권을 보장하고 있다. 다만 제3자의 초상과 일상 정보까지 AI 산출물에 포함될 가능성이 커지면서, 관련 권리 관리 범위를 더욱 정교하게 재설정할 필요성이 커지고 있다.

산업 깃허브, 애니메이션 불법 스트리밍 저장소 900여 개 삭제

▶ 2026년 3월, 삭제 통보 대행업체 리무브유어미디어는 애니메이션 저작권자의 의뢰를 받아, 코드 공유 플랫폼 깃허브에 불법 스트리밍 영상을 자동으로 수집하는 도구가 담긴 저장소들을 삭제할 것을 요청했다. 리무브유어미디어는 이 도구들이 기술적 보호조치를 직접 해제하지는 않지만 접근 통제를 간접적으로 우회하는 것과 같다고 주장하였다. 깃허브는 이 주장을 받아들이지 않았으나 해당 도구들이 악성 시장 목록 등재 사이트를 대상으로 한다는 점에서 저작권 침해를 인정하여 900여 개의 저장소를 삭제하였다. 다만 깃허브가 기각 이유를 공개하지 않아, 간접 우회 방식이 저작권 보호조치 우회에 해당하는지에 대한 해석 기준은 여전히 확립되지 않은 상태이다.

기술 주간 기술 동향

▶ 2026년 4월 앤트로픽의 AI 코딩 도구 소스코드가 유출되면서 AI 생성 코드의 저작권 보호 필요성이 부각되고 있다. 기존 코드 워터마킹 기술은 코드를 수정하면 워터마크가 쉽게 제거되고 성능이 저하되는 문제가 있었다. 이를 해결하기 위해 제안된 MATRIX는 코드의 구조를 바꾸는 방식으로 워터마크를 숨기고, 같은 정보를 여러 곳에 중복 삽입하여 일부가 손상되어도 복원할 수 있도록 설계되었다. MATRIX는 12가지 공격 시나리오에서 95% 이상의 검출률을 기록했으며, 워터마크 삽입 시간은 기존 기술 대비 약 80배 빠른 속도를 기록했다. 이 기술은 AI 코드의 저작권 분쟁 해결과 출처 추적을 가능하게 하며, 향후 신뢰할 수 있는 코드 생태계 구축에 기여할 것으로 기대된다.

구글, 논문 작성 자동화하는 다중 AI 에이전트 페이퍼오케스트라 공개

학술 논문 작성 자동화의 필요성과 기존 기술의 한계

- **연구 성과 문서화 단계의 구조적 병목 현상**

- 학술 연구에서 논문 작성은 문헌 조사, 시각화 생성, 형식 정합성 확보 등 반복 작업으로 인해 상당한 시간을 소모하며, 연구자가 핵심 지적 활동에 집중하기 어렵게 만드는 주요 제약으로 작용함
- 산재된 실험 기록과 연구 노트를 일관된 흐름으로 재구성하고 관련 선행 연구를 배치하며 투고 규정에 맞는 형식을 갖추는 과정은 지적 창의성보다 절차적 숙련도에 의존하는 경향이 강함
- 구글 클라우드 AI팀은 2026년 4월 구조화되지 않은 연구 자료를 투고 가능한 논문으로 자동으로 변환하는 다중 에이전트 프레임워크 페이퍼오케스트라(PaperOrchestra)를 공개함
- 이 시스템은 연구 아이디어와 실험 데이터만으로 평균 40분 내에 학회 제출이 가능한 수준의 완성도 높은 원고를 생성할 수 있다고 밝힘

- **기존 AI 논문 작성 도구의 제한적 기능과 의존성 문제**

- 기존 AI 기반 논문 작성 도구들은 특정 실험 환경에 종속되거나 단편적인 문장 생성 및 단순 문헌 정리 수준에 그치는 한계를 보였으며, 논문 작성의 전체 과정을 포괄하지 못하는 제약이 있었음
- 일반적인 LLM 기반 도구는 텍스트 생성에는 강점을 보이지만, 실험 결과를 시각화하거나 학술 데이터베이스 탐색을 통한 참고문헌 수집 등 다층적 작업을 통합 처리하지 못하는 문제가 있었음
- 페이퍼오케스트라는 이러한 한계를 극복하기 위해 논문 작성의 전 과정을 주요 단계로 나누고, 각 단계를 전문화된 에이전트가 분담하는 방식을 채택함

다중 에이전트 기반 논문 생성 시스템의 구조와 작동 원리

- **5단계 에이전트 파이프라인을 통한 단계별 작업 분리**

- 페이퍼오케스트라는 논문 작성을 논문 구조 설계·시각화·문헌 조사·본문 작성·품질 정제 등 5개 전문 에이전트로 분리하여 각 에이전트가 특정 지적 작업을 독립적으로 수행하도록 설계됨
- 논문 구조 설계 에이전트는 연구 아이디어와 실험 데이터를 분석하여 논문 전체 구조를 설계하고, 필요한 그림과 문헌의 층위, 각 섹션의 서술 내용을 먼저 정리하는 역할을 담당함
- 이러한 단계별 분업 구조는 논문을 한 번에 완성하려는 접근 대신 필요한 판단을 순차 분리하여 각 단계의 품질을 독립적으로 최적화할 수 있다는 설계 철학을 반영함

• 문헌 검증 및 시각화의 병렬 처리 구조

- 시각화 에이전트는 단순 그래프뿐 아니라 개념도까지 생성할 수 있도록 설계되었으며, 구글이 개발한 AI 이미지 생성 모델인 나노 바나나 2(Nano Banana 2) 알고리즘을 활용하여 논문에 필요한 시각화 자료를 자동으로 생성하는 기술이 탑재됨
- 문헌 조사 에이전트는 학술 논문 검색 인터페이스인 시맨틱 스칼라(Semantic Scholar) API를 기반으로 관련 논문을 탐색하되, 검색 결과를 무조건 신뢰하지 않고 실제 논문 존재 여부를 검증한 후 인용 목록을 정리하여 허위 인용 문제를 최소화하는 방식으로 작동함
- 그림 생성과 문헌 탐색이 병렬로 진행되는 구조는 전체 처리 시간을 단축하면서도 각 작업의 독립성을 유지할 수 있게 하며, 이는 논문 작성이 선형적 과정이 아니라 다층적 병렬 작업의 결합임을 보여주는 핵심 설계 요소임

* 시맨틱 스칼라 API(Semantic Scholar API): 미국의 알렌 인공지능 연구소(Allen Institute for AI)가 운영하는 AI 기반 학술 검색 엔진

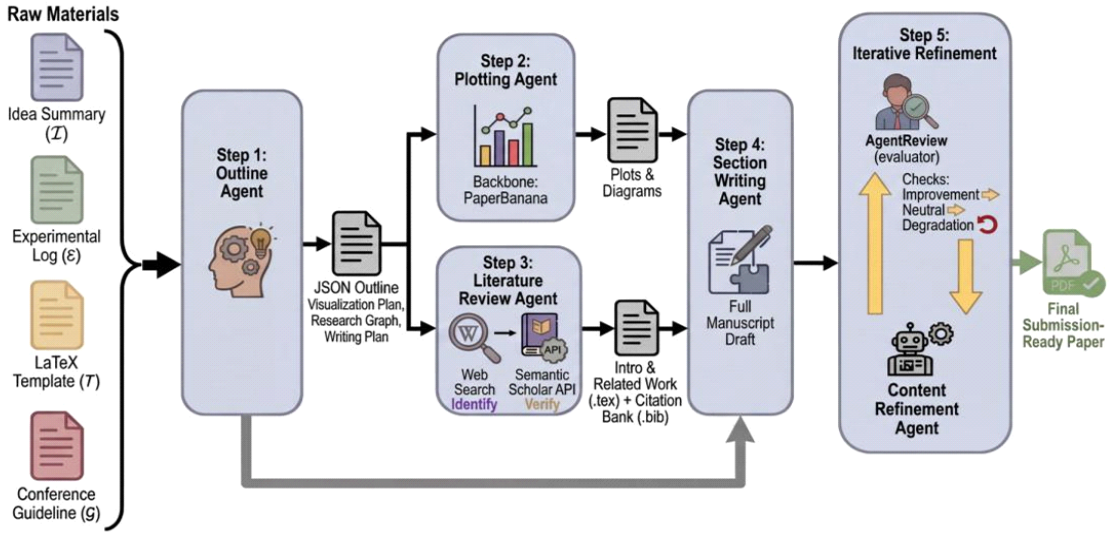
• 반복적 품질 개선을 위한 정제 루프 메커니즘

- 본문 작성 에이전트는 실험 로그에서 수치를 추출하여 표를 생성하고, 앞 단계에서 준비된 그림과 인용을 본문 안에 유기적으로 엮어 초록, 방법론, 실험, 결론 등 전체 논문을 학술 논문 작성용 조판 형식인 LaTeX로 작성함
- 품질 정제 에이전트는 가상의 동료 심사 체계처럼 초안을 검토하되, 수정본의 품질이 향상되었다고 판단될 때만 반영하고 품질이 저하되면 이전 버전으로 되돌리는 방식으로 안정성을 확보함
- 품질 정제 루프를 거친 원고는 초안 대비 자동 평가에서 79~81%의 우세 판정 비율을 기록했으며, 시뮬레이션 학회 심사에서도 컴퓨터 비전 분야 최상위 학회인 CVPR* 기준 19%p, 머신러닝 분야 최상위 학회인 ICLR** 기준 22%p의 채택률 상승을 보여 반복 정제가 품질 향상에 기여함을 입증함
- 전체 시스템은 평균 60~70회의 LLM 호출과 약 40분의 처리 시간을 사용하며, 단일 프롬프트 방식보다 연산 비용이 높지만 품질 향상을 위해 그 비용을 감수하는 설계 선택을 반영함

* CVPR(Computer Vision and Pattern Recognition Conference): IEEE에서 주최하는 컴퓨터 비전과 패턴 인식 분야 학술회의

** ICLR (International Conference on Learning Representations): 기계 학습 및 딥러닝 분야에서 최고 권위를 자랑하는 학회로, 매년 4월 말~5월 초에 개최됨

[그림 1] 페이퍼오케스트라의 다중 에이전트를 활용한 문서 작성 프로세스



출처: Yiwen Song 외 3인, "PaperOrchestra: A Multi-Agent Framework for Automated AI Research Paper Writing", arXiv, 2026.04.06., https://arxiv.org/pdf/2604.05018

• 페이퍼라이팅벤치 기반 성능 평가와 인간 논문 대비 품질 검증

- 연구진은 시스템 성능을 객관적으로 검증하기 위해 2025년 CVPR 및 ICLR에 채택된 논문 200편을 기반으로 아이디어 요약과 실험 로그만 남긴 상태에서 AI가 논문을 재구성하도록 설계된 벤치마크 페이퍼라이팅벤치를 제안함
- 인간 연구자 11명이 참여한 평가에서 페이퍼오케스트라는 기존 방식 대비 문헌 조사 품질에서 50~68% 높은 우세 판정 비율을 기록했고 전체 논문 완성도에서도 14~38% 개선된 결과를 보임
- 자동 평가에서는 문헌 조사 품질이 최대 99%에 가까운 우위를 나타냈으며, 이는 기존 시스템 대비 압도적인 성능 차이를 입증하는 결과임
- AI 기반 논문 평가 프레임워크인 스콜라피어(ScholarPeer)로 페이퍼오케스타의 논문 채택률을 시험한 결과, CVPR 기준 84%, ICLR 기준 81%의 채택률을 기록하여 인간 작성 논문의 채택률인 86%, 94%에 근접한 성능을 보임
- 또한, 페이퍼오케스트라 자체 생성 도표를 포함한 경우에도 절반 이상의 비교 평가에서 인간 결과와 동등하거나 우수한 평가를 받음
- 다만 인간이 작성한 기준 데이터와 비교할 때 여전히 일정한 품질 차이가 확인되었고, 특히 방법론 정의가 촘촘할수록 본문 품질이 향상되는 반면 문헌 조사는 입력 밀도보다 검색 전략과 구조화에 더 크게 좌우된다는 특성이 관찰됨

시사점: AI 보조 학술 생산성 도구의 실효성과 학계의 과제

• 연구 자동화 확산에 따른 윤리적 책임 및 출판 생태계 변화 전망

- 페이퍼오케스트라는 인간을 대체하는 AI 저자가 아니라 연구 생산성을 높이는 보조 도구로 규정되며, 논문의 정확성, 독창성, 윤리적 책임은 여전히 인간 연구자에게 있고 AI가 생성한 결과 역시 검증이 필요하다는 점이 연구진에 의해 명시적으로 강조됨
- 그러나 일각에서는 AI 보조 논문 작성 도구의 확산이 특정 학술 분야의 동료 평가 시스템에 상당한 부담을 주고 있으며, 이를 단순한 기술 활용이 아닌 연구 윤리 차원의 문제로 보는 시각도 존재함
- 페이퍼오케스트라가 제시한 다중 에이전트 기반 자동화 구조는 법률 문서 분석, 금융 모델링 등 다단계 지적 프로세스를 요구하는 다른 영역으로도 확장 가능성이 있으며, 이는 학술 논문 작성을 넘어 지식 생산 방식 전반의 구조적 변화를 예고함
- 따라서 AI가 연구 효율성을 높이는 동시에 학술 출판의 신뢰성과 독창성 기준을 훼손하지 않도록, 기술 개발과 함께 학계 차원의 윤리 지침 및 품질 검증 체계 마련이 병행되어야 한다는 과제가 남음

참고문헌

- Stephen Graves, "Google's PaperOrchestra AI Converts Lab Notes Into Publication-Ready Research Papers", Emerge, 2026.04.10., https://decrypt.co/363837/googles-paperorchestra-ai-converts-lab-notes-into-publication-ready-research-papers
- Yiwen Song 외 3인, "PaperOrchestra: A Multi-Agent Framework for Automated AI Research Paper Writing", arXiv, 2026.04.06., https://arxiv.org/pdf/2604.05018

라이브 스트리밍 불법 유통 탐지를 위한 워터마킹 기반 플랫폼 출시

불법 스트리밍을 실시간으로 탐지 및 차단하는 기술 등장

• 불법 스트리밍 사후 대응 방식의 한계

- 최근 스포츠 및 엔터테인먼트 분야를 중심으로 라이브 스트리밍 콘텐츠의 불법 유통이 확산되고 있음. 이에 따라 신고를 받은 뒤 차단에 나서는 기존 대응 방식으로는 불법 재송출이 이미 퍼진 뒤에야 조치가 이루어진다는 한계가 지적됨
- 이러한 한계를 해결하기 위해, 아일랜드의 디지털 솔루션 기업 스테가웨이브(Stegawave)*는 불법 스트리밍을 실시간으로 탐지하고 차단할 수 있는 플랫폼을 출시함
- 해당 플랫폼은 사전에 원본 라이브 스트리밍에 고유 식별 정보(워터마크)를 삽입해 두고, 불법 스트리밍이 감지되는 즉시 유출 경로를 찾아내 차단하는 구조로 설계되어 있음

* 스테가웨이브(StegaWave): 스트리밍 콘텐츠에 사용자별 식별 정보를 삽입하는 워터마킹 기술을 기반으로, 불법 재송출 영상에서 해당 정보를 추출해 유출 계정을 식별하고 차단하는 보안 플랫폼

작동 구조 및 적용 성과

• 스테가웨이브 플랫폼의 작동 구조

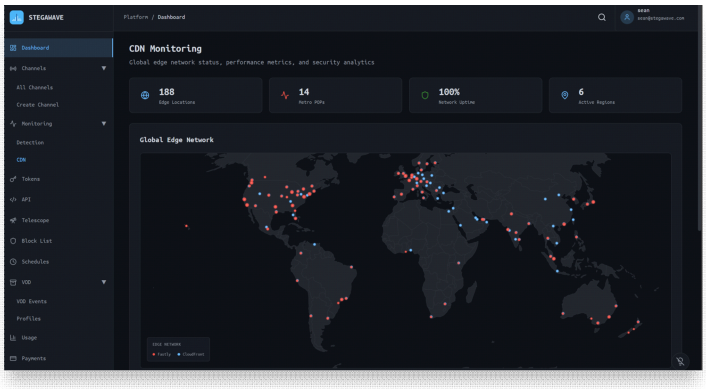
- 플랫폼은 라이브 스트리밍 서비스에 가입한 각 이용자(이하 구독자)에게 고유한 토큰을 발급하여 구독자별 식별이 가능하도록 함
- 해당 토큰을 기준으로 구독자에게 송출되는 각 스트리밍에는 서로 다른 워터마크가 삽입됨
- 라이브 스트리밍이 시작되면, AI가 IPTV 서비스의 채널명을 검색하여 무단으로 송출되고 있는 스트리밍을 탐지하며, 의심 스트리밍에 대해서는 화면을 분석해 원본 라이브 스트리밍과 동일한 방송인지 확인함
- 불법 스트리밍으로 확인되면 해당 영상에 삽입된 워터마크를 분석하여 콘텐츠를 외부로 유출한 구독자의 계정을 특정함
- 유출 계정이 특정되면, 콘텐츠가 구독자에게 전달되는 구간인 CDN(Content Delivery Network)*에서 해당 구독자의 접속을 즉시 차단할 수 있음
- 또한 접속 차단 외에도, 원본 스트리밍 제공자가 미리 준비한 안내 영상 등 대체 콘텐츠로 화면을 전환하는 기능도 지원함

* CDN(Content Delivery Network): 콘텐츠를 사용자에게 전달하는 분산형 네트워크 인프라. 이용자 접속을 제어하거나 차단하는 등 콘텐츠 전달 단계에서 통제가 가능함

- **스테가웨이브 플랫폼의 효과와 적용 성과**

- 이러한 방식은 다수의 불법 스트리밍 서비스가 동일한 유출 계정을 공유하는 구조를 전제로 함 따라서, 단일 계정을 차단하는 것만으로도 여러 불법 스트리밍을 동시에 차단할 수 있음
- 스테가웨이브에 따르면, 해당 기술은 스포츠 라이브 서비스인 클러버 티비(Clubber TV) 적용 사례에서 자사 평가 기준으로 100%의 탐지율을 달성함[1)]
- 이는 해당 기술의 실서비스 적용 가능성을 보여주는 사례임
- 종합해 볼 때, 스테가웨이브의 플랫폼은 아직 초기 적용 단계에 있어 추가 검증이 필요함. 다만 라이브 콘텐츠 보호에 있어 워터마킹 기반 방식이 실효성을 갖추고 있음을 보여줌

[그림1] 스테가웨이브 플랫폼의 CDN 모니터링 화면



출처: NCS, “Stegawave launches anti-piracy platform for live sports streaming”, 2026.04.13., https://www.newscaststudio.com/2026/04/13/stegawave-launches-anti-piracy-platform-for-live-sports-streaming/

[표1] 스테가웨이브 플랫폼의 작동 구조 요약

구분	작동 구조
워터마킹 삽입	① 구독자별 고유 식별 토큰 발급 ② 토큰별로 서로 다른 워터마킹 정보 삽입
불법 스트리밍 탐지	③ AI 프로그램이 채널명을 검색하여 불법 유통 중인 스트리밍을 탐지 ④ 화면 내용을 분석해 원본 라이브 스트리밍과 동일한 방송인지 확인
유출 계정 추적	⑤ 불법 스트리밍 여부가 확인되면 해당 영상에 삽입된 워터마크 정보를 판독 ⑥ 스트리밍을 유출한 구독자 계정을 특정
대응 조치	⑦ CDN에서 해당 구독자의 접속을 즉시 차단하거나, 원본 스트리밍 제공자가 미리 준비한 대체 콘텐츠로 화면을 전환하는 방식 중에서 선택하여 적용 가능

출처: 참고문헌 종합하여 재구성

참고문헌

- NCS,“Stegawave launches anti-piracy platform for live sports streaming”, 2026.04.13., https://www.newscaststudio.com/2026/04/13/stegawave-launches-anti-piracy-platform-for-live-sports-streaming/
- Stegawave Docs, 2026.04.23. 접속 기준, https://stegawave.com/api-docs

1) NCS,“Stegawave launches anti-piracy platform for live sports streaming”, 2026.04.13. https://www.newscaststudio.com/2026/04/13/stegawave-launches-anti-piracy-platform-for-live-sports-streaming/

AI 기반 영상 편집 기술과 동일성 판단 과제

영상 후반 작업에서 기존 기술의 한계

• 기존 영상 인페인팅 기술의 특성과 제약

- 영상 후반 작업에서 불필요한 객체를 삭제하는 인페인팅(inpainting)* 기술은 촬영 장비나 배경 인물의 제거 등에 활용되어 왔으나, 삭제 대상이 주변에 미친 영향을 함께 반영하지 못해 장면을 자연스럽게 이어주는 데 어려움이 있었음
- 예를 들어 인물이 앉아 있던 의자를 제거하면 인물이 공중에 떠 있는 장면이 생성되거나, 물에 뛰어든 사람을 삭제해도 물보라가 그대로 남는 등 삭제된 장면의 연속성이 어긋나는 사례가 발생함
- Runway, ProPainter 등 기존 영상 편집 도구들은 주로 픽셀 단위의 시각적 보정에 집중하며, 화면에서 객체 제거 후 빈 영역을 주변 배경으로 채우는 수준에서 활용됨

* 인페인팅(inpainting): 영상 또는 이미지에서 특정 영역을 제거한 뒤, 주변 정보를 바탕으로 해당 영역을 자연스럽게 채우는 기술

• 실무적 수요와 AI 편집 기능의 확장

- 영상 산업계에서는 촬영 완료 이후 특정 요소를 제거하거나 수정해야 하는 경우가 빈번하여, 재촬영이나 대규모 CG 작업을 대체할 수 있는 기술에 대한 수요가 지속적으로 존재해 왔음
- 이러한 수요를 배경으로 최근에는 AI 기반 편집 기능이 단순한 시각적 보정을 넘어 장면의 구성 요소를 재구성하는 수준으로 확장되고 있으며, 후반 작업에서 구현할 수 있는 변형의 범위가 확대되고 있음

AI 영상 편집 기술의 등장

• VOID 영상 재구성 기술과 확산

- 넷플릭스(Netflix) 연구팀은 2026년 4월 영상에서 특정 객체를 삭제한 뒤, 나머지 장면이 자연스럽게 이어지도록 재구성하는 AI 기반 영상 편집 기술 'VOID(Video Object and Interaction Deletion)'를 공개함
- VOID는 삭제 대상과 그 영향을 받는 영역(그림자, 이동 경로, 접촉 지점 등)을 식별하고, 이를 반영한 쿼드마스크(Quadmask)*를 생성한 뒤 후속 보정을 통해 형태 왜곡을 방지하는 프로세스로 작동함
- 이를 통해 도미노 중간 블록을 삭제할 경우 연쇄 반응이 멈추는 장면을 생성하는 등 객체 간 상호작용이 반영된 결과를 구현하였으며, 사용자 설문에서도 기존 영상 편집 도구 대비 선호도 64.8%를 기록해 2위인 Runway(18.4%)과 성능 차이를 보임[1)]
- 넷플릭스는 VOID를 AI 모델 공유 플랫폼인 허깅페이스(Hugging Face)**에 공개하여 누구나 설치·활용할 수 있도록 배포하였으며, 이는 특정 제작 환경에 한정되지 않는 개방형 확산 구조를 가짐

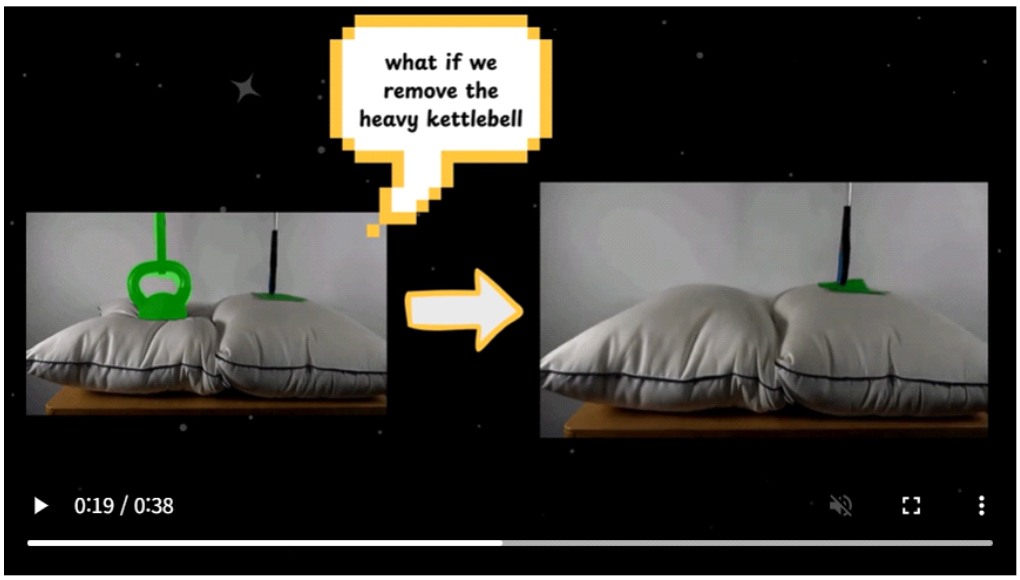
1) Thomas Claburn, "Netflix – yes, Netflix – jumps on the AI bandwagon with video editor", The Register, 2026.04.03., https://www.theregister.com/2026/04/03/netflix_video_ai/

- 이에 따라 독립 제작자, 연구자, 영상 편집 실무자 등 다양한 주체가 해당 기술에 접근할 수 있게 되었으며, 향후 파생 모델의 개발이나 상업적 응용으로 이어질 가능성이 제기되고 있음

* 쿼드마스크(Quadmask): 영상에서 특정 객체를 제거할 때, 그 객체로 인해 영향을 받는 영역을 여러 유형으로 나누어 표시한 뒤, 해당 부분을 자연스럽게 복원할 수 있도록 돕는 마스크 기술

** 허깅페이스(Hugging Face): AI 모델과 데이터셋을 공유·배포할 수 있는 오픈소스 플랫폼으로, 연구자와 개발자가 공개된 모델을 자유롭게 내려받아 활용할 수 있음

[VOID의 객체 삭제 및 물리적 상호작용 예시]



출처: Saman Motamed 외 5인, “VOID: Video Object and Interaction Deletion”, Hugging Face, 2026.04.02., https://huggingface.co/papers/2604.02296

• 편집 환경 내 AI 기능 통합

- 영상 편집 도구 기업 에이비드(Avid)는 2026년 4월 구글 클라우드(Google Cloud)와의 파트너십을 체결하고, 구글의 AI 모델인 제미나이(Gemini)를 자사 편집 소프트웨어인 미디어 컴포저(Media Composer)에 통합함
- 이에 따라 편집자는 키워드 대신 일상적인 문장으로 영상 자료를 검색하거나, AI가 편집 흐름에 맞는 보조 영상을 자동으로 생성하는 기능을 편집 화면 안에서 직접 사용할 수 있게 됨
- 넷플릭스의 VOID가 특정 장면의 물리적 재구성에 초점을 맞춘 기술이라면 에이비드의 사례는 검색·분류·생성 등 편집 전 과정에 AI가 통합되는 흐름을 보여주며, 영상 후반 작업에서 관련 기술이 전반적으로 확장되고 있음을 시사함

시사점: AI 영상 재구성 기술 확산에 따른 과제

• 원본의 변형 가능성과 동일성 판단 과제

- 기존의 영상 편집 기술은 색보정이나 배경 제거 등 시각적 보정에 머물렀으나, VOID와 같이 장면의 물리적 인과관계까지 재구성하는 기술이 등장하면서 촬영된 영상의 사후 변형 가능 범위가 확장되고 있음
- 이에 따라 원본 영상과 편집본 사이의 경계가 모호해질 수 있으며, AI가 장면의 인과관계까지 재구성한 결과물이 저작권법상 저작자의 의사에 반한 변형으로 볼 수 있는지에 대한 판단 기준이 보다 구체적으로 논의될 가능성이 있음

- 특히 관련 기술이 공개 플랫폼을 통해 확산되는 환경에서는 권리자 동의 없이 제3자가 영상을 변형하는 사례가 발생할 수 있어, 변형 결과물의 권리 관계를 어떻게 판단할 것인지에 대한 추가 논의가 필요할 것으로 예측됨

• AI 산출물의 창작 기여도와 저작권 귀속 판단

- 편집 과정에서 보조적 역할에 머물던 기술이 장면 생성 및 재구성 기능까지 수행하게 되면서, 인간의 창작적 기여와 기술적 산출물 간 경계를 구분하기 어려운 상황이 확대되고 있음
- 이에 따라 영상 결과물의 저작권이 누구에게 귀속되는지, 편집자의 창작적 판단과 기술의 기여도를 어떻게 구분할 것인지에 관한 기준 정립이 향후 과제로 남아 있음

참고문헌

• Thomas Claburn, “Netflix – yes, Netflix – jumps on the AI bandwagon with video editor”, The Register, 2026.04.03., https://www.theregister.com/2026/04/03/netflix_video_ai/
• Paul Arnold, “New AI video tool removes objects without breaking the laws of physics”, Tech Xplore, 2026.04.07., https://techxplore.com/news/2026-04-ai-video-tool-laws-physics.html
• Steve Clark, “‘Stay focused on storytelling’: I spoke to Avid about its partnership with Google AI and what it means for creative professionals”, TechRadar, 2026.04.16., https://www.techradar.com/pro/avids-new-google-partnership-brings-agentic-ai-to-the-editing-suite-and-ive-got-the-scoop-on-what-this-really-means-for-creative-professionals
• Saman Motamed 외 5인, “VOID: Video Object and Interaction Deletion”, Hugging Face, 2026.04.02., https://huggingface.co/papers/2604.02296

구글, 개인 사진첩·앱 데이터 활용한 AI 이미지 생성 기능 공개

구글, 개인 사진을 활용한 이미지 생성으로 프롬프트 부담 완화

• 구글, 개인 사진첩과 AI 모델 직접 연동한 맞춤 이미지 생성 기능 공개

- 구글(Google)은 2026년 4월 16일, 이용자 맞춤형 응답 기능인 퍼스널 인텔리전스(Personal Intelligence)*와 이미지 생성 AI 모델 나노 바나나 2(Nano Banana 2)를 결합해, 이용자의 구글 포토 사진첩 사진을 불러와 이미지 생성에 반영하는 기능을 공개함

- 기존 AI 이미지 생성 서비스에서는 이용자가 원하는 결과를 얻기 위해 장문의 프롬프트 작성과 참고 이미지 수동 업로드를 반복해야 했으며, 결과물의 품질도 프롬프트 작성 역량에 크게 좌우됨

- 구글은 이번 기능을 통해 이러한 진입 장벽을 제거하고, 이용자가 짧은 프롬프트만으로도 자신의 취향·일상·주변 인물이 반영된 이미지를 생성할 수 있도록 설계함

  * 퍼스널 인텔리전스(Personal Intelligence): 지메일, 구글 포토 등 구글 앱의 정보를 안전하게 연결해 제미나이가 이용자에게 더 맞춤형 응답을 제공하도록 하는 기능

앱 데이터, 사진첩, 페이스 그룹을 결합한 개인화 메커니즘

• 앱 연동만 되어 있으면 짧은 프롬프트로도 맞춤 결과물 생성

- 퍼스널 인텔리전스는 이용자가 구글 계정에 연결해 둔 앱 데이터를 바탕으로 취향과 선호를 미리 파악해두고, 이미지 생성 요청이 들어오면 이를 자동으로 반영함

- 이미 구글 앱을 연동해둔 이용자는 별도의 추가 설정 없이 "나의 꿈의 집을 디자인해줘"와 같은 짧은 프롬프트만 입력하면 개인 사진첩·앱 정보가 자동 반영된 결과물을 얻을 수 있음

• 페이스 그룹 기능, 가족, 지인, 반려동물까지 이미지 주인공으로 등장

- 이용자가 구글 포토의 개인 사진첩을 퍼스널 인텔리전스에 연결하면, 페이스 그룹(Face Groups)* 기능을 바탕으로 본인, 가족, 친구 등 특정 인물을 구분해 이미지 생성에 반영할 수 있도록 설계됨

- 페이스 그룹은 사람뿐 아니라 반려동물에도 적용할 수 있어, 이용자 주변의 실제 인물이나 동물이 생성 이미지의 주인공으로 등장할 수 있음

  * 페이스 그룹(Face Groups): 구글 포토에서 제공하는 기능으로, 같은 사람이나 반려동물의 얼굴을 자동으로 묶고 이름이나 관계를 붙여 구분할 수 있음

• 비공개 사진은 학습에서 제외, 사전 동의·출처 확인으로 이용자 통제권 보장

- 구글 안내에 따르면 이용자의 비공개 개인 사진첩 자체가 AI 모델 학습에 직접 활용되는 것은 아니지만, 제미나이 앱에 입력되는 프롬프트와 모델 응답 등 일부 정보는 기능 개선을 위해 학습에 활용될 수 있음

- 구글 앱과 제미나이 간 연동은 이용자의 사전 동의가 있어야 활성화되며, 이용자는 설정 화면에서 언제든 연결을 해제하거나 권한 범위를 조정할 수 있음

- 또한 구글은 자동으로 선택된 정보가 이용자의 의도와 일치하지 않을 수 있다는 점을 고려해, 이용자가 소스(Sources) 버튼을 통해 참고된 이미지를 확인하고 다른 참고 사진을 직접 선택하거나 의견을 남겨 결과를 조정할 수 있도록 통제 수단을 마련함

[그림] 개인 사진첩 사진을 반영해 생성한 AI 이미지 예시

출처 Ivan Mehta, "Google adds Nano Banana-powered image generation to Gemini's Personal Intelligence". Tech Crunch, 2026.04.16., https://techcrunch.com/2026/04/16/google-adds-nano-banana-powered-image-generation-to-geminis-personal-intelligence/

'개인정보' 활용 확산, 제3자 초상·학습 데이터 권리 관리 재정비 요구

• 개인 및 제3자 정보 활용 확대에 따른 권리 관리 과제

- 이처럼 개인정보를 활용한 기능이 확산될수록, 데이터의 활용 범위와 목적을 투명하게 고지하는 문제가 지속적인 과제로 남을 수 있음
- 또한 페이스 그룹을 통해 가족, 친구, 반려동물이 이미지 생성에 반영된다는 것은, 이용자 본인뿐 아니라 제3자의 초상과 일상 정보까지 AI 산출물에 포함될 수 있음을 의미함
- 따라서 이러한 기능이 보편화될수록 개인의 저작물은 물론 가족·지인 등 제3자의 초상과 관련한 권리까지 포함하는 방향으로, AI 산출물의 권리 관리 범위를 더욱 정교하게 재설정할 필요성이 있음

참고문헌

- Jay Peters, "Gemini can now pull from Google Photos to generate personalized images", The Verge, 2026.04.17., https://www.theverge.com/tech/913202/gemini-personal-intelligence-images-nano-banana
- Animish Sivaramakrishnan 외 1인, "New ways to create personalized images in the Gemini app", Google, 2026.04.16., https://blog.google/innovation-and-ai/products/gemini-app/personal-intelligence-nano-banana/
- Google Photo Help, "Gemini features in Photos privacy hub", 2026.02.18., https://support.google.com/photos/answer/15344015?hl=en
- Ivan Mehta, "Google adds Nano Banana-powered image generation to Gemini's Personal Intelligence". Tech Crunch, 2026.04.16., https://techcrunch.com/2026/04/16/google-adds-nano-banana-powered-image-generation-to-geminis-personal-intelligence/

깃허브, 애니메이션 불법 스트리밍 저장소 900여 개 삭제

리무브유어미디어, 애니메이션 저작권자의 의뢰 받아 깃허브에 저장소 삭제를 요청

• 저장소 삭제 요청의 배경

- 2026년 3월, 불법 콘텐츠에 대한 삭제 통보를 전문적으로 대행하는 업체 리무브유어미디어(Remove Your Media LLC)는 애니메이션 저작권자의 의뢰를 받아 코드 공유 플랫폼 깃허브(GitHub)에 특정 저장소를 삭제해줄 것을 요청함
- 리무브유어미디어는 이 삭제 요청의 법적 근거로 디지털밀레니엄저작권법(Digital Millennium Copyright Act, 이하 DMCA)의 § 1201(a)(2) 조항*을 제시함

* DMCA § 1201(a)(2): 디지털 콘텐츠에 적용된 기술적 보호조치(예: 암호화, 접근 제한 등)를 우회할 수 있게 하는 기술이나 도구를 만들거나 배포하는 행위를 금지하는 조항

• 불법 스트리밍 영상을 자동 수집하는 기능을 가진 세 개의 저장소 삭제 요청

- 리무브유어미디어가 삭제를 요청한 것은 깃허브에 공개된 세 개의 저장소인 메가클라우드키스(MegacloudKeys), 애니워치(aniwatch), 애니워치 API(aniwatch-api)임
- 이 저장소들에는 애니메이션 불법 스트리밍 사이트인 하이애니메(HiAnime)와 동영상 호스팅 서비스인 메가클라우드(MegaCloud)에서 유료 구독 없이 영상을 자동으로 수집할 수 있도록 만들어진 도구가 담겨 있음
- 세 저장소에 담긴 도구들은 각각 역할을 나누어 서로 연결된 구조로 작동함
- 먼저 메가클라우드키스가 메가클라우드 인프라에서 영상 스트림을 꺼내는 데 필요한 복호화 키를 배포하고, 애니워치가 이 키를 이용해 불법 사이트에 이미 DRM이 제거된 채 게시된 영상을 수집하며, 애니워치 API는 애니워치의 수집 기능을 외부 앱에서도 사용할 수 있도록 제공함
- 리무브유어미디어는 위 세 저장소 외에도 유사한 기능을 가진 저장소 5개를 모두 삭제할 것을 요청하였음. 여기에 이들 8개 저장소를 복제한 포크* 414개 이상이 더해져, 총 삭제 요청 규모는 939개에 달함

*포크(fork): 기존 저장소를 복사해 독립적으로 수정·활용할 수 있도록 만든 복제본

삭제 대상 저장소에 담긴 도구가 접근 통제를 간접적으로 우회한다고 주장

• 리무브유어미디어가 제시한 '간접 우회' 논리

- DRM(Digital Rights Management)은 음원이나 영상 등 디지털 콘텐츠가 무단으로 복제·유출되는 것을 막기 위해 콘텐츠에 적용하는 기술적 보호조치를 말함. 구독 기반 스트리밍 플랫폼은 DRM을 적용해 유료 구독자만 영상을 볼 수 있도록 접근을 통제함

- 리무브유어미디어가 제시한 핵심 논리는 다음과 같음. 삭제 대상 저장소들에 담긴 도구는 정식 플랫폼의 DRM을 직접 해제하지는 않지만, 불법 사이트에서 이미 DRM이 제거된 채 게시된 영상을 자동으로 수집해 줌
- 그 결과 이용자가 유료 구독 없이도 동일한 콘텐츠에 접근할 수 있게 되므로, 이는 접근 통제를 간접적으로 우회하는 것과 같다는 주장임

• 리무브유어미디어, 이번 삭제 요청 건과 유튜브-디엘 사례의 차이점을 강조

- 리무브유어미디어는 삭제 요청문에서 이번 건이 유튜브-디엘(youtube-dl) 사례와 다르다는 점을 명시적으로 강조함
- 유튜브-디엘은 유튜브 등 온라인 플랫폼에서 영상을 내려받을 수 있는 오픈소스 도구로, 2020년 10월 미국음반산업협회(RIAA)의 삭제 요청으로 깃허브에서 일시 제거되었으나, 깃허브가 합법적 용도를 인정하여 복구한 바 있음
- 리무브유어미디어는 이번 삭제 대상 저장소들에 담긴 도구는 유튜브-디엘과 달리 처음부터 불법 스트리밍 사이트만을 대상으로 만들어졌기 때문에, 정당한 이용 목적이 없다고 주장함. 즉, 유튜브-디엘처럼 합법적 용도가 인정될 여지가 없다는 점을 부각한 것임

[깃허브의 삭제 통보문 내 우회 도구 저장소 목록]

V. CIRCUMVENTION DEVICE REPOSITORIES

A. Primary Target: Decryption Key Distribution

Repository: https://github.com/yogesh-hacker/MegacloudKeys

Function: Distributes cryptographic keys enabling extraction of video streams from MegaCloud's piracy infrastructure. These keys are essential for the circumvention tools to function and provide unauthorized access to content that rights holders distribute exclusively through protected, licensed platforms.

§ 1201(a)(2) Analysis: This repository is primarily designed for circumvention, has no commercially significant purpose other than enabling unauthorized access to copyrighted content, and is marketed (through its use by other circumvention tools) for circumventing access controls.

B. Core Circumvention Library

Repository: https://github.com/ghoshRitesh12/aniwatch

Circumvention Code: src/extractors/megacloud.ts

Function: Implements stream extraction from piracy infrastructure, using keys from the MegacloudKeys repository. This library enables automated, scalable access to unauthorized copies of content protected by our clients' access controls.

C. API Implementation

Repository: https://github.com/ghoshRitesh12/aniwatch-api

Forks: 414+ (as of date of notice)

Function: REST API exposing circumvention functionality, enabling third-party applications to programmatically access pirated content without interacting with legitimate, protected platforms.

D. Additional Circumvention Implementations

The following repositories implement similar circumvention functionality:

https://github.com/yahyaMomin/hianime-API

https://github.com/IrfanKhan66/hianime-mapper

https://github.com/Shalin-Shah-2002/Hianime_API

https://github.com/ayanrajpoot10/hianime-api

https://github.com/itzzzme/anime-api

출처: Ernesto Van der Sar, "GitHub Nukes 900+ Anime Piracy Repos and Forks, But Rejects 'Circumvention' Claims", Torrentfreak, 2026.04.22. 접속 기준, https://torrentfreak.com/github-nukes-900-anime-piracy-repos-and-forks-but-rejects-circumvention-claims/

깃허브, 간접 우회 주장은 기각했으나 저작권 침해를 근거로 저장소 삭제

• 삭제 대상 도구가 불법 사이트에서 영상을 수집한다는 점을 근거로 삭제를 집행

- 깃허브는 리무브유어미디어가 제시한 § 1201(a)(2) 위반 주장을 검토한 결과, 이를 인정할 근거가 충분하지 않다고 판단하여 기각함. 다만 깃허브는 기각 이유를 별도로 공개하지 않음
- 깃허브는 간접 우회 주장과는 별개로, 삭제 대상 저장소들에 담긴 도구가 주로 대상으로 삼은 하이애니메와 메가클라우드가 미국영화협회(MPA) 및 미국무역대표부(USTR)가 지정한 악성 시장(Notorious Markets) 목록에 등재된 불법 사이트라는 점에 주목함

- 깃허브는 이를 근거로 저작권 침해 사유를 확인하고, DMCA § 512(c)* 절차에 따라 저장소 삭제를 집행함
- 깃허브는 포크 수가 100개를 초과하는 저장소 네트워크를 묶어 일괄로 삭제를 진행했으며, 그 결과 애니워치 계열 485개, 애니워치 API 계열 310개, 메가클라우드키스 계열 144개가 각각 제거됨

* DMCA § 512(c): 온라인 서비스 제공자가 이용자가 게시한 콘텐츠로 인한 저작권 침해에 대해, 권리자의 삭제 요청에 신속히 응하면 법적 책임을 면할 수 있도록 한 조항

간접 우회 도구가 저작권 보호조치 우회에 해당하는지는 여전히 불분명

• 깃허브가 기각 이유를 공개하지 않아, 해석 기준은 여전히 확립되지 않음

- 깃허브가 기각 이유를 공개하지 않았기 때문에, 간접 우회 방식이 § 1201(a)(2) 위반에 해당하는지에 대한 해석 기준은 여전히 확립되지 않은 상태임
- 이에 따라 향후 유사한 삭제 요청이 제기될 경우, 같은 해석 문제가 반복될 가능성이 있음
- 저작권 집행, 불법 복제 및 유통 이슈를 다루는 전문 매체 토렌트프릭(Torrentfreak)은 과거에도 유사한 서비스들이 이름을 바꿔 다시 등장하는 사례가 반복되었기 때문에, 이번에 삭제된 저장소들에 담긴 도구와 비슷한 서비스가 새로 등장할 가능성을 배제하기 어렵다고 보도함[1)]
- 이번 사례는 간접 우회 방식이 저작권 보호조치 우회에 해당하는지에 관한 해석 기준이 여전히 확립되지 않았음을 보여줌
- 이에 따라 향후 유사 분쟁에서는 플랫폼이 우회 도구 해당 여부와 저작권 침해 여부를 어떤 기준으로 구분해 판단할 것인지가 중요한 과제로 남음

참고문헌

• Ernesto Van der Sar, "GitHub Nukes 900+ Anime Piracy Repos and Forks, But Rejects 'Circumvention' Claims", Torrentfreak, 2026.04.22. 접속 기준, https://torrentfreak.com/github-nukes-900-anime-piracy-repos-and-forks-but-rejects-circumvention-claims/
• Remove Your Media LLC, "2026-03-23-crunchyroll", GitHub DMCA Repository, 2026.04.25. 접속 기준, https://raw.githubusercontent.com/github/dmca/refs/heads/master/2026/03/2026-03-23-crunchyroll.md

1) Ernesto Van der Sar, "GitHub Nukes 900+ Anime Piracy Repos and Forks, But Rejects 'Circumvention' Claims", Torrentfreak, 2026.04.22. 접속 기준, https://torrentfreak.com/github-nukes-900-anime-piracy-repos-and-forks-but-rejects-circumvention-claims/

주간 기술 동향

AI 생성 코드 저작권 보호를 위한 MATRIX 워터마킹 기술

• AI 코드 생성 도구의 확산과 소스코드 유출이 촉발한 지식재산권 보호 논쟁

최근 챗GPT(ChatGPT), 깃허브 코파일럿(GitHub Copilot)과 같은 대규모 언어모델(Large Language Models, 이하 LLM) 기반 코드 생성 도구가 소프트웨어 개발 현장에 도입되면서, AI가 작성한 코드의 저작권 귀속 문제가 새로운 법적·기술적 쟁점으로 부상하고 있다. 개발자들은 AI 도구를 활용해 코딩 시간을 단축하고 있지만, 생성된 코드가 오픈소스 라이선스를 위반하거나 타인의 저작물을 무단으로 학습한 결과물일 수 있다는 우려가 제기되고 있다. 특히 일부 개발자들이 오픈소스 코드를 AI로 약간 변형한 뒤 자신의 창작물로 재배포하는 사례가 증가하면서, 코드 출처를 정확히 추적할 수 있는 기술적 메커니즘의 필요성이 강조되고 있다.

이러한 문제의식은 2026년 4월 앤트로픽(Anthropic)의 AI 코딩 어시스턴트 클로드 코드(Claude Code) 소스코드 유출 사건을 통해 더욱 구체화되었다. 가디언의 보도에 따르면,[1] 소프트웨어 업데이트 과정에서 발생한 인적 오류로 약 2,000개 파일, 50만 줄에 달하는 내부 소스코드가 깃허브에 공개되었다. 또한, 소셜미디어 X에 게시된 유출 코드 링크는 2,900만 회 이상 조회되었다. 유출 코드에는 AI 시스템의 작동 방식과 상업적으로 민감한 기술 정보가 포함되어 있어, 경쟁사인 오픈AI(OpenAI)와 구글(Google)이 이를 분석해 유사 기술을 개발할 수 있다는 우려가 제기되었다. 이번 사건으로 인해 앤트로픽의 보안 취약점이 드러나면서, 코드 유출 방지와 함께 유출된 코드의 출처를 추적하고 저작권을 입증할 수 있는 기술의 중요성이 부각되었다.

기존의 코드 보호 기술은 주로 난독화나 암호화에 의존해왔으나, 이러한 방법들은 코드의 가독성과 유지보수성을 크게 저하시키고 성능에도 부정적 영향을 미친다는 한계를 가지고 있다. 또한 코드 워터마킹 기술의 경우, 코드 리팩토링(refactoring), 최적화, 난독화 등의 변형 공격에 쉽게 제거되거나 손상되는 문제가 있었다. AI가 생성한 코드를 상업적으로 활용하는 사례가 늘어나면서, 코드의 기능성과 성능을 유지하면서도 변형 공격에 강건한 워터마킹 기술에 대한 수요가 증가하고 있다.

본 보고서에서는 이러한 문제를 해결하기 위해 제안된 매트릭스(MATRIX) 워터마킹 기술을 분석한다. MATRIX는 코드에 보이지 않는 식별 정보를 삽입하되, 코드의 실행 결과와 기능은 그대로 유지하는 기술이다. 코드를 수정하거나 변형하더라도 삽입된 워터마크는 남아있어 원작자를 추적할 수 있으며, 이를 통해 AI가 생성한 코드의 저작권을 보호하고 출처를 확인할 수 있다. 이 기술은 LLM 기반 개발 환경에서 실용적으로 활용될 수 있는 가능성을 보여준다.

1) Sanya Mansoor, “Claude’s code: Anthropic leaks source code for AI software engineering tool”, The Guardian, 2026.04.01., https://www.theguardian.com/technology/2026/apr/01/anthropic-claudes-code-leaks-ai

[사례] 변이 기반 변환과 중복 주입을 활용한 MATRIX 코드 워터마킹 기술

• 기술 개요 및 배경

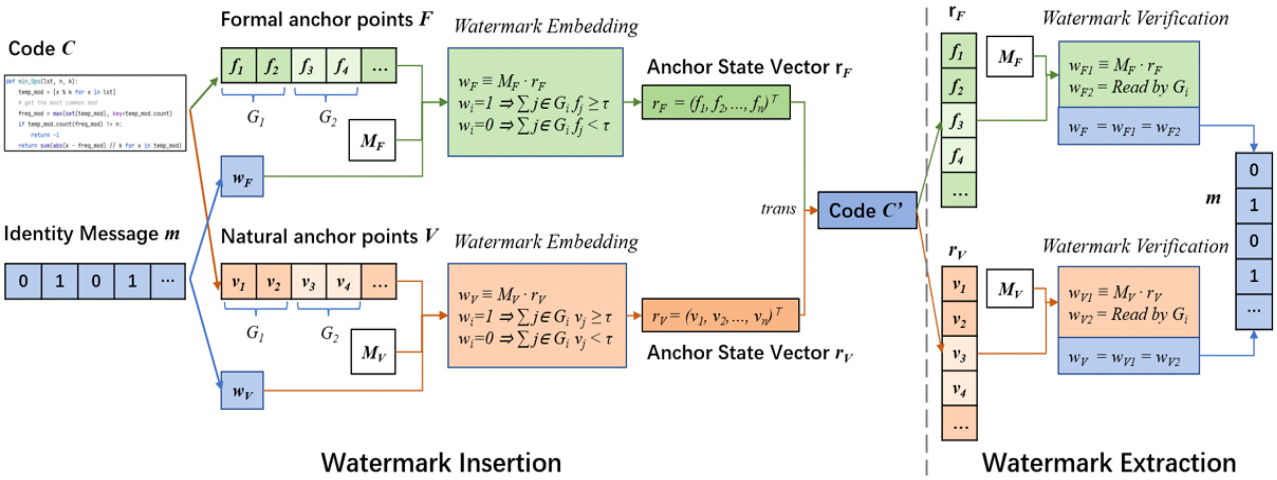
- 기존 코드 워터마킹 기술은 공격자가 코드를 변형하거나 재구성하면 워터마크가 쉽게 손상되거나 제거되는 문제가 있었으며, 워터마크 삽입 과정에서 코드의 실행 성능을 저하시키거나 가독성을 떨어뜨려 실제 개발 환경에서 사용하기 어려웠음
- 이러한 문제를 해결하기 위해 연구자들은 코드의 구조적 특성을 활용하여 변형 공격에도 견딜 수 있는 강건한 워터마킹 기법인 MATRIX를 개발함
- MATRIX는 변이 기반 변환과 중복 주입이라는 두 가지 핵심 기술을 결합하여, 코드의 기능과 성능을 유지하면서도 난독화*, 리팩토링**, 최적화 등 다양한 공격에 강한 워터마크를 삽입할 수 있음

* 난독화(obfuscation): 역공학이나 분석을 방지하기 위해 코드의 실행 결과는 그대로 유지하면서 코드를 읽기 어렵게 만드는 기법
** 리팩토링(refactoring): 소프트웨어의 외부 동작은 변경하지 않으면서 내부 구조를 개선하는 작업으로, 코드의 가독성과 유지보수성을 높이는 과정

• 변이 기반 변환을 활용한 워터마크 삽입

- MATRIX는 코드의 추상 구문 트리를 분석하여 실행 결과는 동일하지만 구조가 다른 코드로 변환하는 방식으로 워터마크를 삽입함. 예를 들어 for 반복문을 while로 변환하거나 if 조건문 순서를 재배치하는 등의 방법으로 워터마크 정보를 코드 구조에 내재화함
- 변형된 코드는 원본과 완전히 동일한 실행 결과를 출력하도록 의미적 동등성을 보장하며, 자동화된 단위 테스트를 통해 워터마크 삽입이 코드 기능에 영향을 주지 않음을 검증함
- 공격자가 코드를 변형하거나 재구성하더라도 워터마크는 코드의 심층 구조에 내재되어 있어 제거가 어려우며, 완전한 제거를 위해서는 공격자가 코드의 논리 구조를 바꿔 처음부터 재작성해야 함
- MATRIX는 여러 변형 기법을 무작위로 조합하여 사용하므로 공격자가 대량의 샘플을 수집하여 워터마크 생성 패턴을 학습하고 자동 제거 도구를 개발하기 어려움

[그림 1] MATRIX의 워터마크 삽입 및 추출 프로세스



출처: Yuqing Nie 외 6인, "MATRIX: Multi-Layer Code Watermarking via Dual-Channel Constrained Parity-Check Encoding", arXiv, 2026.04.17., https://arxiv.org/pdf/2604.16001

• 중복 주입을 활용한 워터마크 보호

- MATRIX는 동일한 워터마크 정보를 코드의 여러 위치에 반복 삽입하여, 일부 워터마크가 공격으로 손상되거나 제거되더라도 나머지 부분을 통해 원본 식별 정보를 완전히 복원할 수 있도록 설계함
- 이 기법은 오류 정정 코드를 활용하여 워터마크를 인코딩하며, 워터마크의 일부가 손상되어도 나머지 정보를 바탕으로 원본 메시지를 자동으로 복구할 수 있음
- 워터마크는 변수 이름, 주석, 코드의 제어 흐름 등 다양한 레이어에 분산 배치되어, 공격자가 모든 워터마크 위치를 찾아서 동시에 제거하는 것을 사실상 불가능하게 만듦
- 워터마크 시퀀스와 원본 메시지 간의 관계가 암호학적으로 복잡하게 설계되어 있어, 공격자가 다수의 샘플을 수집하여 분석하더라도 워터마크 생성 규칙을 역추적하기 어려움

• 워터마크 검출 및 보안

- MATRIX는 워터마크 검출 시 코드에서 워터마크가 삽입된 위치들을 식별한 후, 각 위치에서 워터마크 비트를 추출하여 다수결 방식으로 최종 워터마크를 결정함
- 이 방식은 일부 워터마크가 손상되거나 변형되어도 나머지 워터마크들의 정보를 종합하여 정확한 검출을 가능하게 하며, 노이즈에 강건한 특성을 보임
- 워터마크 시퀀스와 식별 메시지 간의 대응 관계가 암호학적으로 복잡하게 설계되어 있어, 공격자가 대량의 워터마크 샘플을 수집하여 패턴을 학습하더라도 생성 규칙을 파악하기 어려움
- 공격자가 코드의 일부만 복사하거나 여러 출처의 코드를 혼합하는 경우에도, MATRIX는 부분적으로 남아있는 워터마크 조각들을 분석하여 원본 출처를 추적할 수 있음
- MATRIX는 암호학적 서명 기법을 활용하여 워터마크의 진위 여부를 검증하며, 이를 통해 공격자가 가짜 워터마크를 생성하거나 기존 워터마크를 조작하는 것을 효과적으로 차단함

결론 및 시사점

• 성능 및 기술적 성과

- MATRIX는 코드 난독화, 재작성 등 12가지 공격 시나리오에서 평균 95% 이상의 워터마크 검출률을 기록했으며, 기존 기법 대비 변형 공격에 대한 강건성이 평균 42% 향상된 것으로 나타남
- 또한, 코드의 기능과 실행 성능에 미치는 영향을 최소화하면서도 기존 기술 대비 워터마크를 빠르게 삽입할 수 있어, AI 생성 코드에 실용적으로 적용 가능한 저작권 보호 기술임을 보임

• 산업적 의의와 향후 과제

- MATRIX는 AI 생성 코드의 저작권 분쟁 해결과 출처 추적이라는 문제에 대한 기술적 해법을 제시하며, 오픈소스 커뮤니티와 상업적 개발 환경 모두에서 신뢰할 수 있는 코드 생태계 구축에 기여할 수 있음
- 향후 과제로는 고도로 최적화된 컴파일러 환경에서의 강건성 향상, 대규모 코드베이스 처리 시 계산 비용 최적화, 법적 분쟁 시 워터마크의 증거 능력 확보를 위한 표준화 등이 남아 있음

참고문헌

- Yuqing Nie 외 6인, "MATRIX: Multi-Layer Code Watermarking via Dual-Channel Constrained Parity-Check Encoding", arXiv, 2026.04.17., https://arxiv.org/pdf/2604.16001
- Sanya Mansoor, “Claude’s code: Anthropic leaks source code for AI software engineering tool”, The Guardian, 2026.04.01., https://www.theguardian.com/technology/2026/apr/01/anthropic-claudes-code-leaks-ai