Biweekly Report
통권 제82호(2026. 5-2)

www.copyright.or.kr

# 저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

# CONTENTS

## 저작권 이슈 트렌드

Biweekly Report | 통권 제82호(2026. 5-2)

# 에이전틱 AI 번역 시스템의 등장과 인간 번역가의 대체 가능성

## 뉴스 브리프

생성형 AI의 확산은 유럽 번역 시장에 직접적인 충격을 가하고 있다. 프랑스와 영국의 최근 조사에 따르면, 번역가의 79%가 AI로 인한 일자리 대체 위협을 체감하며, 84%는 수요 감소와 임금 하락을 예상한다. 이러한 변화는 번역 산업의 구조 자체를 재편하는 방향으로 이어지고 있다. 그러나 대규모 언어 모델 기반 기계번역은 비용과 속도 면에서 압도적 우위를 점하지만, 문맥 이해 실패와 창의적 표현 부재라는 한계를 동시에 드러낸다. 이러한 배경에서 번역학 이론을 생성형 AI의 명령 체계로 전환한 에이전틱 번역 시스템이 등장했다. 이 시스템은 번역 목적과 독자를 사전에 구조화하고, 오류 검증과 문서 수준 일관성 유지 메커니즘을 통합하여 기계번역의 구조적 한계를 보완한다. 본 보고서는 이 기술의 작동 원리와 산업적 함의를 분석하고, 인간 번역가의 역할이 어떻게 재정의되는지 전망한다.

## 뉴스 플러스

### I. 서론 : 기계번역 시대, 인간 번역가의 존재 이유

- **기술 확산과 번역 노동의 재편**

2024년 이후 유럽의 번역 산업은 생성형 AI의 확산과 함께 급격한 변화를 겪고 있다. 프랑스 저작권 단체와 작가 협회의 공동 조사 결과, 번역가의 79%가 AI 기술이 자신의 업무를 대체할 위협으로 인식한다고 응답했다. 영국에서 실시된 2025년 조사에서도 84%의 번역가가 인간 번역의 수요 감소로 인한 수입 하락을 전망했다.[1] 이러한 수치는 단순한 기술적 변화에 대한 우려가 아니라, 본인의 직업과 관련 산업 자체에 대한 불안을 반영한다. 번역 시장은 200개 이상의 언어가 공존하고 있는 유럽에서 필수적인 산업이지만, 대규모 언어 모델(Large Language Model, 이하 LLM)을 활용한 기계번역 도구의 비용 효율성은 전통적인 인간 번역의 수요를 빠르게 잠식하고 있다.

1) Philip Oltermann, "'Being human helps': despite rise of AI is there still hope for Europe's translators?", The Guardian, 2026.05.08., https://www.theguardian.com/technology/2026/may/08/being-human-helps-despite-rise-of-ai-is-there-still-hope-for-europes-translators

그러나 기술 발전 속도만큼이나 기계번역의 한계도 명확히 드러나고 있다. 기존의 기계번역은 입력된 텍스트를 받아 목표 언어로 변환된 텍스트를 출력하는 단방향 처리 방식으로 작동한다. 이 과정에서 시스템은 문장과 단락 단위로 언어 간 대응 관계를 계산하지만, 번역이 수행되는 목적이나 텍스트가 도달해야 할 예상 독자층에 대한 정보는 처리 과정에 포함되지 않는다. 동일한 원문이라도 누가 읽을 것인지, 어떤 상황에서 사용될 것인지에 따라 번역 결과는 달라져야 하지만, 현재의 도구들은 이러한 변수를 입력받을 경로 자체가 없다. 결과적으로 생성된 번역문은 일정 수준의 문법적 정확성은 확보할 수 있으나, 실제 사용 맥락에서는 부적절한 표현이나 어조를 담게 된다. 이는 기계번역 시스템이 텍스트의 표면적 의미만을 처리할 뿐, 그 텍스트가 사용될 구체적 상황이나 독자의 기대를 파악할 수 없기 때문이다.

이러한 기존 기계번역의 한계를 보완하기 위해 등장한 에이전틱 번역 시스템(agentic translate system)은 이러한 간극을 메우기 위해 번역 과정을 다단계 처리 사이클로 설계되었다. 이 시스템은 번역 작업을 시작하기 전에 사용자와 대화를 통해 목적, 예상 독자, 장르 및 어조 등을 구조화된 데이터 형식으로 수집하고, 이를 생성 과정 전반에 반영한다. 또한 생성된 번역문을 다차원 품질 평가((Multidimensional Quality Metrics, MQM) 프로토콜로 자동 검증하고, 오류를 재입력하여 번역문을 개선하는 반복 구조를 갖추고 있다. 이는 번역을 단순 언어 변환에서 상황과 목적에 맞는 소통 작업으로 전환하려는 시도이며, 인간 번역가의 판단 과정을 기술적으로 구현하려는 접근이다.

## II. 본론: 번역 패러다임의 전환과 기술적 가능성

### • 단방향 변환 구조의 기술적 한계

현재 널리 사용되는 기계번역 도구들은 번역 과정에서 텍스트의 사용 목적이나 예상 독자에 대한 정보를 요구하지 않는다. 사용자가 원문을 입력하면 시스템은 즉시 언어 간 대응 관계를 계산하여 번역문을 생성하지만, 그 텍스트가 어떤 상황과 목적에서 읽힐 것인지는 고려 대상이 아니다. 이러한 단방향 처리 방식은 입력 데이터와 출력 데이터 사이에 중간 조정 단계가 존재하지 않는다는 것을 의미한다. 원문이 시스템에 입력되는 순간, 번역 방향과 스타일에 대한 모든 결정은 사전 학습된 가중치에 의해 자동으로 처리된다.

이로 인해 동일한 원문이 서로 다른 맥락에서 사용될 때 요구되는 뉘앙스 차이를 반영할 수 없다. 기술 문서는 정확성과 용어 일관성을 우선하지만, 마케팅 자료는 설득력과 감정적 호소를 요구한다. 그러나 기존 시스템은 이러한 구분 없이 통계적으로 가장 빈번한 번역 패턴을 적용한다. 결과적으로 자동 번역된 텍스트는 문법적으로는 정확할 수 있으나, 실제 사용 맥락에서는 적절하지 않은 표현이나 어조를 담을 가능성이 높다. 사용자는 번역 결과를 받은 뒤에야 그것이 자신의 의도와 맞지 않음을 발견하고, 직접 수정하는 과정을 거쳐야 한다.

이러한 문제는 번역이 단순히 언어 기호를 치환하는 작업이 아니라, 특정 목적을 달성하기 위해 텍스트를 재구성하는 행위라는 점을 간과한 데서 비롯된다. 번역 요청이 발생하는 순간, 그 뒤에는 반드시 의도된 목적과 예상 독자가 존재한다. 그러나 기존 기계번역 도구들은 이러한 정보를 입력받을 설계를 갖추지 못했다. 사용자는 '이 텍스트는 누구를 위한 것인가', '어떤 효과를 내야 하는가'를 시스템에 입력할 방법이 없으며, 시스템 역시 그 질문을 던지지 않는다. 번역 과정의 모든 판단은 알고리즘에 위임되고, 사용자는 결과물을 받아 사후에 조정하는 수밖에 없다.

### • 4단계 에이전틱 사이클의 설계 원리

일본 릿쿄 대학교(立教大学, Rikkyo University) 연구진이 개발한 에이전틱 번역 시스템은 번역 과정을 4개의 독립된 단계로 분리하고, 각 단계가 순환하며 결과물을 개선하는 구조로 설계되었다.

첫 번째 단계는 식별 단계(identification)로, 사용자가 원문을 입력하면 시스템은 10개 항목으로 구성된 구조화된 문서를 제안한다. 이 과정은 번역 실행 전 제약 조건을 명시적으로 정의하는 단계로, 기존 도구와 차별화되는 고유한 절차다. 두 번째는 생성 단계(generation)로, 수립된 조건이 프롬프트에 주입되어 번역문이 산출된다. 세 번째는 검증 단계로, 생성된 번역문은 다차원 품질 평가 프로토콜에 따라 자동으로 분석된다. 네 번째는 개선 단계로, 검증에서 탐지된 오류가 다시 생성 프롬프트에 반영되어 번역문이 수정된다. 이 사이클은 품질 기준을 충족하거나 사용자가 중단할 때까지 피드백 루프를 통해 반복된다.

**[그림 1] 에이전틱 번역 시스템의 조건 설정 10개 항목**

| 항목 | 내용 |
|---|---|
| 목적(skopos) | 번역이 달성해야 할 효과와 의도된 결과 명시 |
| 예상 독자(audience) | 번역문을 읽을 대상의 배경 지식 수준, 전문성, 문화적 맥락 설정 |
| 어조(register and voice) | 화자가 독자에게 전달하려는 태도와 격식 수준 설정 |
| 장르(genre) | 텍스트 유형과 장르별 관습, 형식적 제약 지정 |
| 용어 지침(terminology guidance) | 특정 단어나 개념의 번역 방식 고정으로 문서 전체 일관성 유지 |
| 문체 결정(style decisions) | 문장 구조, 능동태/수동태 선택, 문장 길이 등 문체적 특성 조정 |
| 보존할 요소(things to preserve) | 원문에서 반드시 유지해야 할 특성 명시 |
| 현지화할 요소(things to localise) | 목표 언어 문화권 관습에 맞게 조정할 부분 지정 |
| 피해야 할 표현(things to avoid) | 번역문에서 사용 금지할 어휘, 구문, 표현 방식 명시 |
| 미결 질문(open questions) | 시스템이 자동 판단하기 어려운 사항에 대한 사용자 확인 요청 |

출처: Masaru Yamada, "Agentic AI Translate: An Agentic Translator Prototype for Translation as Communication Design", arXiv, 2026.05.16., https://arxiv.org/pdf/2605.17041

조건 설정 단계의 10개 항목은 스코포스 이론이나 레지스터 이론을 비롯한 목적 중심의 번역학 개념들을 실행할 수 있는 형식으로 재구성한 것이다. 각 항목은 번역가가 암묵적으로 수행하던 판단을 명시적 절차로 변환한 형태로, 사용자가 직접 편집하거나 시스템과 대화를 통해 수정할 수 있으며, 만족할 때까지 반복 조정이 가능하다. 확정된 조건은 구조화된 데이터 형식으로 변환되어 생성 프롬프트에 주입되며, 생성 모델은 이를 참조하여 각 문장의 어휘 선택, 문장 길이, 어조를 결정한다.

긴 문서의 경우 일정 길이의 단위로 나뉘어 처리되며, 각 단위는 독립적으로 번역되지만 이전 단위에서 확립된 용어 목록과 문맥 정보를 이어받는다. 시스템은 각 단위를 번역할 때마다 세 가지 데이터를 생성한다. 첫째, 해당 단위에서 등장한 고유명사와 전문 용어의 번역 쌍을 기록한 용어 목록이다. 둘째, 지금까지 번역된 내용의 핵심을 요약한 문서 수준 요약이다. 셋째, 바로 직전 단위의 번역문이다. 이 세 가지 데이터는 다음 단위를 번역할 때 프롬프트에 명시적으로 주입된다. 예를 들어 첫 번째 단위에서 "Natsume Soseki"가 "나쓰메 소세키"로 번역되었다면, 이 대응 관계가 용어 목록에 저장되고, 이후 단위에서 동일 인물이 등장할 때 자동으로 "나쓰메 소세키"로 번역된다. 이를 통해 문서 전체의 용어 일관성과 문맥 연속성이 유지된다.

조건 설정 단계를 별도의 처리 과정으로 분리한 이유는 두 가지다. 첫째, 상황 분석 결과를 구조화된 데이터로 산출함으로써 번역학 개념을 명시적 항목으로 가시화한다. 둘째, 사용자가 생성 이전에 시스템이 수행한 상황 분석을 확인하고 수정할 수 있도록 한다. 이는 번역가가 작업 전에 수행하는 판단 과정을 알고리즘이 실행할 수 있는 형태로 재구성한 것이며, 번역을 단순 언어 변환에서 상황과 목적에 맞는 소통 작업으로 전환하려는 시도다.

**• 다차원 품질 평가와 반복 개선 메커니즘**

생성 단계에서 산출된 번역문은 검증 단계로 넘어가 평가를 거친다. 이 단계에서 사용되는 GEMBA-MQM(GPT Estimation Metric Based on Multidimensional Quality Metrics) 프로토콜은 GPT 모델을 활용하여 다차원 품질 측정 체계에 따른 번역 평가를 자동화한 방식이다. 다차원 품질 측정 체계는 번역 오류를 정확성, 유창성, 용어, 문체 및 현지화 등 여러 차원으로 분류하고 각 오류에 점수를 부여하는 구조화된 평가 방식을 의미한다. 쉽게 말해, 번역문의 문제를 유형별로 나누어 점검하고 각각 얼마나 심각한지 수치로 표시하는 것이다. GPT-4는 LLM 중 하나로, 텍스트를 이해하고 평가할 수 있는 능력을 갖춘 AI 모델이다. GEMBA-MQM은 이 평가 체계를 GPT-4가 자동으로 수행하도록 설계되어, 생성된 번역문이 조건 설정 단계에서 정한 조건들을 충족하는지 체계적으로 검증한다.

검증 과정은 원문과 번역문을 동시에 입력받아 문장 단위로 비교하며 진행된다. 시스템은 각 문장에서 오역, 누락, 추가, 용어 불일치 및 문체 부적합 등을 탐지하고, 각 오류에 대해 경미함, 보통, 심각함의 3단계 심각도를 부여한다. 예를 들어 "system"을 "systems"로 잘못 번역한 경우 용어 불일치로 분류되며, 문맥상 의미 전달에 큰 영향을 미치지 않으면 경미한 오류로 표시된다. 반면 핵심 개념을 완전히 다른

의미로 번역한 경우 정확성 오류로 분류되며 심각한 오류로 기록된다. 모든 오류는 발생 위치, 유형, 심각도와 함께 목록으로 정리되어 개선 단계로 전달된다.

개선 단계에서는 검증 단계에서 탐지된 오류 목록이 다시 생성 프롬프트에 추가되어 번역문이 재생성된다. 오류 목록은 "3번째 문장에서 용어 불일치 발견: 'system'을 'systems'로 번역함"과 같이 구체적으로 명시되며, 생성 모델은 이를 수정 지침으로 해석하여 해당 부분만 조정한다. 이 과정은 단순히 오류를 지적하는 것이 아니라, 왜 오류인지에 대한 맥락 정보를 함께 제공한다. 예를 들어 문체 부적합 오류의 경우, 조건 설정 단계에서 지정한 어조와 실제 번역문의 어조가 어떻게 다른지를 명시한다. 재생성된 번역문은 다시 검증 단계로 넘어가 평가를 받으며, 이 사이클은 설정된 품질 기준을 충족하거나 사용자가 수동으로 중단할 때까지 반복된다. 이러한 순환 구조는 번역을 일회성 산출 작업이 아니라 점진적 품질 개선 과정으로 전환하며, 각 반복마다 이전 오류 정보가 누적되어 시스템의 판단 정확도가 향상된다.

## III. 결론 및 전망: 협업 모델로서의 미래

### • 기술 발전이 제시하는 새로운 역할 분담

에이전틱 번역 시스템은 기계번역의 한계를 기술적으로 보완하려는 시도지만, 동시에 전문 번역가의 역할 중요성을 재확인하는 결과를 낳는다. 이 시스템이 조건 설정 단계를 별도로 분리한 이유는 번역이 단순 언어 변환이 아니라 고도의 판단과 맥락 해석이 요구되는 작업임을 고려했기 때문이다. 10개 항목으로 구조화된 조건 설정 과정은 번역가가 수행하는 상황 분석, 독자 고려, 전략 수립 과정을 명시적 절차로 전환한 것이며, 이는 역설적으로 전문 번역가의 판단이 아직 알고리즘으로 대체될 수 없음을 보여준다. 시스템은 조건을 입력받아 실행할 수 있지만, 그 조건을 설정하는 주체는 여전히 인간이다. 결국 기계번역 기술의 발전은 번역가의 역할을 대체하는 것이 아니라, 단순 반복 작업에서 벗어나 전략적 판단과 품질 관리에 집중할 수 있도록 재배치하는 효과를 낳는다.

이러한 변화는 번역 교육과 산업 구조에도 영향을 미칠 것으로 보인다. 제네바 대학교(Université de Genève)의 번역통역학부가 생성형 AI 등장 이후 감소했던 지원자 수가 다시 회복 추세를 보인다고 밝힌 것은,[2] 시장이 초기 충격을 넘어 기술과 인간 역량의 차이를 새롭게 인식하기 시작했음을 시사한다. 향후 번역가에게 요구되는 역량은 언어 능력뿐 아니라 AI 도구 활용 능력, 조건 설정 및 품질 평가 능력, 커뮤니케이션 설계 능력으로 확장될 것이다. 동시에 번역 산업은 AI가 처리할 수 있는 표준화된 번역과 인간 판단이 필수적인 고부가가치 번역으로 분화될 가능성이 크다. 이 과정에서 번역 노동의 가치를 어떻게 재평가하고, 기술 발전의 이익을 어떻게 분배할 것인지에 대한 사회적 합의가 필요할 것이다.

2) Philip Oltermann, "'Being human helps': despite rise of AI is there still hope for Europe's translators?", The Guardian, 2026.05.08., https://www.theguardian.com/technology/2026/may/08/being-human-helps-despite-rise-of-ai-is-there-still-hope-for-europes-translators

## 참고문헌


- Philip Oltermann, "'Being human helps': despite rise of AI is there still hope for Europe's translators?", The Guardian, 2026.05.08., https://www.theguardian.com/technology/2026/may/08/being-human-helps-despite-rise-of-ai-is-there-still-hope-for-europes-translators
- Masaru Yamada, "Agentic AI Translate: An Agentic Translator Prototype for Translation as Communication Design", arXiv, 2026.05.16., https://arxiv.org/pdf/2605.17041

# 독일 최대 이미지 에이전시, iMATAG의 비가시적 워터마킹 기술 채택

## 뉴스 브리프

생성형 인공지능 기술의 확산으로 실제 촬영 이미지와 AI 산출물을 구분하기 어려워지면서, 이미지 콘텐츠의 진위 검증이 산업계의 핵심 과제로 떠올랐다. 독일 최대 이미지 에이전시 dpa Picture Alliance는 IMATAG의 비가시적 워터마킹 기술을 도입해 3억 장 이상의 이미지 카탈로그에 식별자를 삽입하고, 라이선스 집행과 콘텐츠 진위성 강화에 나섰다. 산업계 대응과 동시에 학계에서는 AI 생성 이미지를 자동으로 표시하기 위한 워터마킹 기술 연구가 진행 중이다. 잠재 확산 모델의 생성 과정에 워터마크를 삽입하는 가이던스 기반 접근법은 이미지 품질을 유지하면서도 높은 검출 성능을 달성했다. 진본 이미지 보호와 합성 이미지 식별이라는 양방향 워터마킹 기술은 이미지 산업에서 신뢰할 수 있는 검증 기술 중 하나로 자리 잡을 것으로 예상된다.

## 뉴스 플러스

### I. 서론: 생성형 AI 시대, 이미지 진위 검증의 새로운 과제

**• 진본과 합성의 구분 문제와, 산업계의 대응**

2024년 이후 생성형 인공지능 기술이 대중화되면서 실제 촬영 이미지와 AI 산출 이미지를 구분하는 일이 어려워졌다. 이미지 콘텐츠 산업은 저작권 보호와 함께 콘텐츠 진위 검증이라는 과제에 직면하게 되었다. 독일 최대 이미지 에이전시 중 하나인 dpa Picture Alliance는 2026년 5월 6일, 프랑스의 워터마킹 솔루션 기업 iMATAG와 협력해 자사가 보유한 3억 장 이상의 이미지 카탈로그에 비가시적 식별자를 일괄 삽입하기로 결정했다. 이는 라이선스 집행과 무단 사용 추적을 넘어, 해당 이미지가 실제 촬영된 것임을 증명하려는 조치로 해석된다.

dpa Picture Alliance 측은 이번 결정의 배경을 "신뢰할 수 있는 시각 정보의 출처로서 우리의 역할은 사진작가의 권리를 보호하는 동시에, 우리가 배포하는 이미지의 무결성을 보장하는 것"이라고 설명했다. 생성형 AI가 만들어낸 이미지가 뉴스 사진이나 상업용 이미지로 유통될 경우, 콘텐츠 이용자가 진위를 판단하기 어려워진다. 특히 수십억 장의 이미지가 매일 유통되는 소셜 네트워크와 검색 엔진에서, 자동화된 검증 체계에 대한 필요성이 제기되고 있다.

**• 비가시적 워터마킹 기술의 부상 배경**

디지털 워터마킹은 이미지에 육안으로 확인할 수 없는 식별 정보를 삽입하고, 전용 검출기로 이를 추출하는 기술이다. 이 기술은 원래 저작권 보호, 시청률 측정, 콘텐츠 식별 및 수익화, 방송 모니터링 등에 활용되어 왔으며, 최근 AI 산출 콘텐츠의 식별로 적용 범위가 확대되고 있다. 미국 국가안보국(National Security Agency, NSA)은 2025년 발표한 보고서에서 워터마킹 기술의 주요 활용 시나리오로 두 가지를 제시했다. 첫째는 소셜 네트워크나 검색 엔진 이용자에게 해당 이미지가 실제가 아님을 알리는 것이고, 둘째는 향후 생성형 AI의 학습 데이터에서 AI 산출 이미지를 걸러내 모델 성능 저하를 방지하는 것이다. 이는 AI가 생성한 합성 데이터가 재학습 과정에 반복 투입될 경우, 데이터의 다양성이 상실되고 오류가 증폭되는 '모델 붕괴(Model Collapse)' 현상을 방지하기 위함이다.

두 시나리오 모두 수십억 장의 이미지를 자동으로 분석해야 하므로, 높은 탐지 신뢰도와 낮은 오탐률이 요구된다. AI 산출물 식별을 위한 워터마킹에서 오탐(false positive)은 워터마크가 없는 진본 이미지를 AI 산출물로 잘못 검출하는 경우를 의미한다. 반대로 미탐(false negative)은 워터마크가 삽입된 AI 산출 이미지를 검출하지 못하는 경우다. 특히 오탐률이 높을 경우 실제 촬영 이미지가 AI 산출물로 오인되어 부당하게 필터링될 수 있으므로, 산업 적용을 위해서는 통계적으로 검증 가능한 수준의 낮은 오탐률 달성이 필요하다.

## II. 본론: 잠재 확산 모델과 워터마킹 메커니즘

**• 잠재 확산 모델의 작동 원리와 이미지 생성 과정**

잠재 확산 모델(Latent Diffusion Model)은 이미지를 직접 생성하는 대신, 압축된 잠재 공간에서 작업한 뒤 이를 다시 이미지로 변환하는 방식으로 작동한다. 이 과정은 크게 세 단계로 구성된다. 먼저 인코더가 고해상도 이미지를 저차원 잠재 표현으로 압축하고, 확산 모델이 잠재 공간에서 노이즈 제거 과정을 반복하며, 마지막으로 디코더가 잠재 표현을 다시 이미지로 복원한다. 이러한 접근법은 계산 효율성을 크게 향상시켜, 스테이블 디퓨전 2(Stable Diffusion 2), 플럭스(Flux), 사나(Sana) 같은 이미지 생성 모델의 실용화를 가능하게 했다.

확산 과정의 핵심은 점진적인 노이즈 제거다. 모델은 가우시안 노이즈(Gaussian noise) 상태에서 시작해 수십 단계에 걸쳐 조금씩 노이즈를 제거하며 이미지를 형성한다. 각 단계에서 모델은 현재 상태와 텍스트 프롬프트를 참고해 다음 상태를 예측한다. 이 반복적 과정을 통해 표준 정규분포를 따라 무작위로 배치된 노이즈가 점차 의미 있는 이미지로 변환되며, 텍스트 조건에 따라 생성 방향이 조절된다. 이러한 단계별 생성 과정은 워터마크를 삽입할 수 있는 여러 시점을 제공한다는 점에서 저작권 보호 기술 적용에 중요한 의미를 지닌다.

**• 생성 단계 워터마크 삽입, 가이던스 원리 접근법**

생성 과정 중 워터마크를 삽입하는 가이던스 원리 접근법은 기존 워터마크 디코더를 활용해 확산 과정을 유도하는 방식이다. 이 방법은 별도의 워터마크 인코더를 학습시킬 필요 없이, 이미 검증된 디코더만으로 워터마크를 삽입할 수 있다는 장점이 있다. 구체적으로는 각 노이즈 제거 단계에서 생성 중인 이미지를 워터마크 디코더에 통과시켜 검출 점수를 계산하고, 이 점수가 높아지는 방향으로 생성 과정을 조정한다.

논문에서 제시된 G-SSig와 G-VS 두 가지 방법은 이러한 가이던스 원리를 구현한 대표적 사례다. G-SSig는 스테이블 시그니처(Stable Signature) 디코더를 활용하며, G-VS는 비전 실딩(Vision Shielding) 디코더를 사용한다. 두 방법 모두 생성 과정의 마지막 단계들에서만 가이던스를 적용하는데, 이는 초기 단계에서의 과도한 개입이 이미지 구조 형성을 방해할 수 있기 때문이다. 실험 결과, G-VS는 사나 모델에서 기존 사후 삽입 방식보다 우수한 -96.4의 로그 오탐률 성능을 달성했으며, 이는 10의 96.4제곱분의 1이라는 극히 낮은 오탐률을 의미한다.

가이던스 강도와 적용 시점의 조절은 이미지 품질과 워터마크 강도 사이의 균형을 결정한다. 너무 강한 가이던스는 이미지 품질을 손상시키고, 너무 약한 가이던스는 워터마크 검출을 어렵게 만든다. 연구진은 다양한 실험을 통해 15단계 가이드 사용 시, 마지막 5~15단계에서 적절한 강도로 가이던스를 적용할 때 최적의 결과를 얻을 수 있음을 확인했다. 이러한 세밀한 조정 과정은 워터마킹 기술이 단순한 표시 기능을 넘어, 콘텐츠의 미적 가치를 보존하면서도 저작권 정보를 효과적으로 담아내는 기술로 발전함을 시사한다.

**[표 1] AI 이미지 생성 모델 별 G-SSig와 G-VS의 워터마크 성능 비교표**

| LDM | WM | FID (↓) | CLIP (↑) | PSNR (↑) | LPIPS (↓) | Capacity(↑) | $P_D$ @ $10^{-10}$ $P_{FA}$ | $-\log_{10}(P_{FA})$ @ $P_D = 0.9$ |
|---|---|---|---|---|---|---|---|---|
| SD2 | | 5.0 | 0.330 | | | | | |
| SD2 | G-SSig | 2.3 | 0.332 | 19.6 | 0.22 | **27.7 (+19.3)** | **0.99 (+0.5)** | **16.3 (+12.2)** |
| SD2 | G-VS | 2.2 | 0.332 | 18.5 | 0.28 | **212.2 (+37.7)** | **1.0 (+0.0)** | **105.6 (+61.8)** |
| Flux | | 9.5 | 0.271 | | | | | |
| Flux | G-SSig | 9.3 | 0.271 | 25.4 | 0.07 | **26.6 (+16.6)** | **0.99 (+0.46)** | **16.6 (+12.8)** |
| Flux | G-VS | 9.3 | 0.269 | 26.0 | 0.07 | **192.5 (+16.0)** | **1.0 (+0.0)** | **72.8 (+24.3)** |
| Sana | | 4.3 | 0.346 | | | | | |
| Sana | G-SSig | 4.2 | 0.347 | 28.6 | 0.02 | **26.5 (+17.0)** | **0.98 (+0.41)** | **15.5 (+10.6)** |
| Sana | G-VS | 4.1 | 0.346 | 23.5 | 0.07 | **207.5 (+28.8)** | **1.0 (+0.0)** | **96.4 (+49.2)** |

출처: Enoal Gesny, "Guidance Watermarking for Diffusion Models", arXiv, 2026.05.07., https://arxiv.org/pdf/2509.22126

• **이미지 품질과 검출 성능 사이의 기술적 균형**

워터마킹 기술의 실용성은 이미지 품질 유지와 검출 성능 확보라는 두 가지 상충하는 목표를 얼마나 잘 균형을 맞추느냐에 달려 있다. 현대 워터마킹 기술은 이 두 요구사항을 동시에 만족시키기 위해 다양한 접근법을 시도하고 있다.

논문의 실험 결과가 보여주는 가장 중요한 발견은 워터마크가 이미지의 시각적 품질에 미치는 영향이 생각보다 적다는 점이다. 워터마크가 삽입된 이미지와 원본 이미지를 나란히 놓고 봐도 색조나 형태의 미세한 차이만 있을 뿐, 전체적인 구도나 의미는 동일하게 유지된다. 이는 기존의 노이즈 형태 워터마크나 트리링(Tree-Ring) 방식이 이미지 구성을 크게 변경하는 것과 대조적이다. 워터마크가 이미지의 잠재 공간에 자연스럽게 녹아들어, 사람의 눈으로는 감지할 수 없지만 전용 검출기로는 명확히 식별 가능한 상태를 만드는 것이다.

오탐률 제어는 산업 적용에서 특히 중요한 의미를 지닌다. 예를 들어 하루에 백만 장의 이미지를 처리하는 플랫폼에서 0.01%의 오탐률도 매일 100장 수준의 잘못된 판정 가능성을 의미한다. 이는 인간 작가의 사진이 AI 산출물로 오인되거나, 반대로 AI 이미지가 진본으로 인증되는 문제를 야기할 수 있다. 논문에서 제시한 방법들은 이론적으로 매우 낮은 오탐률을 달성했는데, 이는 대규모 이미지 플랫폼에서도 신뢰할 수 있는 자동 검증 시스템 구축이 가능함을 시사한다.

• **모델별 성능 차이와 오탐률 최소화 전략**

워터마킹 성능은 적용되는 확산 모델의 특성에 따라 달라진다. 스테이블 디퓨전 2, 플럭스, 사나는 각각 다른 아키텍처와 학습 방식을 가지고 있어, 동일한 워터마킹 기법을 적용해도 서로 다른 결과를 보인다. 스테이블 디퓨전 2는 가장 널리 사용되는 모델로 워터마크 삽입과 검출이 안정적이며, 플럭스는 최신 기술을 적용해 이미지 품질이 우수하지만 워터마크 강도가 상대적으로 약하다. 사나는 효율적인 구조로 빠른 생성이 가능하면서도 워터마크 용량이 가장 크다. 이러한 차이는 각 모델의 노이즈 제거 과정과 잠재 공간 표현 방식의 차이에서 비롯된다.

오탐률을 최소화하기 위한 핵심 전략은 워터마크 적용 시점과 강도를 세밀하게 조절하는 것이다. 연구진은 생성 과정의 마지막 5단계, 10단계, 15단계에서 각각 워터마크를 적용하는 실험을 진행했다. 너무 이른 단계에서 워터마크를 삽입하면 이미지 형성 자체를 방해하고, 너무 늦은 단계에서 적용하면 워터마크가 충분히 반영되지 않는다. 따라서 최적 지점은 이미지의 주요 구조가 형성된 후, 세부 사항이 다듬어지는 단계다.

실제 산업 환경에서는 다양한 변형 공격에 대한 강건성도 고려해야 한다. 이미지는 배포 과정에서 압축, 크기 조정, 색상 보정 등 여러 처리를 거친다. 워터마크가 이러한 변형 후에도 검출 가능해야 실용적

가치가 있다. 본 보고서에서 소개한 방법은 JPEG 압축 같은 일반적인 변형에 대해 높은 생존율을 보였다. 특히 G-VS 방식은 다양한 공격 시나리오에서도 90% 이상의 검출률을 유지했는데, 이는 워터마크 정보가 이미지의 여러 영역에 분산 저장되어 일부가 손상되어도 복구 가능하기 때문이다. 이러한 강건성은 저작권 보호 기술이 실제 콘텐츠 유통 환경에서 신뢰할 수 있는 도구로 자리 잡는 데 필수적인 요소다.

## III. 결론: 이미지 산업의 인프라로 자리잡는 워터마킹 솔루션

**• 진본 인증과 합성 식별, 양방향 워터마킹 생태계의 가능성**

생성형 AI 시대의 워터마킹 기술은 진본 인증과 합성 식별의 양방향에서 이미지 콘텐츠의 신뢰성을 확보하는 역할을 수행하고 있다. 한편으로는 AI가 생성한 이미지에 자동으로 워터마크를 삽입해 합성 콘텐츠임을 표시하고, 다른 한편으로는 실제 촬영된 이미지에 워터마크를 삽입해 진본임을 증명한다. 이러한 접근은 결국 모든 디지털 이미지가 출처와 생성 방식을 명확히 밝히는 체계로 이어진다. dpa Picture Alliance와 iMATAG의 협업 사례는 이미 산업 현장에서 이러한 변화가 시작되었음을 보여주며, 가이던스 기반 워터마킹 기술의 발전은 AI 산출물 표시의 기술적 토대를 제공한다. 이를 통해 콘텐츠 창작자의 권리 보호와 소비자의 알 권리 보장이 동시에 가능해진다.

향후 워터마킹 기술의 발전 방향은 표준화와 상호운용성 확보에 있다. 현재는 각 기업과 연구 그룹이 독자적인 워터마킹 방식을 개발하고 있지만, 장기적으로는 C2PA(Coalition for Content Provenance and Authenticity)와 같은 산업 표준에 통합될 필요가 있다. 기술적 측면에서는 더 작은 계산 비용으로 더 강건한 워터마크를 삽입하는 방법, 다양한 미디어 형식에 일관되게 적용 가능한 범용 워터마킹 프레임워크 개발이 과제로 남아 있다. 제도적으로는 워터마크 삽입 의무화, 검증 체계 구축, 위반 시 제재 방안 등이 논의되어야 한다. 이러한 기술적, 제도적 보완이 이루어질 때, 워터마킹은 단순한 보호 기술을 넘어 디지털 이미지 경제 전체의 신뢰할 수 있는 검증 체계로 자리잡을 수 있을 것이다.

## 참고문헌

• Enoal Gesny, "Guidance Watermarking for Diffusion Models", arXiv, 2026.05.07., https://arxiv.org/pdf/2509.22126
• iMATAG, "dpa Picture Alliance Selects iMATAG as Preferred Invisible Watermarking Technology", 2026.05.06., https://www.imatag.com/blog/dpa-picture-alliance-selects-imatag-as-preferred-invisible-watermarking-technology

# C2PA와 신스ID, AI 산출물 식별 기술의 작동 원리

## 뉴스 브리프

유럽연합 AI 법의 투명성 의무 시행을 앞두고, AI 산출물을 기계 판독 가능한 형식으로 표시하는 기술의 중요성이 커지고 있다. 단순한 문구 표시만으로는 AI 생성 여부를 자동으로 판별하기 어렵기 때문에, 콘텐츠의 생성 이력과 편집 과정을 기록하거나 콘텐츠 내부에 식별 신호를 삽입하는 기술이 주요 대응 수단으로 부상하고 있다. 이러한 흐름 속에서 메타데이터 기반 출처 증명 표준인 C2PA와 구글 딥마인드의 비가시 워터마킹 기술인 신스ID가 상호 보완적 기술로 주목받고 있다. 두 기술은 작동 방식과 파일 변환 과정에서의 보존성 측면에서 서로 다른 강점과 한계를 가진다. 본 보고서는 C2PA와 신스ID의 작동 원리와 파일 변환 과정에서의 보존성 차이를 살펴보고, 메타데이터 기반 출처 관리와 신호 기반 워터마킹을 결합한 이중 검증 구조의 필요성을 검토한다.

## 뉴스 플러스

### I. 서론: AI 산출물 표시 의무와 식별 기술의 필요성

**• 유럽연합 AI 법 시행과 AI 산출물 표시 의무화**

2026년 8월부터 유럽연합 AI 법(EU AI Act)의 주요 의무가 본격 적용되며, AI 시스템이 생성하거나 조작한 이미지·오디오·비디오 및 텍스트 등 합성 콘텐츠에 대한 투명성 의무가 중요한 규제 과제로 부상하고 있다. 유럽연합 AI 법은 AI 시스템 제공자가 AI 산출물의 경우 인공지능에 의해 생성 또는 변형된 것임을 기계 판독 가능한 형식(machine-readable format)으로 표시하도록 요구한다.

이러한 규제 흐름은 단순히 “AI가 만들었습니다”라는 문구를 표시하는 수준을 넘어, 자동화된 시스템이 콘텐츠의 생성 여부를 판별할 수 있는 기술적 장치를 요구한다는 점에서 기존 자율 표시 방식과 차이가 있다. 이에 따라 미드저니(Midjourney), 구글 제미나이(Google Gemini), 어도비 파이어플라이(Adobe

Firefly) 등 주요 AI 이미지 생성 플랫폼들은 이미 산출물에 식별 정보를 삽입하는 방식을 도입하고 있다. 이러한 흐름 속에서 AI 생성 콘텐츠의 출처를 기록하고 검증하기 위한 기술로 C2PA(Coalition for Content Provenance and Authenticity) 표준과 구글 딥마인드(DeepMind)의 비가시 워터마킹 기술인 신스ID(SynthID)가 주목받고 있다.

**• 파일 변환 과정에서 발생하는 출처 정보 손실 문제**

사용자들은 이미지를 소셜미디어에 업로드하거나 공유하기 위해 파일 형식을 변환하고, 용량을 줄이기 위해 압축하며, 화면 크기에 맞추기 위해 리사이징한다. 이러한 과정은 일상적인 파일 처리로 인식되지만, 실제로는 픽셀 값이 다시 계산되고 생성일, 수정일, 파일 크기 및 해상도 등 메타데이터가 변경되는 작업이다. 문제는 이 과정에서 AI 산출물임을 식별하는 데 필요한 출처 정보나 식별 신호가 의도치 않게 손실되거나 약화될 수 있다는 점이다.

이러한 문제는 AI 산출물 표시 기술이 생성 단계에서 삽입되는 것만으로는 충분하지 않고, 실제 유통·변환 과정에서 얼마나 유지되는지도 함께 검토해야 한다는 점을 보여준다. 이에 따라 본문에서는 C2PA와 신스ID의 작동 방식과 파일 변환 과정에서의 보존성 차이를 중심으로 AI 산출물 식별 기술을 살펴본다.

## II. 본론: C2PA와 신스ID의 기술적 메커니즘과 한계

**• C2PA의 메타데이터 기반 출처 검증 방식**

C2PA는 디지털 콘텐츠의 출처와 변경 이력을 기록하고 검증하기 위한 표준이다. 이미지·영상·문서 등 디지털 콘텐츠가 어떤 도구로 생성되었는지, 이후 어떤 편집이나 변환을 거쳤는지, 해당 정보가 중간에 조작되지 않았는지를 확인할 수 있도록 하는 것이 핵심이다. 이를 위해 C2PA는 콘텐츠에 매니페스트(manifest)라는 구조화된 정보를 결합한다.

C2PA 매니페스트에는 콘텐츠 생성 도구, 편집 작업, 콘텐츠 바인딩 정보, 디지털 서명 등이 포함될 수 있으며, 이를 통해 콘텐츠의 출처와 변경 이력을 검증할 수 있다. 쉽게 말해 C2PA는 콘텐츠에 "이 파일이 어떻게 만들어지고 수정되었는지"를 설명하는 검증 가능한 이력 정보를 붙이는 방식이다.

예를 들어 AI 이미지 생성 도구로 이미지를 만든 뒤 편집 프로그램에서 색상 보정이나 리사이징을 수행했다면, C2PA 매니페스트에는 생성, 색상 조정, 크기 조정, 변환 등 주요 작업이 작업 이력 정보 형태로 기록될 수 있다. C2PA 명세는 생성 작업을 나타내는 c2pa.created, 색상 조정을 나타내는 c2pa.adjustedColor, 리사이징을 나타내는 c2pa.resized, 인코딩 변환을 나타내는 c2pa.transcoded 등의 작업 이력 정보를 제시한다. 다만 작업 이력 정보만으로 실제 작업 순서를 항상 완전하게 확인할 수 있는 것은 아니다. 따라서 C2PA는 '시간 순서대로 모든 과정을 기록하는 기술'로 보기 어렵다. 그보다는 콘텐츠에 수행된 주요 작업과 출처 정보를 검증 가능한 방식으로 기록하는 기술로 설명하는 것이 적절하다.

C2PA의 강점은 출처 정보를 상세히 제시하고 검증할 수 있다는 점이다. C2PA는 단순히 AI 생성 여부만 표시하는 것이 아니라, 콘텐츠가 어떤 도구를 거쳐 생성·수정되었는지에 대한 변경 이력을 제공할 수 있다. 또한 디지털 서명과 콘텐츠 바인딩(content binding)*을 통해 매니페스트가 해당 콘텐츠에 붙은 정보가 맞는지, 중간에 조작되지 않았는지 확인할 수 있다.

* 콘텐츠 바인딩(content binding): 매니페스트와 콘텐츠가 실제로 일치하는지, 그리고 콘텐츠가 변조되지 않았는지 암호학적 해시값을 통해 검증하기 위한 메커니즘

### • 신스ID의 비가시 워터마킹 방식

신스ID는 이미지 콘텐츠 자체에 비가시 워터마크 신호를 삽입하는 방식이다. 이미지 품질에 거의 영향을 주지 않는 범위에서 픽셀 값의 분포를 미세하게 조정하고, 이후 전용 검증 알고리즘이 해당 패턴을 감지하도록 설계된다. 워터마크가 콘텐츠 내부에 포함되기 때문에, 메타데이터를 제거하더라도 쉽게 사라지지 않는다는 특징이 있다.

구글 딥마인드는 신스ID를 생성 모델 안에 직접 넣는 방식이 아니라, 생성이 끝난 이미지 위에 별도로 워터마크를 입히는 방식으로 설계했다. 따라서 특정 생성 모델에 종속되지 않고, 여러 생성 모델의 이미지에 적용할 수 있다. 이러한 구조는 다양한 서비스에 적용하기 쉽다는 장점이 있지만, 워터마크가 이미지 품질을 해치지 않으면서도 압축이나 편집 이후까지 남아 있어야 한다는 과제를 함께 가진다.

### • 파일 변환 과정의 출처 정보 보존 한계

파일 변환 과정에서 C2PA와 신스ID는 서로 다른 방식으로 영향을 받는다. C2PA 매니페스트는 파일 내부의 메타데이터 영역에 저장되므로, 변환 도구가 이를 읽고 새로운 파일 형식에 다시 삽입하도록 설계되어 있어야 보존된다. 그러나 일반적인 이미지 변환 도구는 픽셀 데이터만 처리하고, 메타데이터는 생략하거나 일부 표준 메타데이터만 복사하는 경우가 많다. 특히 온라인에서 파일을 업로드해 변환하는 웹 기반 변환 도구나, 프로그램 명령어로 이미지 형식·크기·압축률을 조정하는 명령행 변환 도구는 C2PA 매니페스트를 인식하지 못해 변환 과정에서 출처 정보를 누락시키는 것으로 보고되며, 이 경우 생성 모델, 편집 이력, 서명 정보 등 C2PA 기반 출처 정보가 손실될 수 있다.

신스ID와 같은 신호 기반 워터마킹은 픽셀 데이터에 직접 삽입되기 때문에, 파일을 열고 새로운 형식으로 다시 저장하는 과정에서 워터마크 신호도 이미지 콘텐츠와 함께 처리된다. 무손실 변환처럼 픽셀 값이 그대로 유지되는 경우에는 신호 보존 가능성이 높고, 일반적인 수준의 손실 압축에서도 일정 수준의 감지 가능성이 유지될 수 있다. 다만 극단적인 저품질 압축, 과도한 리사이징, 강한 크롭(Crop) 등은 신호 강도를 낮춰 검출 성능을 떨어뜨릴 수 있다.

### • C2PA와 신스ID의 이중 검증 구조

C2PA와 신스ID는 AI 산출물 식별 과정에서 서로 다른 역할을 수행할 수 있다. C2PA는 콘텐츠의 생성 도구, 생성 시점, 편집 작업, 서명 정보 등 상세한 출처 이력을 확인하는 데 활용된다. 반면 신스ID는

이러한 상세 이력을 제공하기는 어렵지만, 메타데이터가 제거된 이후에도 콘텐츠 내부 신호를 통해 AI 산출물 여부를 확인하는 보완 수단으로 활용될 수 있다.

이중 검증 구조에서는 먼저 업로드된 콘텐츠에 C2PA 매니페스트가 포함되어 있는지 확인할 수 있다. 매니페스트가 존재하고 서명이 유효하다면 생성 도구와 편집 이력 등 상세 정보를 검증할 수 있다. 반대로 매니페스트가 없거나 손상된 경우에는 신스ID를 비롯한 신호 기반 워터마크 검출을 추가로 수행해 해당 콘텐츠가 AI 산출물인지 여부를 확인할 수 있다. 이처럼 C2PA는 출처 정보를 투명하게 증명하는 역할을, 신스ID는 출처 정보가 손실된 상황에서 식별 가능성을 보완하는 역할을 맡는다.

다만 이중 검증 구조도 완전한 해결책은 아니다. C2PA 매니페스트가 유통 과정에서 제거되면 상세한 출처 이력은 유지되기 어렵고, 신스ID 역시 극단적 압축이나 의도적 제거 공격을 거치면 검출 성능이 저하될 수 있다. 따라서 AI 산출물 식별 체계는 복수의 검증 기술을 병행하는 동시에, 파일 변환 도구와 플랫폼 차원에서 출처 정보 보존 기능을 함께 강화할 필요가 있다.

## III. 결론 및 전망: AI 산출물 식별 기술의 향후 방향

**• 출처 정보 보존을 위한 기술·플랫폼 연계 필요**

C2PA와 신스ID로 대표되는 AI 산출물 식별 기술은 유럽연합 AI 법의 투명성 의무와 맞물려 중요성이 커지고 있다. 다만 실제 유통 환경에서는 다양한 처리 과정에서 출처 정보나 식별 신호가 손실되거나 약화될 수 있다. 따라서 향후 AI 산출물 추적 체계는 단일 기술에 의존하기보다, 메타데이터 기반 출처 관리와 신호 기반 워터마킹을 결합한 구조로 설계될 필요가 있다.

이를 위해서는 기술적 보완과 산업적 협력이 함께 이루어져야 한다. 두 기술의 결합은 AI 산출물 식별의 정확성을 높이는 수단이 될 수 있지만, 생성 도구, 편집 소프트웨어, 플랫폼, 변환 도구 전반의 연계가 함께 이루어져야 실효성을 가질 수 있다.

기술적 측면에서는 변환 도구와 플랫폼이 C2PA 매니페스트를 안정적으로 보존할 수 있도록 지원해야 하며, 신호 기반 워터마킹은 일반적인 유통 환경에서도 감지 가능성을 유지할 수 있도록 강건성을 개선해야 한다. 산업적 측면에서는 생성·편집 도구와 소셜미디어 플랫폼, 검색 엔진 등이 출처 정보 보존과 검증 절차를 일관되게 지원할 필요가 있다.

결국 AI 산출물 식별 체계의 핵심은 단일 워터마킹 기술의 완성도가 아니라, 다양한 출처 증명 수단을 결합해 실제 유통 환경에서 식별 가능성을 유지하는 데 있다. C2PA는 콘텐츠의 생성·편집 맥락을 설명하는 역할을 하고, 신스ID는 메타데이터가 사라진 상황에서도 AI 생성 여부를 확인할 수 있는 보완 장치로 기능할 수 있다. 이러한 접근은 AI 기술의 활용을 제한하기보다, 생성 콘텐츠의 투명성과 책임성을 확보하는 실무적 기반으로 활용될 수 있을 것으로 보인다.

## 참고문헌


- Fileza Team, "C2PA and SynthID: How AI Watermarking Changes What Happens When You Convert a File", fileza, 2026.04.18., https://fileza.io/articles/ai-watermarking-c2pa-synthid-file-conversion#google_vignette
- Sven Gowal 외 25명, "SynthID-Image: Image watermarking at internet scale", arxiv, 2025.10.10., https://arxiv.org/abs/2510.09263
- C2PA, "Content Credentials : C2PA Technical Specification", 2026.05.19. 접속 기준, https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html


## 기술용어

| 순번 | 용어 | 설명 |
|---|---|---|
| 1 | **콘텐츠 바인딩 (content binding)** | 매니페스트와 콘텐츠가 실제로 일치하는지, 그리고 콘텐츠가 변조되지 않았는지를 암호학적 해시값을 통해 검증하기 위한 메커니즘 |