논문 2026-1-4 http://dx.doi.org/10.29056/jsf.2026.03.04

온디바이스 AI 환경의 하드웨어 기반 콘텐츠 출처 인증 프레임워크 연구

신수홍*, 고병수**†

A Study on Hardware-Based Content Provenance Authentication Framework for On-Device AI

Su-Hong Shin*, Byoung-Soo Koh**†

요 약

오늘날 생성형 AI와 온디바이스 AI의 보편화로 인해 스마트폰, 스마트 글래스, IoT 기기 등 엣지 디바이스에서 콘텐츠를 직접 생성·캡처·편집하는 환경이 확산되고 있다. 그러나 기존 소프트웨어 기반 보안 방식은 OS 수준의 우회 가능성, 제한된 자원에서의 암호 연산 부담, 오프라인 인증 공백 등 구조적 한계로 인해 온디바이스 환경의 출처 인증 요구를 충족하지 못한다. 본 연구에서는 칩 수준의 콘텐츠 출처 인증을 지향하는 하드웨어 기반 프레임워크인 HW-CPA(Hardware-based Content Provenance Authentication)를 제안한다. HW-CPA는 NPU 기반 콘텐츠 유형 분류, TEE 기반 C2PA 매니페스트 생성, Hardware Root of Trust 기반 디바이스 신원 보증, 오프라인 해시 체인 기반 지연 검증의 4계층으로 구성된다. 위협 모델링과 상용 기술 매핑을 통해 설계 타당성을 논증하고, 소프트웨어 기능 모사 프로토타입을 통해 평가한 결과, 변조·조작에 대한 일관된 검출과 소비자급 환경에서 수용 가능한 서명·검증 지연 시간을 확인하였다. 본 연구는 향후 SoC 통합 연구의 기초로서 온디바이스 AI 시대의 콘텐츠 신뢰성 확보에 기여 할 수 있을 것으로 기대한다.

Abstract

The proliferation of generative AI and on-device AI has enabled edge devices such as smartphones, smart glasses, and IoT devices to generate, capture, and edit content locally. However, software-based security mechanisms face structural limitations including OS-level bypass risks, cryptographic overhead under constrained resources, and offline authentication gaps. This study proposes HW-CPA (Hardware-based Content Provenance Authentication), a hardware-oriented framework comprising four layers: NPU-based content classification, TEE-based C2PA manifest generation, Hardware Root of Trust-based device identity assurance, and offline hash chain-based deferred verification. Threat modeling, commercial technology mapping, and a software emulation prototype confirm consistent tamper detection and acceptable signing/verification latency for consumer-grade environments. This work provides a foundation for future SoC-level integration toward trustworthy generative AI content.

한글키워드 : 하드웨어 보안, 콘텐츠 출처 인증, 생성형 AI, 온디바이스 AI, C2PA

keywords : Hardware Security, Content Provenance Authentication, Generative AI, On-device AI, C2PA

* ㈜디지캡 비즈니스솔루션사업실 모빌리티BE개발팀
** 한신대학교 AI.SW대학
† 교신저자: 고병수(email: bskoh@hs.ac.kr)

접수일자: 2026.03.11. 심사완료: 2026.03.18.
게재확정: 2026.03.20.

1. 서 론

1.1 필요성

오늘날 디지털 미디어 생태계는 생성형 AI 기술의 발달로 콘텐츠의 신뢰성 위기에 직면해 있다. 과거에는 전문적인 기술이 필요했던 이미지 합성과 조작이 텍스트 프롬프트 몇 줄로 가능한 시대가 도래했기 때문이다. 2025년 기준 탐지된 딥페이크 사례는 약 800만 건에 달하며, 이는 2023년 대비 약 900%나 증가한 수치이다[1].

여기에 온디바이스 AI의 성장은 문제를 한층 더 복잡하게 만들고 있다. NPU(Neural Processing Unit)의 발전으로 스마트폰, 스마트 글래스, IoT 기기 같은 엣지 디바이스에서 생성형 AI를 직접 구동하는 것이 가능해졌기 때문이다. 이러한 이유로 2025년에는 NPU 라이선스 계약이 폭발적으로 증가하면서 다양한 산업군에서 온디바이스 AI가 주요 전략으로 자리잡았다[3].

문제는 디바이스 영역에서 생성되거나 촬영된 콘텐츠의 진위성을 보장할 메커니즘이 부재하다는 점이다. 소프트웨어 기반 보안은 루트 권한 획득 시 우회되고, 제한된 시스템 자원 환경에서 암호 연산이 성능 저하를 일으키며, 또한 오프라인 상태에서는 클라우드 및 외부 인증기관(CA) 등에 접근이 불가능하다. Synopsys(2026)가 지적한 것처럼, AI 시스템의 보안은 소프트웨어만으로는 부족하며 SoC 설계 단계부터 하드웨어에 내장해야 한다고 말하고 있다[4].

1.2 연구 목표

본 논문은 온디바이스 AI 환경에서 콘텐츠 출처 인증, 생성형 AI 콘텐츠 식별, 오프라인 무결성 보증을 칩 수준에서 수행할 수 있는 하드웨어 지향 프레임워크를 설계하고, 핵심 절차의 기능적 타당성을 평가하는 것을 목표로 한다.

본 논문에서 말하는 핵심 연구에 대한 질문은 아래와 같다.

- RQ1: HW-CPA의 Layer 2~4 핵심 절차는 소프트웨어 기능 모사 환경에서도 변조 및 조작을 탐지할 수 있는가

- RQ2: 오프라인 환경에서 해시 체인 및 모노토닉 카운터 구조(Monotonic Counter, 값이 증가만하고 되돌릴 수 없는 구조)는 삭제·삽입·순서 변경에 대해 순서 무결성을 보장하는가

- RQ3: 서명 기반 provenance manifest 생성·검증 오버헤드는 소비자급 환경에서 허용 가능한 수준인가

이를 위해 STRIDE 기반 위협 모델링, C2PA·TEE·NPU·HRoT를 결합한 4계층 설계, 기존 상용 기술과의 매핑, 소프트웨어 기능 모사 프로토타입 평가를 수행한다. 본 논문은 실제 SoC 구현이 아닌, 하드웨어 통합을 지향하는 설계와 기능 모사를 통해 후속 연구 방향을 제시하는 데 초점을 두었다.

2. 관련 연구

2.1 하드웨어 기반 AI 보안 메커니즘

최근 하드웨어 아키텍처에 보안 기능을 직접 내장하려는 연구가 활발히 진행되고 있다. 사실 이는 이전에도 다양한 분야에서도 많은 연구가 진행되었고, 또 실제 제품도 많이 출시가 되곤 했었다. 다만 최근 AI 환경이 각광을 받으면서 더욱 더 하드웨어 기반 보안 칩에 대한 요구사항이 갈수록 커지고 있다. 한 예로 CNAS의 Aarne 등(2024)은 'on-chip governance' 개념을 도입하여 TEE를 활용해 AI 칩의 무단 사용을 방지하고 추론 무결성을 보증하는 방안을 제안하기도 하였다 [5]. UK ARIA의 flexHEG 프로젝트(2025)는 AI 칩의 데이터 경로에 검증 전용 보증

프로세서를 물리적으로 연동시켜, 메인 OS를 신뢰하지 않고도 연산 무결성을 담보할 수 있는 아키텍처를 발표하였으며[6], Reuel 등(2025)은 이러한 하드웨어 기반 검증 메커니즘이 책임 있는 AI 개발의 기술적 기반이 될 수 있음을 논의한 바도 있다[16]. 이러한 동향은 보안의 초점이 애플리케이션 계층에서 하드웨어 계층으로 활발히 확장되고 있음을 보여준다.

2.2 C2PA 표준과 온디바이스 AI

C2PA(Coalition for Content Provenance and Authenticity)는 Adobe, Microsoft, Intel 등이 주도하여 2021년 결성된 디지털 콘텐츠 출처 인증 표준 기구로, 콘텐츠의 출처와 변경 이력을 암호학적으로 서명된 매니페스트에 기록하여 무결성을 보장한다[7]. CISA와 NSA도 Content Credentials의 활용을 권고하였다[15]. C2PA의 하드웨어 구현이 빠르게 확산되고 있는데, Leica M11-P(2023)는 전용 보안 칩으로 C2PA 서명을 부여하는 최초의 소비자 카메라이고[8], Google Pixel 10(2025)은 Tensor G5와 Titan M2 보안 칩을 조합하여 C2PA Assurance Level 2를 획득하며 오프라인에서도 온디바이스 신뢰 타임스탬프를 하드웨어로 구현하였다[9].

온디바이스 AI 측면에서는 Google의 Coral NPU가 RISC-V ISA 기반 오픈소스 플랫폼으로 CHERI 기술을 활용한 하드웨어 메모리 격리를 제공하며[2], MediaTek Kompanio Ultra NPU는 수십 TOPS 성능으로 생성형 모델의 온디바이스 구동을 지원한다[10]. 한편 스마트 글래스의 경우 사용자가 인지하지 못하는 상황에서도 데이터를 수집·처리할 수 있어 보안 과제가 특히 첨예한데[11], 수집 데이터가 서드파티 AI 서비스로 전송되는 문제가 지적되고 있다[12]. Gallardo 등(2023)은 AR 글래스의 데이터 수집에 대한 프라이버시 우려를 분석하였으며[13], TEE 기반 보안 칩을 통한 키 보호가 권고되고 있으나[14] 콘텐츠 출처 인증까지 확장한 통합 프레임워크는 아직 제안된 바 없다.

2.3 기존 연구와의 차별성

기존 연구들은 클라우드나 범용 모바일 기기의 소프트웨어적 C2PA 구현에 집중하였고, flexHEG 같은 하드웨어 보안 연구도 AI 연산 검증에 그쳤을 뿐 결과물을 C2PA와 결합하는 프레임워크로 발전하지 못했다. 본 논문이 제안하는 HW-CPA는 (1) 센서 및 NPU 데이터 경로에 하드와이어드(Hardwired) 분류 로직을 삽입하여 생성 단계부터 소프트웨어 개입을 차단하는 구조를 설계하고, (2) 하드웨어 타이머와 해시 체인을 결합한 오프라인 지연 검증 프로토콜을 통합 설계하였다는 점에서 차별성이 있음을 제안한다.

그리하여 기존 연구들의 한계를 종합하면, 하드웨어 기반 AI 보안 연구(CNAS, flexHEG, Reuel 등)는 AI 칩의 무단 사용 방지나 연산 무결성 검증에 초점을 맞추고 있으나, 출력 콘텐츠에 대한 출처 인증 표준(C2PA)과의 결합으로 확장되지 않았다. C2PA 하드웨어 구현(Leica M11-P, Pixel 10)은 특정 디바이스에 한정된 사례이며, 온디바이스 생성형 AI 출력물에 대한 하드웨어 수준의 출처 표시와 오프라인 검증까지 포괄하는 체계적 프레임워크는 부재하다. CISA/NSA는 Content Credentials 도입을 권고하고 있으나 구체적인 하드웨어 아키텍처를 제시하지는 않았다. 아래 표 1은 기존 접근 방식과 본 연구의 구조적 차별성을 비교하여 보여준다.

표 1. 기존 콘텐츠 출처 인증 접근 방식과 HW-CPA의 비교
Table 1. Comparison of Existing Content Provenance Approaches and HW-CPA

비교항목	SW C2PA	Pixel 10	flexHEG	HW-CPA(제안)
NPU 통합 분류	X	X	X	설계
TEE 내 서명	X	Titan M2	X	기능 모사 검증
오프라인 검증	X	O	X	기능 모사 검증
디바이스 신원	X	StrongBox	X	기능 모사 검증
AI 출력 태깅	X	부분	X	설계
End-to-end 통합	X	카메라 한정	연산 검증	4계층 설계

연구가 제안하는 HW-CPA의 학술적 기여는 콘텐츠 생성 시점의 유형 분류에서부터 오프라인 사후 검증에 이르는 end-to-end 신뢰 사슬을 단일 SoC 수준 아키텍처로 통합 설계하였다는 점에 있다. NPU 데이터 경로 내 콘텐츠 분류, C2PA 매니페스트의 TEE 내 완전 생성, 오프라인 해시 체인 기반 지연 검증이라는 세 요소의 결합은 기존 연구에서 다루어지지 않은 구성이다.

3. HW-CPA 프레임워크 설계

3.1 위협 모델 및 설계 원칙

온디바이스 AI 환경에서 콘텐츠 출처 인증이 직면하는 위협을 STRIDE 프레임워크 기반으로 정의한다. T1(출처 위조)은 AI 생성 콘텐츠를 실촬영으로 위장하는 것, T2(콘텐츠 변조)는 서명 이후 무결성을 훼손하는 것, T3(부인)은 생성 사실을 부정하는 것, T4(정보 유출)는 서명 키의 외부 유출, T5(저작권 침해)는 타인 저작물의 무단 캡처·재생성, T6(오프라인 위조)는 네트워크 단절 시 위조 콘텐츠 생성이다.

소프트웨어 기반 보안은 세 가지 구조적 한계를 갖고 있다. 첫째는 신뢰 경계의 부재이다. 운영체제의 루트 권한이 탈취되면 서명 키를 메모리 덤프로 추출하거나 서명 함수를 훅킹(Hooking)하여 위조 매니페스트를 발행할 수 있다. 흔히 말하는 '신뢰의 근원(RoT, Root of Trust)'이 소프트웨어 스택에 있는 한 이 문제는 근본적으로 해결되기가 어렵다. 둘째는 실시간 처리의 병목 현상이다. NPU가 이미지를 생성할 때마다 소프트웨어가 SHA-256 해시와 ECDSA 서명을 수행하면 메인 CPU 자원이 과도하게 소모되어 추론 프레임 드랍과 배터리 급격 소진을 유발한다. 특히 단순 이미지가 아닌 실시간 영상에서 더욱더 해당 문제가 대두된다. 셋째는 오프라인 인증의 불가이다. 예를 들면 C2PA 규격은 온라인 TSA(Time Stamp Authority)의 타임스탬프를 요구하므로, 클라우드 연결이 끊기면 재난 현장이나 오지에서 생성된 콘텐츠의 법적 증거 능력이 상실되는 인증의 공백이 발생하게 된다.

3.2 HW-CPA 4계층 구조

HW-CPA(Hardware-based Content Provenance Authentication)는 온디바이스 SoC 내에 콘텐츠 출처 인증을 하드웨어 수준으로 통합하는 것을 목표로 하는 4계층 프레임워크이다. 아래 그림 1은 제안하는 HW-CPA 프레임워크의 전체 구성도를 나타낸다.

그림 1에서 확인할 수 있듯이 센서 입력과 온디바이스 생성형 AI로부터 유입된 콘텐츠는 Layer1에서 분류된 후, 보안 채널을 통해 Layer 2~4를 순차적으로 거쳐 출처 인증이 완성된다. 각 계층의 구성요소와 핵심 기능 그리고 대응하

는 위협 유형은 아래 표 2에 정리하였다.

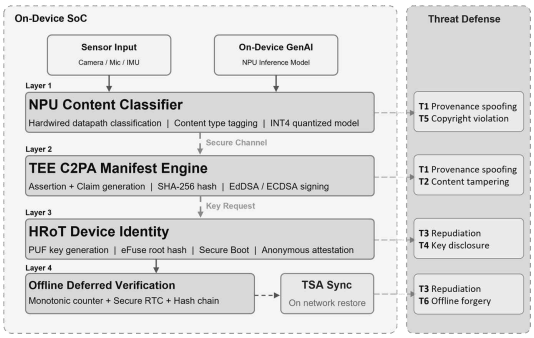

그림 1. HW-CPA 4계층 아키텍처 및 위협 방어 매핑

Fig. 1. HW-CPA four-layer architecture and threat defense mapping

표 2. HW-CPA 4계층 프레임워크 구성

Table 2. HW-CPA Four-Layer Framework

계층	구성요소	핵심기능	대응 위협
L1	NPU 콘텐츠 분류기	AI/센서 분류 및 태깅	T1, T5
L2	TEE C2PA	출처 서명 및 해시 결합	T1, T2
L3	HRot 신원 모듈	키 생성, 저장, 증명	T3, T4
L4	오프라인 검증	해시 체인, 타임 스탬프	T3, T6

Layer 1(NPU 내장형 AI 콘텐츠 분류)은 NPU 추론 파이프라인에 콘텐츠 분류 모듈을 직접 통합하는 설계를 나타낸다. 콘텐츠 생성 시점에서 캡처 원본, AI 생성 합성, AI 편집 혼합 데이터를 실시간으로 분류한다. 분류 판단이 NPU 데이터 경로(Datapath) 내부에서 이루어지므로 소프트웨어 수준에서 분류 결과를 조작하기 어려워진다. 분류 모델은 INT4 양자화된 경량 네트워크를 사용하며, 결과는 보안 채널을 통해 Layer 2의 TEE로 직접 전달된다. 다만 Layer 1은 실제 NPU datapath 통합이 요구되므로 본 연구에서는 설계 원칙만을 제시한다. 적대적 입력(adversarial example)에 대한 강건성과 관련하여, Layer 1의 분류는 콘텐츠의 의미적 특징이 아닌 물리적 데이터 경로에 주로 의존하므로 소프트웨어 수준의 적대적 공격 벡터와는 성격이 다르나, 양자화된 분류 모델 자체에 대한 적대적 강건성 평가 및 분류 정확도·오탐률 측정은 향후 과제로 남긴다.

Layer 2(TEE 기반 C2PA 매니페스트 생성)는 TEE(Trusted Execution Environment) 내부에서 C2PA 규격의 Content Credentials를 생성·서명한다. TEE는 ARM TrustZone이나 RISC-V PMP를 통해 구현되며, 메인 OS가 손상되더라도 내부 코드와 데이터의 기밀성을 보장하는 격리 실행 환경이다[17][18]. 매니페스트 생성은 4단계로, Layer 1으로부터 콘텐츠 유형 태그와 해시를 수신하고, 어서션(Assertion) 세트를 구성한 뒤, 클레임(Claim)을 생성하여 콘텐츠 해시와 하드바인딩을 수행하고, Layer 3의 HRoT 개인 키로 디지털 서명을 생성한다. TEE 내부에는 전용 암호화 가속기를 배치하여 SHA-256 해싱과 EdDSA/ECDSA 서명을 하드웨어로 처리함으로써 CPU 점유를 최소화하는 것이 목표이다.

Layer 3(HRoT 기반 디바이스 신원 인증)은 디바이스 자체의 진정성과 고유성을 보증하는 하드웨어 앵커(Hardware Trust Anchor)이다. PUF(Physically Unclonable Function)는 반도체 제조 과정의 미세한 물리적 변이를 이용하여 칩마다 고유한 디지털 지문을 생성하며, eFuse에 인증서 체인의 루트 해시를 영구 기록하고, Secure Boot Chain으로 변조된 펌웨어의 실행을 방지한다. C2PA 인증서 체계와의 통합을 위해 익명 키 증명(Anonymous Key Attestation)을 도입하여, 사용자의 프라이버시를 보호하면서 해당 매니페스트가 진본 하드웨어 기기에서 생성되었

음을 보증할 수 있다.

Layer 4(오프라인 지연 검증 프로토콜)는 네트워크가 끊긴 환경에서도 시간적 무결성을 보장하기 위한 계층이다. 세 가지 메커니즘을 통합하는데, 첫째 하드웨어 모노토닉 카운터(Hardware Monotonic Counter)는 소프트웨어적으로 되돌릴 수 없는 단조 증가 카운터로 콘텐츠 생성의 상대적 순서를 보증하고, 둘째 보안 실시간 시계(Secure RTC)는 TEE 내부에 독립 전원으로 구동되어 메인 시스템 시계와 독립적인 시간 기준을 유지하며, 셋째 해시 체인 기반 순서 증명은 오프라인 기간의 모든 매니페스트를 해시 체인으로 연결하여 삭제·삽입·순서 변경을 검출 가능하게 한다. 네트워크 복구 시 TSA 동기화로 오프라인 기간 전체의 시간적 무결성이 사후 보증된다. 한편, 하드웨어 보안 아키텍처 자체에 대한 잠재적 공격 가능성도 고려할 필요가 있다. TEE에 대해서는 캐시 기반 부채널 공격(cache-based side-channel attack)이 알려져 있다. Muñoz 등(2023)의 조사에 따르면 ARM TrustZone 기반 TEE는 타이밍 부채널, 캐시 부채널 등의 공격에 노출될 수 있으며[20], 이에 대해 상수 시간 암호 구현(constant-time implementation), 캐시 파티셔닝 등의 대응 기법이 연구되고 있다. HW-CPA에서는 TEE 내부의 전용 암호화 가속기가 메인 캐시와 분리된 경로에서 동작하도록 설계함으로써 부채널 위험을 저감하는 것을 목표로 한다. 또한 Layer 1의 센서 스푸핑 공격, 즉 센서 입력 자체를 물리적으로 위조하는 공격에 대해서는 현재 설계만으로 완전한 방어가 어려우며, 센서 모듈 인증 체계 도입이나 이중 센서 교차 검증 등 추가적인 방어 계층이 향후 연구에서 다루어져야 할 과제이다.

4. 구현 및 평가

4.1 프로토타입 구현

HW-CPA의 각 계층은 기존에 검증된 상용 기술들의 체계적 통합으로 설계되었다. Layer 1은 Google Coral NPU의 RISC-V 오픈소스 아키텍처와 CHERI 기술, Layer 2는 Google Pixel 10의 Titan M2 기반 C2PA 구현, Layer 3의 PUF/eFuse/Secure Boot는 Apple Secure Enclave과 Android StrongBox, Layer 4의 모노토닉 카운터와 Secure RTC는 TPM 규격에 이미 포함된 기능이다. 다시 말해, 각 계층에 대응하는 기술은 독립적으로 이미 산업에서 활용되고 있으며, 본 논문의 기여는 이들을 하나의 출처 인증 파이프라인으로 통합 설계한 데 있다.

실제 ASIC이나 FPGA 구현에 앞서 HW-CPA의 핵심 절차를 검증하기 위해, 소프트웨어 기능 모사 프로토타입을 구현하였다. 평가 환경은 아래 표 3과 같다.

표 3. 프로토타입 평가 환경
Table 3. Prototype Evaluation Environment

항목	사양
Platform	Apple Silicon M1Pro(10Core)
OS	MacOS
Language	Python 3.12.4
Hash	SHA-256
Signature	Ed25519(cryptography Lib.)

프로토타입은 Layer 2의 provenance manifest 생성·서명, Layer 3의 디바이스 결속형 키 사용, Layer 4의 해시 체인 기반 오프라인 순서 검증을 구현하였으며, Layer 1은 소프트웨어 라벨 입력으로 대체하였다. 본 프로토타입은 C2PA의 완전한 JUMBF 임베딩 대신, 콘텐츠 해시·유형·디바이스 식별자·타임스탬프·모노토닉 카운터·이전 체인 루트를 포함하는 정규화 된 manifest를 생성하고 디지털 서명을 부여하는 방식으로 동작한다.

평가는 4가지 시나리오로 구성하였다. (1) 정상 시나리오에서는 서명 및 해시 검증이 정상적으로 수행되는지 확인하고, (2) 콘텐츠 변조 시나리오에서는 서명 이후 파일 바이트를 수정하여 해시 불일치가 탐지되는지 측정하였다. (3) 매니페스트 조작 시나리오에서는 콘텐츠 유형 라벨, 타임스탬프, 카운터 등 주요 필드를 변경하여 서명 검증 실패 여부를 확인하였으며, (4) 오프라인 연속 생성 시나리오에서는 해시 체인 구성 후 중간 항목 삭제, 비정상 삽입, 순서 변경에 대한 검출 가능성을 분석하였다. 성능 평가는 64KB, 256KB, 1MB, 5MB, 20MB 파일을 각 30회 반복하여 mean, std, p50, p95를 산출하였다.

4.2 실험 결과

정상 검증 시나리오에서 모든 실험 파일(50건)은 서명 및 해시 검증에 성공하여 100.0%의 성공률을 보였다. 콘텐츠 변조 시나리오에서도 바이트 수정이 전수 탐지되어 100.0%를 기록하였고, 매니페스트 조작 시나리오에서도 콘텐츠 유형, 타임스탬프, 카운터, 디바이스 ID, 체인 루트 등 모든 필드의 변경이 서명 불일치로 검출되어 100.0%의 검출률을 보였다. 이는 RQ1에 대한 기능적 타당성을 뒷받침하는 결과이다.

오프라인 해시 체인 검증에서는 20건의 연속 생성 매니페스트에 대해 중간 항목 삭제, 비연속 카운터 삽입, 항목 순서 변경이 모두 탐지되었다. 이는 RQ2에 대해 제안한 해시 체인 및 모노토닉 카운터 구조가 네트워크 단절 환경에서도 순서 무결성을 보증할 수 있음을 보여준다.

RQ3에 대해, 표 4에서 확인할 수 있듯이 1MB 파일 기준 평균 서명 지연은 0.583ms, 검증 지연은 0.686ms로 측정되었으며, 최대 크기인 20MB에서도 p95 지연은 9.633ms를 초과하지 않았다. 이는 소비자급 온디바이스 환경에서도 과도한 부하 없이 provenance manifest 생성·검증이 가능함을 보여준다. 다만 본 평가 환경은 Apple Silicon의 SHA-256 하드웨어 가속이 반영된 결과이므로, 실제 모바일 AP 환경에서는 클럭 주파수와 메모리 대역폭의 차이로 인해 지연 시간이 다소 상이 할 수 있다. 또한 이번 결과는 실제 TEE나 PUF, eFuse를 직접 사용한 하드웨어 실측값이 아닌, 해당 절차를 소프트웨어로 기능 모사한 상대적 평가 결과로 해석되어야 한다.

표 4. 파일 크기별 서명 및 검증 오버헤드(ms)
Table 4. Signing/Verification Overhead by File Size(ms)

크기	평균 서명	평균 검증	p95 서명	p95 검증
64KB	0.133	0.242	0.171	0.303
256KB	0.221	0.328	0.282	0.385
1MB	0.583	0.686	0.657	0.748
5MB	2.373	2.419	2.558	2.616
20MB	9.236	9.261	9.627	9.633

적용 시나리오 측면에서, 분쟁 현장의 저널리스트가 HW-CPA 탑재 스마트 글래스로 촬영하면 센서 원본 분류와 오프라인 해시 체인으로 콘텐츠 신뢰성을 확보할 수 있고, 온디바이스 생성형 AI로 만든 이미지는 NPU 하드와이어드 로직이 'AI 생성'으로 강제 태깅하여 EU AI Act 같은 글로벌 규제를 기술적으로 충족할 수 있다. 그리고 재난 및 오지 환경에서 Layer 4의 오프라인 검증 프로토콜이 시간적 무결성을 보증하여, 촬영된 콘텐츠의 법적 증거 능력을 강화하는 데 기여 할 수 있을 것으로 보인다.

4.3 하드웨어 환경 적용 시 예상 분석

본 논문에서 제안하는 프로토타입은 데스크톱급 프로세서 기반의 기능 모사 환경에서 수행되었으므로, 실제 모바일 AP 또는 웨어러블 AP 환경과의 물리적 차이를 고려할 필요가 있다. 첫째, ARM TrustZone 기반 TEE의 월드 전환(world

switch)에는 일반적으로 수 μs 수준의 오버헤드가 발생하는 것으로 보고되고 있다[17]. 이는 본 연구에서 측정한 서명 지연(0.5~7ms)에 비해 상대적으로 적은 비중이므로 TEE 진입 자체가 전체 파이프라인의 주된 병목이 될 가능성은 낮다고 보인다.

둘째, 현재 주요 모바일 SoC에는 SHA-256, AES, ECDSA 등의 암호 연산을 하드웨어로 처리하는 전용 암호화 가속기가 탑재되어 있다. Google Pixel 10의 Titan M2는 Common Criteria AVA_VAN.5 인증을 획득한 보안 칩으로서, 소프트웨어 대비 암호 연산의 처리 속도를 높이면서 메인 CPU 점유를 제거한다[9]. HW-CPA 설계에서 TEE 내 전용 가속기를 활용할 경우, 본 연구에서 측정한 서명·검증 지연 시간은 더욱 단축될 것으로 예상한다.

셋째, 스마트 글래스 등의 웨어러블 AP는 열 제어(thermal throttling)로 인해 메인 CPU의 소프트웨어 암호 연산 성능이 데스크톱 대비 저하되는 반면, 전용 하드웨어 가속기는 고정된 저전력 클럭에서도 일관된 처리 성능을 유지하므로, 저전력 환경일수록 하드웨어 오프로딩에 의한 상대적 이점이 커질 것으로 예상된다. 다만 이상의 분석은 문헌 기반 추정이며, 정확한 성능 특성은 향후 FPGA 프로토타입 또는 실제 SoC 환경에서의 실측을 통해 확인할 필요가 있다.

5. 결 론

본 논문은 온디바이스 AI 시대에 생성·촬영되는 콘텐츠의 신뢰성을 확보하기 위한 하드웨어 지향 출처 인증 프레임워크인 HW-CPA를 제안하였다. 기존에는 별개로 논의되던 하드웨어 보안(TEE, HRoT), 콘텐츠 출처 인증(C2PA), 온디바이스 AI(NPU)를 단일 4계층 프레임워크로 통합하였으며, 소프트웨어 기반 접근의 구조적 한계를 위협 모델링을 통해 분석하고 하드웨어 기반 접근의 필요성을 논증하였다. 이는 NIST AI 위험관리 프레임워크[19]가 제시하는 AI 시스템의 신뢰성 원칙과도 맞닿아 있어, 향후 규제 대응의 기술적 기반으로 활용될 수 있을 것으로 기대된다.

그러나 본 논문에는 다음의 한계가 존재한다. 프로토타입은 실제 TEE, PUF, eFuse, Secure RTC를 사용한 하드웨어 실측이 아닌 소프트웨어 기능 모사에 기반하므로, 본문에서 제시한 지연 시간이 실제 칩의 절대 성능을 의미하지는 않는다. Layer 1의 NPU 내장형 분류기는 실제 datapath 통합과 적대적 예제(Adversarial Example)에 대한 강건성 평가를 포함하지 않았다. 또한 C2PA의 완전한 JUMBF 임베딩 및 상호운용성 검증, 다중 디바이스 간 체인 병합 프로토콜, 칩 면적·전력·제조 비용 분석은 본 논문의 범위를 벗어난다. 넷째, TEE 환경에서의 부채널 공격(side-channel attack)[20], 센서 스푸핑 공격 등에 대한 하드웨어 수준의 강건성 평가가 추가로 필요하다.

향후 연구에서는 FPGA 기반 하드웨어 시제품 구현, NPU 통합형 Layer 1의 설계 및 적대적 공격 강건성 평가, 실제 Secure Enclave/TEE 연동 실험, 완전한 C2PA 포맷 내장, 다중 디바이스 체인 확장, Post-Quantum Cryptography 통합 설계를 수행할 계획이다. 궁극적으로 HW-CPA는 스마트 글래스, 모바일 기기, 산업용 엣지 디바이스에서 생성형 AI 콘텐츠의 투명성을 담보하는 핵심 인프라로 성장할 수 있을 것으로 기대한다.

참 고 문 헌

[1] The Traceability Hub, “Digital Provenance

& Content Authentication: Trust in AI Media", 2026, https://thetraceabilityhub.com/digital-provenance-why-content-authentication-matters-in-2026/
[2] Google Research, "Coral NPU: A Full-Stack Platform for Edge AI", Google Developers Blog, 2025, https://research.google/blog/coral-npu-a-full-stack-platform-for-edge-ai/
[3] Ceva, Inc., "Ceva Highlights Breakthrough Year for AI Licensing and Physical AI Adoption in 2025", 2026, https://aicompetence.org/latest-ai-news/?rkey=20260217SF88744&filter=27632
[4] Synopsys, "Securing AI at the Silicon Level for Comprehensive Protection", 2026, https://www.synopsys.com/blogs/chip-design/ai-security-silicon-level.html
[5] O. Aarne, T. Fist, C. Withers, "Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing", CNAS, 2024, https://www.cnas.org/publications/reports/secure-governable-chips
[6] N. Ammann, "Flexible, Hardware-Enabled Guarantees (flexHEG) Report Series", UK ARIA, 2025, https://www.flexheg.com/report-1.pdf
[7] C2PA, "Content Credentials: C2PA Technical Specification v2.2", 2025, https://spec.c2pa.org/specifications/specifications/2.2/index.html
[8] Content Authenticity Initiative, "5,000 Members: Building Momentum for a More Trustworthy Digital World", 2025, https://contentauthenticity.org/blog/5000-members-building-momentum-for-a-more-trustworthy-digital-world
[9] Google Security Blog, "How Pixel and Android are bringing trust to images with C2PA Content Credentials", 2025, https://security.googleblog.com/2025/09/pixel-android-trusted-images-c2pa-content-credentials.html
[10] Google Developers Blog, "MediaTek NPU and LiteRT: Powering the Next Generation of On-device AI", 2025, https://developers.googleblog.com/en/mediatek-npu-and-litert/
[11] Workplace Privacy Report, "Compliance Concerns with AI Smart Glasses, Part 4: Data Security", 2026, https://www.workplaceprivacyreport.com/2026/01/articles/artificial-intelligence/the-hidden-legal-minefield-compliance-concerns-with-ai-smart-glasses-part-4-data-security-breach-notification-and-third-party-ai-processing-risks/
[12] Help Net Security, "Smart glasses are back, privacy issues included", 2026, https://www.helpnetsecurity.com/2026/02/05/ai-smart-glasses-privacy-risk/
[13] A. Gallardo et al., "Speculative Privacy Concerns About AR Glasses Data Collection", Proc. PETs, pp.1-18, 2023, https://doi.org/10.56553/popets-2023-0117
[14] Inairspace, "Augmented Reality Glasses Security: Navigating the New Frontier", 2026, https://inairspace.com/blogs/learn-with-inair/augmented-reality-glasses-security-navigating-the-new-frontier-of-personal-and-corporate-data-protection?srsltid=AfmBOoqo6Fe0P-nHx78mqejoELQ6m83sq4K_FrVt_BBLEKTLD74wG2Em
[15] CISA/NSA, "Content Credentials", U/OO/109191-25, 2025, https://www.nsa.gov/Press-Room/Cybersecurity-Advisories-Guidance/
[16] A. Reuel et al., "Hardware-Enabled Mechanisms for Verifying Responsible AI Development", arXiv:2505.03742, 2025, https://doi.org/10.48550/arXiv.2505.03742
[17] ARM, "ARM TrustZone Technology", ARM Security Technology White Paper, 2023,

https://developer.arm.com/documentation/
[18] GlobalPlatform, “TEE System Architecture v1.3”, 2022, https://globalplatform.org/specs-library/tee-system-architecture/
[19] E. Tabassi, “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, NIST, 2023, https://doi.org/10.6028/NIST.AI.100-1
[20] A. Muñoz, R. Rios, R. Roman, J. Lopez, “A survey on the (in)security of trusted execution environments”, Computers & Security, Vol.129, pp.1-23, 2023, https://doi.org/10.1016/j.cose.2023.103180

저 자 소 개

신수홍(Su-Hong Shin)

2011.2 호서대학교 컴퓨터공학부 졸업
2013.2 호서대학교 컴퓨터공학 석사
2013.2-현재: ㈜디지캡 비즈니스솔루션사업실 모빌리티BE개발팀 팀장
<주관심분야> 정보보안, 인프라, AI

고병수(Byoung-Soo Koh)

2004.8 대전대학교 컴퓨터공학과 박사
2011.9~2020.2 한국공학대학교 겸임교수
2020.3~2025.7 한국콘텐츠진흥원 저작권 PD
2025.8-2026.02 : ㈜디지캡 미래성장연구실 실장
2026.03-현재 : 한신대학교 AI.SW대학 교수
<주관심분야> 저작권, 포렌식, AI