TruthLens v4.4 사용자 가이드

AI 기반 딥페이크 탐지 플랫폼 운영 매뉴얼

항목	내용
문서 버전	Rev 4.1
대상 시스템	TruthLens v4.4.0
작성일	2026년 4월 14일
작성	Brian Lee | AI R&D Center | A3 Security Co.,Ltd.

머리말 — 왜 딥페이크 탐지가 필요한가

2024년 이후 전 세계적으로 딥페이크(Deepfake) 기술을 악용한 사기 피해가 급증하고 있습니다. 2025년 대한민국에서만 딥페이크 관련 사이버 범죄 신고가 전년 대비 340% 증가하였으며, 금융기관 대상 보이스피싱에 AI 음성 합성이 활용된 사례, 공직자를 사칭한 딥페이크 영상으로 인한 여론 조작 사건, 기업 임원의 화상 회의 영상을 위조하여 수십억 원의 송금을 유도한 BEC(Business Email Compromise) 사건이 연이어 보도되었습니다.

이러한 사회적 위협에 대응하기 위하여, TruthLens는 40개 독립 탐지 모듈, 8개 멀티 에이전트 AI 프레임워크, OWL 온톨로지 기반 파이프라인 검증 시스템을 통합한 차세대 멀티모달 딥페이크 탐지 플랫폼으로 개발되었습니다. 본 플랫폼은 한국 AI 기본법 제15조의 "AI 판단 근거 설명 의무"를 준수하며, 모든 판정 결과에 대하여 설명 가능한 AI(XAI) 시각화와 SHA-256 해시 체인 기반 감사 로그를 제공합니다.

TruthLens는 정부기관, 금융기관, 방송사, 수사기관, 그리고 기업의 정보보안 담당자가 별도의 AI 전문 지식 없이도 의심 미디어를 신속하게 분석하고 과학적 근거에 기반한 판정을 내릴 수 있도록 설계되었습니다.

목차

전체 메뉴 구조 — 10개 섹션 / 42개 워크스페이스

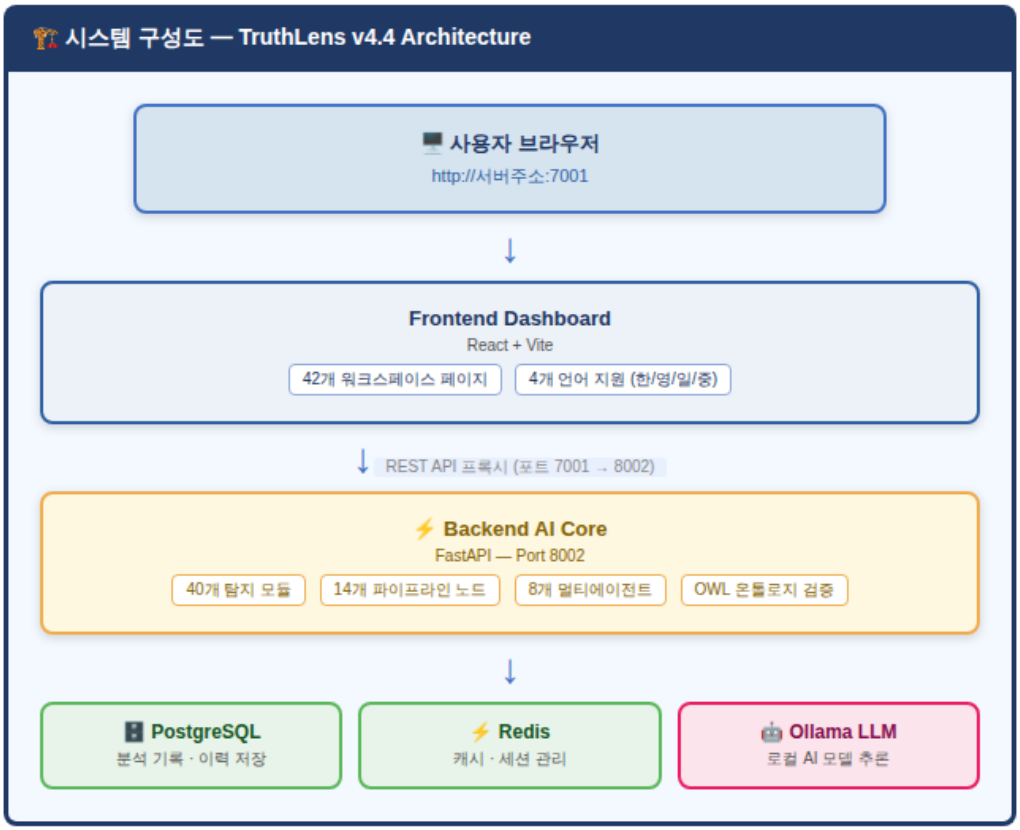
---

1. 시스템 개요

1.1 TruthLens란

TruthLens는 영상, 음성, 이미지, 문서 등 다양한 형식의 미디어 파일에 대하여 AI 생성 또는 조작 여부를 자동으로 판별하는 딥페이크 탐지 플랫폼입니다. 40개의 독립적인 탐지 모듈이 각기 다른 관점에서 미디어를 분석하고, 그 결과를 멀티 에이전트 AI 시스템이 종합하여 최종 판정을 내립니다.

1.2 시스템 구성




```
사용자 브라우저 (7001번 포트)
    │  (API 프록시)
    ▼
Frontend Dashboard (React + Vite) — 42개 워크스페이스
    │
    ▼
Backend AI Core (FastAPI, 8002번 포트)
│   40개 탐지 모듈 + 14개 파이프라인 노드
│   8개 멀티에이전트 + OWL 온톨로지 검증 시스템
    ├── PostgreSQL (분석 기록)
    ├── Redis (캐시)
    └── Ollama LLM (AI 모델)
```

1.3 판정 결과의 의미

TruthLens는 모든 분석에 대하여 세 가지 판정 중 하나를 제공합니다.

판정	의미	배지 색상
REAL (진본)	분석 결과 AI 생성 또는 조작의 흔적이 발견되지 않았음	초록색
FAKE (위조)	AI 생성 또는 조작의 흔적이 유의미하게 발견됨	빨간색
UNCERTAIN (판단 유보)	판단을 내리기에 증거가 불충분하거나 모호함	노란색

각 판정에는 신뢰도(Confidence) 점수가 함께 제공됩니다. 0%에서 100% 사이의 값으로, 높을수록 판정에 대한 시스템의 확신이 강합니다.

> *주의사항: UNCERTAIN 판정은 "분석 실패"가 아닙니다. 시스템이 결론을 내리기에 근거가 불충분하다는 의미이며, 이 경우 전문가의 추가 검토를 권고합니다.*

---

2. 로그인 및 초기 화면

2.1 로그인

웹 브라우저에서 http://서버주소:7001에 접속하면 로그인 화면이 표시됩니다. 관리자로부터 발급받은 이메일과 비밀번호를 입력하여 로그인합니다.

2.2 사용자 역할

역할	권한
VIEWER (열람자)	분석 실행 및 결과 조회
ANALYST (분석가)	열람자 권한 + API 키 관리 + 보고서 생성
ADMIN (관리자)	분석가 권한 + 사용자/조직 관리 + 시스템 설정

2.3 화면 구성

- 좌측 사이드바: 10 개 섹션으로 분류된 메뉴 항목. 각 섹션은 펼침/접힘이 가능합니다.
- 상단 헤더: 현재 페이지명, 언어 선택, 사용자 정보
- 중앙 콘텐츠: 선택한 메뉴에 해당하는 워크스페이스

---

3. 탐지 섹션

이 섹션은 TruthLens의 핵심 기능으로, 미디어 파일을 업로드하여 분석하고 결과를 확인하는 워크스페이스입니다.

3.1 분석 (Analyze)

메뉴 경로: 사이드바 > 분석 | 경로: /

이 워크스페이스는 딥페이크 탐지의 시작점입니다. 다음 두 가지 방식으로 미디어를 분석할 수 있습니다.

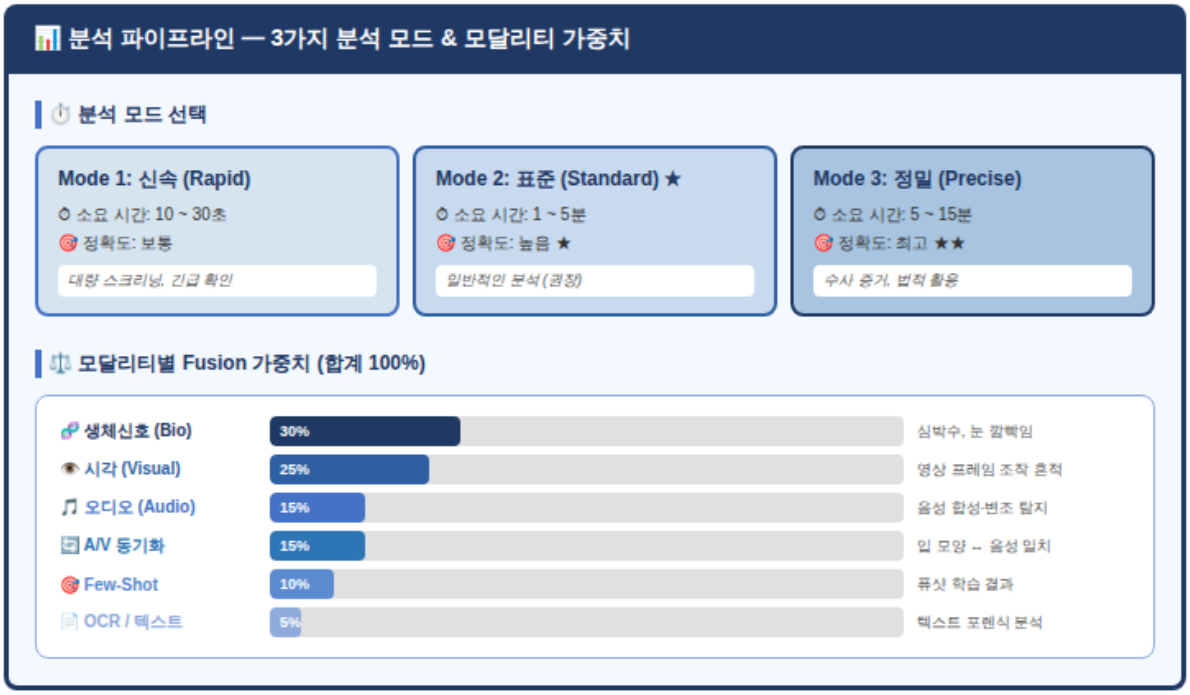
파일 업로드 방식

- "파일 선택" 버튼을 클릭하여 분석할 영상, 음성, 또는 이미지 파일을 선택합니다.
- 지원 형식: MP4, AVI, MOV, MKV, WebM (영상) / MP3, WAV, FLAC (음성) / JPG, PNG, WebP (이미지)
- 최대 파일 크기: 500MB

URL 입력 방식

- YouTube, Instagram, TikTok 등 소셜 미디어 URL 을 입력합니다.
- 시스템이 자동으로 미디어를 다운로드하여 분석합니다.
- 1,000 개 이상의 사이트를 지원합니다.

분석 모드 선택



모드	소요 시간	정확도	권장 상황
신속 (Rapid)	10~30초	보통	대량 스크리닝, 긴급 확인
표준 (Standard)	1~5분	높음	일반적인 분석 (권장)
정밀 (Precise)	5~15분	최고	수사 증거, 법적 활용

용어 해설

용어	설명
딥페이크(Deepfake)	딥러닝 기술로 사람의 얼굴, 목소리, 동작 등을 합성하거나 변조한 미디어
모달리티(Modality)	분석 대상의 유형. 시각, 청각, 생체신호 등 각기 다른 분석 채널
신뢰도(Confidence)	시스템이 판정에 대해 갖는 확신의 정도 (0~100%)

3.2 분석 결과 (Results)

메뉴 경로: 분석 완료 후 자동 이동 | 경로: /results/:jobId

분석이 완료되면 이 워크스페이스에 종합 결과가 표시됩니다.

판정 배지: 화면 상단에 REAL(초록) / FAKE(빨강) / UNCERTAIN(노란) 배지와 신뢰도가 크게 표시됩니다.

모달리티별 점수 (레이더 차트)

방사형 차트(레이더 차트)로 각 모달리티의 분석 결과가 표시됩니다.

- 시각 분석(Visual): 영상 프레임의 얼굴 조작 흔적 분석 결과
- 오디오 분석(Audio): 음성 합성 또는 변조 흔적 분석 결과
- 생체신호 분석(Biological): 심박수, 눈 깜빡임 등 생체 반응의 자연스러움
- A/V 동기화(A/V Sync): 입 모양과 음성의 일치 여부

그래프 판독법

레이더 차트는 중심에서 바깥쪽으로 갈수록 위조 의심 점수가 높아지는 방사형 그래프입니다. 각 축의 값이 0.5(50%)를 초과하면 해당 모달리티에서 위조 의심 징후가 발견된 것입니다. 값이 높을수록 위조 가능성이 높으며, 0.8 이상은 강한 위조 증거로 해석합니다. 모든 축이 0.3 이하로 중심 근처에 밀집되어 있으면 진본일 가능성이 높습니다.

증거 체인(Evidence Chain): 판정 근거가 된 구체적인 증거 항목이 심각도(HIGH/MEDIUM/LOW)와 함께 나열됩니다.

3.3 분석 이력 (History)

메뉴 경로: 사이드바 > 분석 이력 | 경로: /history

지금까지 수행한 모든 분석의 이력을 시간순으로 조회할 수 있습니다.

- 파일명: 분석한 미디어 파일의 원본 이름
- 판정: REAL / FAKE / UNCERTAIN

- 신뢰도: 0~100%
- 분석 시각: 분석이 완료된 일시
- 소요 시간: 분석에 걸린 시간 (밀리초)

검색 기능을 통해 파일명 또는 판정 유형으로 이력을 필터링할 수 있습니다.

3.4 BI 대시보드

메뉴 경로: 사이드바 > BI 대시보드 | 경로: /bi

비즈니스 인텔리전스(BI) 대시보드는 분석 통계를 시각적으로 요약합니다.

- 판정 분포 파이차트: REAL/FAKE/UNCERTAIN 의 비율
- 일별 분석 추이: 날짜별 분석 건수 추이 그래프
- 평균 신뢰도: 기간별 평균 신뢰도 변화

그래프 판독법

파이차트에서 FAKE 비율이 전체의 30%를 초과하면 해당 기간에 딥페이크 유입이 매우 활발했음을 의미하며, 유입 경로와 시기에 대한 긴급 점검이 필요합니다. 정상적인 운영 환경에서 FAKE 비율은 통상 5~15% 범위를 유지합니다. 평균 신뢰도가 지속적으로 낮아지는 추세라면 탐지 대상 딥페이크의 품질이 고도화되고 있다는 신호일 수 있으므로 모델 재학습이나 탐지 임계값 재조정을 검토해야 합니다.

---

4. 공간/주파수 탐지 섹션

이 섹션은 영상의 시각적 특성을 분석하는 전문 워크스페이스입니다. 각 도구는 서로 다른 기법으로 영상의 조작 흔적을 탐지합니다.

4.1 DM 탐지 센터

메뉴 경로: 사이드바 > 공간/주파수 탐지 > DM 탐지 센터 | 경로: /dm-detection

기능: Stable Diffusion, Midjourney, DALL-E 등 확산 모델(Diffusion Model)로 생성된 이미지를 전문적으로 탐지합니다. 확산 모델은 노이즈에서 시작하여 점진적으로 이미지를 복원하는 방식으로 동작하며, 이 과정에서 고유한 통계적 패턴(DIRE)이 내재됩니다. TruthLens는 이 패턴을 3단계 파이프라인(스크리닝 → 고속 → 정밀)으로 분석 깊이를 높여 탐지합니다.

화면 구성

- 탐지 결과 카드: 각 분석 건에 대한 판정, 신뢰도, 소요 시간
- DM 확률 점수: 확산 모델로 생성되었을 확률 (0~1)
- 스테이지별 결과: 스크리닝/고속/정밀 각 단계의 점수

그래프 판독법

DM 확률 점수 0.7 이상이면 확산 모델 생성 의심도가 높고 즉각적인 추가 분석이 권고됩니다. 스테이지별 점수가 단계를 거칠수록 상승하는 패턴은 위조 가능성이 높은 신호입니다. 0.3~0.7 구간은 판단 유보 구간으로, 반드시 ViT 분석 및 주파수 분석 결과와 교차 검증이 필요합니다.

용어 해설

용어	설명
확산 모델(Diffusion Model)	이미지를 점진적으로 생성하는 AI 기법. Stable Diffusion, DALL-E 등이 대표적
스크리닝(Screening)	대량의 이미지를 빠르게 1차 필터링하는 과정
DIRE	Diffusion Reconstruction Error — 확산 모델이 남기는 고유한 오차 패턴을 탐지하는 기법

4.2 ViT 분석

메뉴 경로: 사이드바 > 공간/주파수 탐지 > ViT 분석 | 경로: /vit-analysis

기능: 비전 트랜스포머(Vision Transformer, ViT) 모델을 활용하여 이미지의 조작 흔적을 탐지합니다. ViT는 이미지를 작은 패치로 분할한 뒤 트랜스포머의 어텐션 메커니즘으로 파악하는 최신 아키텍처로, Swin-V2, CrossViT, TIMM 등 세 가지 ViT 변형 모델의 결과를 앙상블하여 탐지 정확도를 높입니다.

화면 구성

- 모델별 점수: Swin-V2 Score, CrossViT Score, TIMM Score, Ensemble Score
- 어텐션 히트맵: 모델이 주목한 영역을 색상으로 표시한 지도 (파란색=정상, 빨간색=의심)
- 조작 영역: 감지된 조작 의심 영역의 위치와 이상 점수

그래프 판독법

어텐션 히트맵에서 빨간색이 얼굴 경계부, 머리카락-이마 경계, 턱선 주변에 집중되어 있으면 페이스스왑의 전형적인 패턴입니다. 세 모델의 앙상블 점수가 0.7 이상이면서 히트맵의 의심 영역이 일치하면 판정의 신뢰도가 매우 높습니다. 세 모델 간 점수 편차가 크면 해당 이미지에 대한 추가 전문가 검토가 필요합니다.

용어 해설

용어	설명
비전 트랜스포머(ViT)	이미지를 작은 조각(패치)으로 나누어 문맥을 파악하는 AI 모델
어텐션 히트맵	AI 모델이 이미지의 어느 부분에 주목하는지를 색상으로 시각화한 지도
앙상블(Ensemble)	여러 모델의 결과를 종합하여 하나의 판단을 내리는 기법

4.3 블렌딩 경계 분석

메뉴 경로: 사이드바 > 공간/주파수 탐지 > 블렌딩 경계 | 경로: /blending-analysis

기능: 얼굴을 합성할 때 생기는 경계면의 부자연스러움을 탐지합니다. LAA(Learned Adaptive Attention) 기법으로 경계면의 이상 패턴을 감지하고, Poisson 블렌딩 알고리즘의 수학적 흔적을 역추적하는 방법이 함께 사용됩니다. 색상 및 조명 불일치 점수를 교차 분석하여 조잡한 딥페이크와 정교한 딥페이크를 모두 탐지합니다.

화면 구성

- LAA 점수: Learned Adaptive Attention — 학습된 적응적 주의력 점수
- Poisson 점수: 푸아송 블렌딩 흔적 점수
- 색상 불일치: 합성된 얼굴과 원본 배경 사이의 색감 차이
- 조명 불일치: 얼굴과 배경의 조명 방향 차이

그래프 판독법

LAA 점수와 Poisson 점수가 동시에 0.6 이상이면 합성 경계가 존재할 가능성이 매우 높습니다. 색상·조명 불일치는 낮지만 LAA 점수만 높은 경우는 정교하게 처리된 고급 딥페이크를 의미하므로 오히려 더 위험한 케이스로 간주해야 합니다.

용어 해설

용어	설명
블렌딩(Blending)	두 이미지를 자연스럽게 합치는 기법. 딥페이크에서 얼굴을 교체할 때 사용됨
LAA	경계면의 이상 패턴을 학습한 AI가 감지한 의심도
푸아송 블렌딩(Poisson Blending)	이미지 합성 시 경계를 매끄럽게 처리하는 수학적 기법

4.4 주파수 분석

메뉴 경로: 사이드바 > 공간/주파수 탐지 > 주파수 분석 | 경로: /frequency-analysis

기능: 이미지를 주파수 도메인으로 변환하여 AI 생성 이미지 특유의 패턴을 탐지합니다. FFT(고속 푸리에 변환), DCT(이산 코사인 변환), Wavelet 변환 등 세 가지 독립적인 수학적 변환을 병렬로 수행하여 교차 검증합니다. GAN 계열 생성 모델은 업샘플링 과정에서 "체커보드(Checkerboard)" 패턴이라 불리는 격자형 주파수 아티팩트를 남기며, 이것이 GAN 핑거프린트 탐지의 핵심 근거가 됩니다.

화면 구성

- FFT 점수: 고속 푸리에 변환 기반 이상 점수
- DCT 점수: 이산 코사인 변환 기반 이상 점수
- Wavelet 점수: 웨이블릿 변환 기반 이상 점수
- GAN 핑거프린트 점수: GAN 특유의 주파수 패턴 점수
- 스펙트럼 이상치: 주파수 스펙트럼에서 발견된 비정상 패턴

그래프 판독법

GAN 핑거프린트 점수가 0.7 이상이면 GAN으로 생성된 이미지일 가능성이 높습니다. FFT, DCT, Wavelet 세 점수가 모두 0.6 이상으로 높으면 여러 분석법에 의해 교차 검증된 것이므로 신뢰도가 크게 강화됩니다. 스펙트럼 이상치에서 고주파 대역의 갑작스러운 단절은 JPEG 재압축, 리샘플링 등 후처리 흔적을 나타냅니다.

용어 해설

용어	설명
주파수 도메인(Frequency Domain)	이미지를 밝기의 변화 패턴(주파수)으로 분해한 표현. 고주파는 세부 디테일
FFT(Fast Fourier Transform)	이미지를 주파수 성분으로 분해하는 수학적 변환
GAN(Generative Adversarial Network)	생성자와 판별자가 경쟁하며 이미지를 생성하는 AI 기법
핑거프린트(Fingerprint)	AI 모델이 생성물에 남기는 고유한 패턴

4.5 Foundation 모델

메뉴 경로: 사이드바 > 공간/주파수 탐지 > Foundation 모델 | 경로: /foundation-models

기능: 최신 대규모 사전학습 모델(Foundation Model)을 활용하여 딥페이크 및 AI 생성 콘텐츠를 탐지합니다. UniFD(범용 위조 탐지), LNCLIP-DF(CLIP 기반 딥페이크 특화), DINOv2(Meta의 자기지도 학습 비전 모델)의 세 가지 이질적인 모델을 병렬로 실행하여 미지의 생성 방식도 포착할 수 있습니다.

화면 구성

- UniFD 점수: 범용 위조 탐지 모델의 판정 점수
- LNCLIP-DF 점수: CLIP 기반 딥페이크 특화 모델의 점수
- DINOv2 점수: Meta 의 자기지도 학습 비전 모델의 점수
- 앙상블 점수: 세 모델을 종합한 최종 점수

그래프 판독법

세 모델의 점수가 모두 0.6 이상이면 여러 관점에서 교차 검증된 위조 가능성이 높습니다. UniFD만 높으면 전통적인 GAN 계열, LNCLIP-DF만 높으면 텍스트-이미지 생성 모델(Stable Diffusion, Midjourney 등) 계열일 가능성이 있습니다.

용어 해설

용어	설명
파운데이션 모델 (Foundation Model)	대규모 데이터로 사전 학습된 범용 AI 모델
CLIP	OpenAI가 개발한 이미지-텍스트 연결 모델

자기지도 학습(Self-Supervised Learning)	사람이 정답을 제공하지 않아도 데이터의 구조를 스스로 학습하는 방법

---

5. 시간/생체 분석 섹션

이 섹션은 영상의 시간적 흐름과 인물의 생체 신호를 분석합니다. 딥페이크는 개별 프레임에서는 완벽해 보일 수 있지만, 시간에 따른 변화 패턴에서 부자연스러움이 드러나는 경우가 많습니다.

5.1 시간 도메인 분석

메뉴 경로: 사이드바 > 시간/생체 분석 > 시간 도메인 분석 | 경로: /temporal-analysis

기능: 영상의 연속된 프레임 간 시간적 일관성을 분석합니다. VideoMAE와 TALL 알고리즘을 활용하여 초당 30프레임 기준으로 각 프레임의 시간적 이상도를 측정합니다. 딥페이크 제작 시 사용되는 프레임 스티칭(Frame Stitching) 기법은 특정 구간에서만 얼굴을 교체하므로, 교체 전후 프레임에서 급격한 점수 변동이 발생하는 것이 탐지의 핵심 단서가 됩니다.

화면 구성

- 시간적 일관성 점수: 프레임 간 변화의 자연스러움 (높을수록 의심)
- 프레임별 점수 그래프: 각 프레임의 위조 의심도를 시간축으로 표시
- 감지된 생성기: 추정되는 딥페이크 생성 도구

그래프 판독법

가로축이 영상의 시간(초), 세로축이 위조 의심도(0~1)인 시계열 그래프입니다. 전체적으로 0.3 이하로 안정적인 흐름을 보이면 진본일 가능성이 높습니다. 갑작스러운 급등 구간은 해당 시점에서 얼굴 합성이 불안정하게 이루어진 것을 의미하며 해당 전후 프레임을 반드시 확인해야 합니다.

용어 해설

용어	설명
시간적 일관성(Temporal Consistency)	연속된 프레임 간의 자연스러운 변화 유지 여부
VideoMAE	영상의 시간적 패턴을 학습하는 자기지도 학습 비전 트랜스포머
TALL	Temporal Attention Learned Locally — 지역적 시간 주의력 기법

5.2 립싱크 포렌식

메뉴 경로: 사이드바 > 시간/생체 분석 > 립싱크 포렌식 | 경로: /lip-forensics

기능: 입술 움직임과 음성의 동기화 정확도를 밀리초 단위로 정밀 분석합니다. Wav2Lip, VideoRetalking 등 주요 립싱크 생성 도구의 특성 패턴을 학습하여 어떤 도구가 사용되었는지를 역추적합니다. 음소별로 기대되는 입 모양 패턴과 실제 영상의 입 모양을 비교하는 음성-시각 상관 분석도 수행됩니다.

화면 구성

- 동기화 편차 그래프: 시간축에 따른 입술-음성 동기화 차이
- 동기화 점수: 전체 평균 동기화 품질 (낮을수록 의심)
- 감지된 립싱크 도구: Wav2Lip, VideoRetalking 등 추정 도구

그래프 판독법

동기화 편차 그래프는 가로축이 시간(초), 세로축이 편차(밀리초, ms)로 표시됩니다. 자연스러운 영상에서는 편차가 ±20ms 이내에서 균일하게 분포하는 것이 정상입니다. 지속적으로 50ms 이상의 편차가 나타나면 음성 변조의 강력한 징후이며, 100ms 이상이면 거의 확실한 음성 교체의 증거입니다.

용어 해설

용어	설명
립싱크(Lip-Sync)	입술 움직임과 음성의 동기화
Wav2Lip	음성에 맞춰 입술 움직임을 합성하는 딥러닝 모델
동기화 편차(Sync Offset)	입 모양과 실제 발음 사이의 시간 차이 (밀리초 단위)

5.3 미세 표정 분석

메뉴 경로: 사이드바 > 시간/생체 분석 > 미세 표정 분석 | 경로: /micro-expression

기능: 0.04~0.5초 사이에 나타나는 미세 표정(Micro-Expression)의 자연스러움을 분석합니다. FACS(Facial Action Coding System) 기반의 AU(Action Unit) 분석을 통해 43개의 얼굴 근육 움직임을 개별적으로 추적하며, 각 AU의 발생 빈도, 지속 시간, 강도를 통계적으로 분석합니다. 딥페이크 영상에서는 미세 표정이 전혀 나타나지 않거나 과도하게 과장되는 양극단의 패턴이 주로 관찰됩니다.

화면 구성

- 감지된 미세 표정 수: 분석 구간에서 발견된 미세 표정의 횟수
- 표정별 자연스러움 점수: 각 감지된 표정의 자연성 평가
- 이상 구간: 부자연스러운 표정 전환이 발생한 시간대

그래프 판독법

진본 영상에서는 통계적으로 분당 2~5회의 미세 표정이 감지되는 것이 정상 범주입니다. 분당 0~1회이면 딥페이크로 인해 표정이 억제된 것을 의심해야 합니다. 반대로 분당 10회 이상이면 딥페이크 생성 모델이 표정을 과장되게 렌더링한 경우입니다.

용어 해설

용어	설명
미세 표정(Micro-Expression)	무의식적으로 매우 짧은 시간 동안 나타나는 얼굴 표정. 의식적으로 통제하기 어려움
AU(Action Unit)	얼굴 근육의 개별 움직임 단위. FACS(Facial Action Coding System)에서 정의

5.4 양안(시선) 분석

메뉴 경로: 사이드바 > 시간/생체 분석 > 양안 분석 | 경로: /gaze-analysis

기능: 양쪽 눈의 시선 방향 일관성과 생체적 정상성을 분석합니다. 실제 사람의 양안은 하나의 신경 시스템에 의해 동기화되어 항상 동일한 지점을 주시하지만, 딥페이크에서는 두 눈을 독립적으로 합성하는 과정에서 시선이 미세하게 어긋나는 현상이 발생합니다. 눈 깜빡임 주기(보통 3~5초 간격), 눈 깜빡임 지속 시간(보통 150~300ms), 홍채 반사광의 일관성 등 다양한 생체 지표를 종합적으로 평가합니다.

화면 구성

- 시선 추적 궤적: 양쪽 눈의 시선 방향을 시간축으로 표시
- 양안 편차: 왼쪽 눈과 오른쪽 눈의 시선 차이
- 시선 일관성 점수: 전체 평균 일관성 (높을수록 정상)

그래프 판독법

시선 추적 궤적 그래프에서 왼쪽 눈(파란 선)과 오른쪽 눈(빨간 선)이 밀착하여 겹쳐 보이면 정상적인 양안 동기화입니다. 양안 편차가 지속적으로 5도 이상이면 딥페이크의 강력한 증거로 해석합니다. 시선 일관성 점수가 0.7 이상이면 양안 동기화가 자연스럽고, 0.4 이하이면 강한 의심 판정에 해당합니다.

용어 해설

용어	설명
양안 시선(Binocular Gaze)	양쪽 눈이 동시에 같은 지점을 향하는 자연스러운 움직임
시선 편차(Gaze Deviation)	양쪽 눈의 시선 방향 차이 (각도 단위)

5.5 머리 포즈 역학

메뉴 경로: 사이드바 > 시간/생체 분석 > 머리 포즈 역학 | 경로: /head-pose

기능: 머리의 회전(Yaw), 기울임(Pitch), 좌우 기울기(Roll)의 역학적 자연스러움을 물리 법칙 기반으로 분석합니다. 딥페이크 영상에서는 생성 알고리즘의 한계로 인해 머리 움직임이 지나치게 부드럽거나, 반대로 물리적으로 불가능한 속도의 급격한 전환이 발생할 수 있습니다.

화면 구성

- 3 축 회전 그래프: Yaw(좌우), Pitch(상하), Roll(기울기)의 시간별 변화
- 변위량(Displacement): 각 시점에서의 머리 이동량
- 이상치 표시: 비정상적인 급격한 움직임 구간

그래프 판독법

자연스러운 영상에서는 세 축 모두 완만하게 변화하며 미세한 떨림(noise)이 포함된 곡선을 보입니다. 물리적으로 불가능한 속도의 회전이 감지되면 합성 영상의 강력한 증거입니다. 반대로 세 그래프가 완벽하게 매끄러운 곡선을 그리면 딥페이크 생성 과정에서 노이즈가 제거된 것으로 강한 의심 징후입니다.

용어 해설

용어	설명
Yaw	머리의 좌우 회전 (고개를 양옆으로 돌리는 동작)
Pitch	머리의 상하 기울임 (고개를 끄덕이거나 젖히는 동작)
Roll	머리의 좌우 기울기 (어깨 쪽으로 고개를 기울이는 동작)

---

6. 오디오 포렌식 섹션

이 섹션은 음성의 진위를 판별하는 전문 도구를 제공합니다. AI 음성 합성(TTS), 보이스 클로닝 등의 흔적을 탐지합니다.

6.1 오디오 포렌식

메뉴 경로: 사이드바 > 오디오 포렌식 > 오디오 포렌식 | 경로: /audio-forensics

기능: 음성 파일의 종합적인 진위를 다각도로 분석하는 핵심 오디오 분석 도구입니다. MFCC(멜 주파수 켑스트럼 계수) 분석으로 음성의 고유한 음색과 조음 패턴의 자연스러움을 평가하고, 스펙트럼 분석으로 주파수 대역별 에너지 분포의 이상을 탐지합니다. 영교차율(ZCR), 스펙트럴 센트로이드, 에너지 엔벨로프 등 20여 가지 음향 특성 지표를 병렬로 산출합니다.

화면 구성

- 판정 결과: 음성의 REAL/FAKE/UNCERTAIN 판정
- MFCC 표준편차: 음성 특성의 변동성 (낮으면 합성 의심)
- 영교차율(ZCR): 음성 신호가 0 을 지나는 빈도
- 스펙트럼 분석: 주파수 대역별 에너지 분포

그래프 판독법

MFCC 표준편차가 0.5 이하로 비정상적으로 낮으면 AI 합성 음성의 특징으로 자연 음성의 풍부한 변동성이 결여된 상태입니다. 스펙트럼 시각화에서 4kHz 이상의 고주파 대역이 갑자기 평탄해지거나 끊기는 패턴은 TTS 또는 보코더 처리의 흔적입니다.

용어 해설

용어	설명
MFCC	멜 주파수 켑스트럼 계수 — 음성의 고유한 특성을 수치화한 지표
ZCR(Zero-Crossing Rate)	음성 신호가 0을 지나는 빈도. 자연 음성과 합성 음성에서 차이를 보임
스펙트럼(Spectrum)	음성을 주파수 성분별로 분해한 표현
TTS(Text-to-Speech)	텍스트를 음성으로 변환하는 기술

6.2 SSL 음성 탐지

메뉴 경로: 사이드바 > 오디오 포렌식 > SSL 음성 탐지 | 경로: /audio-ssl

기능: 자기지도 학습(Self-Supervised Learning, SSL) 기반 대규모 사전학습 모델을 활용하여 합성 음성을 탐지합니다. wav2vec 2.0, HuBERT, WavLM 등 Meta, Microsoft가 개발한 수억 개의 음성 데이터로 사전 학습된 모델이 고차원 특성을 자동으로 추출합니다. 특히 학습 데이터에 없던 신규 TTS 모델이나 보이스 클로닝 서비스에 대한 일반화 탐지 성능이 강점입니다.

화면 구성

- SSL 점수: 자기지도 학습 모델의 위조 의심도
- 클러스터 분석: 음성 특성의 분포 시각화

용어 해설

용어	설명
SSL(Self-Supervised Learning)	레이블 없이 데이터의 내재적 구조를 학습하는 방법
wav2vec	Meta가 개발한 음성 자기지도 학습 모델

6.3 보코더 식별

메뉴 경로: 사이드바 > 오디오 포렌식 > 보코더 식별 | 경로: /vocoder-id

기능: AI 음성 합성에 사용된 보코더(Vocoder)의 종류를 식별하는 역추적 분석 도구입니다. WaveNet, WaveGlow, HiFi-GAN, MelGAN, BigVGAN 등 20종 이상의 대표적인 보코더 모델에 대한 특성 데이터베이스를 보유하고 있으며, 식별된 보코더 정보는 수사기관이 음성 위조의 출처와 제작 도구를 추적하는 디지털 포렌식 과정에서 중요한 단서로 활용됩니다.

화면 구성

- 식별된 보코더: WaveNet, WaveGlow, HiFi-GAN 등 추정 보코더
- 보코더별 유사도: 각 보코더 유형과의 일치도 (바 차트)
- 자연 음성 유사도: 실제 사람 목소리와의 유사도

그래프 판독법

보코더별 유사도 바 차트에서 특정 보코더의 유사도가 0.7 이상이면 해당 보코더로 합성되었을 가능성이 매우 높습니다. 1순위와 2순위 보코더 간 유사도 격차가 0.3 이상이면 식별의 신뢰도가 높고, 격차가 0.1 미만이면 여러 보코더의 특성이 혼재하는 것으로 추가 분석이 필요합니다.

용어 해설

용어	설명
보코더(Vocoder)	음성 합성의 마지막 단계에서 파형을 생성하는 모듈
WaveNet	Google DeepMind가 개발한 음성 합성 보코더

HiFi-GAN	고품질 음성 파형을 빠르게 생성하는 GAN 기반 보코더

---

7. 포렌식 도구 섹션

이 섹션은 분석 결과를 심층적으로 해석하고, 법적 증거로 활용 가능한 보고서를 생성하는 전문 도구를 제공합니다.

7.1 포렌식 보고서

메뉴 경로: 사이드바 > 포렌식 도구 > 포렌식 보고서 | 경로: /forensic-report

기능: 분석 결과를 법적 증거력을 갖춘 포렌식 보고서로 자동 생성합니다. PDF 또는 Excel 형식으로 다운로드할 수 있으며, 국과수(국립과학수사연구원) 제출 형식과 수사기관 표준 포맷을 지원합니다. SHA-256 해시 체인으로 보호되는 감사 로그와 타임스탬프가 포함되어 보고서의 무결성과 생성 시각이 법적으로 검증 가능합니다.

화면 구성

- 보고서 목록: 생성된 보고서 이력
- SHAP 분석: 각 탐지 모듈이 판정에 기여한 정도
- 베이지안 불확실성: 각 모듈의 판정 신뢰 구간
- 해시 체인 검증: 보고서의 무결성 검증 상태

그래프 판독법

SHAP 분석의 폭포(Waterfall) 차트는 기저값(Base Value = 0.5)에서 시작하여 각 탐지 모듈이 최종 판정을 FAKE(+) 또는 REAL(-) 방향으로 얼마나 밀었는지를 순서대로 누적하여 보여줍니다. 빨간색 바는 FAKE 방향, 파란색 바는 REAL 방향으로의 기여를 나타냅니다. 해시 체인 상태가 "VALID"이어야 보고서의 무결성이 확인된 것이며, "BROKEN"이면 법적 효력이 없습니다.

용어 해설

용어	설명
SHAP(SHapley Additive exPlanations)	각 입력 요소가 결과에 얼마나 기여했는지를 수학적으로 분해하는 설명 기법
베이지안 불확실성(Bayesian Uncertainty)	판정의 불확실성을 확률 분포로 표현하는 방법
해시 체인(Hash Chain)	각 기록의 무결성을 이전 기록의 해시값으로 연결하여 보장하는 기법

7.2 핑거프린트 DB

메뉴 경로: 사이드바 > 포렌식 도구 > 핑거프린트 DB | 경로: /fingerprint-db

기능: 분석된 미디어의 디지털 핑거프린트(Digital Fingerprint)를 데이터베이스에 체계적으로 저장하고 관리하여, 동일하거나 변형된 콘텐츠의 재유포 여부를 자동으로 추적합니다. 각 미디어 파일에서 고유한

지각적 해시(Perceptual Hash)를 추출하여 저장하며, 원본 딥페이크 파일이 약간 변형되더라도 동일 출처 콘텐츠를 추적할 수 있게 합니다.

7.3 위협 인텔리전스

메뉴 경로: 사이드바 > 포렌식 도구 > 위협 인텔리전스 | 경로: /threat-intel

기능: 딥페이크 위협 현황을 매트릭스 형태로 시각화하여 기관의 보안 전략 수립을 지원합니다. 공격 유형별(페이스스왑, 보이스 클로닝, 전체 생성, 텍스트-이미지 등), 대상별(금융, 정치, 방송, 기업) 위협 수준을 색상 코드로 직관적으로 파악할 수 있습니다.

화면 구성

- 위협 매트릭스: 공격 유형(행) × 대상(열)의 위험도 격자
- 최근 위협: 최근 감지된 위협 이벤트 목록
- 추세 그래프: 위협 유형별 시간 추이

7.4 XAI 심층 분석

메뉴 경로: 사이드바 > 포렌식 도구 > XAI 심층 분석 | 경로: /xai-deep-dive

기능: 판정의 근거를 설명 가능한 AI(XAI) 기법으로 심층 분석하는 고급 분석 도구입니다. "왜 이 영상이 FAKE로 판정되었는가?"라는 질문에 시각적이고 정량적인 방식으로 답합니다. SHAP 폭포 차트로 각 탐지 모듈의 기여도를 수치화하고, Grad-CAM++ 히트맵으로 이미지의 어느 픽셀이 판정에 결정적 영향을 미쳤는지를 시각화합니다. 한국 AI 기본법 제15조의 AI 판단 근거 설명 의무를 직접적으로 충족시키는 핵심 도구입니다.

화면 구성

- SHAP 폭포 차트: 각 모듈의 판정 기여도 (양방향 바 차트)
- Grad-CAM++ 히트맵: 모델이 주목한 영역의 열지도 (파란색→노란색→빨간색)
- 베이지안 불확실성: 각 모듈의 판정 신뢰 구간 (평균, 표준편차, 90% CI)
- 모듈 기여도: 각 탐지 모듈의 기여 비율 (바 차트)

그래프 판독법

SHAP 폭포 차트: 기저선(Base=0.5)에서 출발하여 각 모듈이 FAKE(+) 또는 REAL(-) 방향으로 얼마나 밀었는지를 보여줍니다. 최종 도달점이 판정 결과입니다.

Grad-CAM++ 히트맵: 파란색→초록→노란→빨간색으로 의심 강도가 높아집니다. 빨간색이 얼굴 경계에 집중되면 합성 흔적의 강력한 증거입니다.

베이지안 불확실성: 신뢰구간이 좁을수록 해당 모듈의 판단이 확실하고, 신뢰구간이 0.5를 포함하는 모듈의 결과는 판정 근거로 사용하기 어렵습니다.

용어 해설

용어	설명
XAI(eXplainable AI)	AI의 판단 과정을 인간이 이해할 수 있도록 설명하는 기술

Grad-CAM++	신경망의 기울기(Gradient)를 활용하여 중요 영역을 시각화하는 기법
신뢰구간(Confidence Interval)	실제 값이 존재할 것으로 예상되는 범위

7.5 모델 핑거프린터

메뉴 경로: 사이드바 > 포렌식 도구 > 모델 핑거프린터 | 경로: /model-fingerprinter

기능: AI 생성 콘텐츠가 어떤 생성 모델(Sora, Kling, Runway, SDXL 등)로 만들어졌는지를 역추적하는 6단계 핑거프린팅 파이프라인입니다. 노이즈 추출 → 주파수 분석 → 인과 패턴 → 시간적 특성 → 위험도 평가 → 최종 귀속의 6단계를 거쳐 점진적으로 후보 모델 범위를 좁혀 나갑니다. 500개 이상의 AI 생성 모델 레지스트리와 연계합니다.

화면 구성

- 귀속 결과: 추정된 생성 모델명과 확신도
- 6 단계 파이프라인 결과: 각 분석 단계의 점수
- 후보 모델 순위: 상위 5 개 후보 모델과 유사도

7.6 T-GD 출처 분석

메뉴 경로: 사이드바 > 포렌식 도구 > T-GD 출처 분석 | 경로: /tgd-attribution

기능: 전이학습 기반 생성 탐지(Transfer Learning for Generative Detection, T-GD) 모듈로, 500개 이상의 AI 생성 모델 데이터베이스를 기반으로 콘텐츠의 출처를 정밀하게 특정합니다. 이진 탐지와 다중 클래스 귀속을 두 개의 독립 헤드로 동시에 수행하는 이중 헤드 아키텍처를 채택합니다. AUROC 95% 이상의 탐지 성능을 목표로 학습되었으며, 탐지 결과에 법적 증거력 점수(LES)를 산출하여 수사기관 제출 기준(75점 이상)을 충족하는지 자동으로 판단합니다.

화면 구성

- 탐지 헤드(Detection Head): 합성/진본 판별 결과 (AUROC 95%+)
- 귀속 헤드(Attribution Head): 500 개 모델 중 출처 추정 결과
- 법적 증거력 점수(LES): 0~100 점 (수사기관 제출 기준: 75 점 이상)
- 모델 레지스트리: 검색 가능한 500 개 모델 데이터베이스

LES 점수 구성

구성 요소	최대 점수	설명
Detection 확신도	30점	T-GD Detection Head의 판별 확신도
Attribution 확신도	35점	출처 모델 특정의 확신도
Top-1/2 마진	20점	1순위와 2순위 후보 간 점수 격차
다중 합의	15점	기존 모듈과의 판정 일치율

용어 해설

용어	설명
T-GD	Transfer Learning for Generative Detection — 전이학습을 활용한 AI 생성물 탐지 기술
LES(Legal Evidence Score)	법적 증거력 점수. 수사기관에 증거로 제출할 때의 신뢰도 지표
AUROC	Area Under the ROC Curve — 분류 모델의 성능 지표. 1.0에 가까울수록 우수

---

8. 방어 및 분류 섹션

8.1 3-Class 분류

메뉴 경로: 사이드바 > 방어 & 분류 > 3-Class 분류 | 경로: /three-class

기능: 미디어를 REAL(진본) / FAKE(위조) / ANTI-FORENSIC(반포렌식)의 세 가지 클래스로 분류합니다. 기존 이진 분류에서는 감지되지 않는 반포렌식 조작 유형을 별도 클래스로 분리하여, 탐지 회피를 시도하는 고급 공격에 대한 조기 경보 기능을 제공합니다. 탐지 회피 시도가 감지된 경우 즉각적인 경고와 함께 ADAG 레드팀 분석을 권고합니다.

용어 해설

용어	설명
반포렌식(Anti-Forensic)	딥페이크 탐지를 회피하기 위해 의도적으로 조작 흔적을 제거하는 기법

8.2 사전 방어

메뉴 경로: 사이드바 > 방어 & 분류 > 사전 방어 | 경로: /proactive-defense

기능: 원본 미디어에 비가시적 워터마크(Invisible Watermark)를 삽입하여, 향후 딥페이크로 변조되더라도 원본 여부를 과학적으로 증명할 수 있게 하는 사전적 보호 도구입니다. PGD, FGSM 등 6종의 적대적 방어(Adversarial Defense) 기법을 지원하며, PSNR 40dB 이상의 고품질 설정을 적용하면 육안으로 워터마크 삽입 전후를 구별할 수 없는 수준의 품질이 유지됩니다.

용어 해설

용어	설명
적대적 방어(Adversarial Defense)	AI 공격에 대응하여 원본을 보호하는 기법
PGD/FGSM	적대적 섭동(Perturbation)을 생성하는 알고리즘
PSNR	원본과 방어 처리된 이미지의 품질 차이를 측정하는 지표. 높을수록 품질 손실이 적음

---

9. 실시간 및 인프라 섹션

9.1 실시간 모니터

메뉴 경로: 사이드바 > 실시간 & 인프라 > 실시간 모니터 | 경로: /realtime-monitor

기능: 웹캠 또는 RTSP 스트림을 실시간으로 모니터링하여 딥페이크 영상이 감지되면 즉시 알림을 발생시킵니다. 화상 회의, 라이브 방송, 실시간 심문·조사 상황 등 즉각적인 검증이 필요한 환경에서 사용합니다. 초당 5~30프레임을 실시간으로 분석하며, 처리 지연(Latency)은 통상 200~500ms 내에서 유지됩니다. 감지 이벤트 발생 시 MoltBot(Telegram) 또는 이메일을 통한 즉각 알림을 지원합니다.

화면 구성

- 탐지 포인트 타임라인: 시간축에 따른 탐지 이벤트 표시
- 알림(Alert): 심각도별 알림 목록
- 프레임 메트릭: 실시간 분석 중인 프레임의 품질 지표
- 뷰티 필터 분석: SNS 뷰티 필터 적용 여부 탐지

9.2 모델 증류

메뉴 경로: 사이드바 > 실시간 & 인프라 > 모델 증류 | 경로: /model-distillation

기능: 서버급 GPU에서 실행되는 대형 탐지 모델을 경량화하여 모바일 기기, 엣지 장비, 임베디드 시스템에서도 딥페이크 탐지가 가능하도록 최적화하는 모델 압축 도구입니다. 교사(Teacher) 모델의 지식을 학생(Student) 모델로 전달하는 증류 기법과 INT8 양자화를 결합하여 모델 크기를 최대 10분의 1로 줄이면서도 탐지 정확도는 90% 이상 유지합니다. ONNX 형식으로 내보내기하여 TensorRT, OpenVINO, TFLite 등 다양한 추론 엔진에 배포할 수 있습니다.

용어 해설

용어	설명
모델 증류(Model Distillation)	대형 AI 모델의 지식을 소형 모델로 전달하여 경량화하는 기법
ONNX	다양한 AI 프레임워크 간 모델을 호환시키는 개방형 표준
INT8 양자화	모델의 정밀도를 낮추어 속도를 높이는 기법

---

10. 에이전트 AI 섹션

이 섹션은 TruthLens의 멀티 에이전트 AI 시스템을 관리하고 모니터링하는 고급 워크스페이스입니다. 일반 사용자는 이 섹션을 사용하지 않아도 분석에 지장이 없으나, 시스템의 내부 동작을 이해하고 튜닝하고자 할 때 활용합니다.

10.1 멀티에이전트 프레임워크

메뉴 경로: 사이드바 > 에이전트 AI > 멀티에이전트 프레임워크 | 경로: /multi-agent

기능: TruthLens에 통합된 8개 멀티에이전트 AI 프레임워크의 상태와 성능을 실시간으로 모니터링하고 관리합니다. 8개 프레임워크는 LangGraph(그래프 기반 워크플로우), CrewAI(역할 기반 전문가 팀), AutoGen(적대적 토론), 그리고 5개의 특화 에이전트 파이프라인으로 구성됩니다. 각 프레임워크는 독립적으로 판정을 산출하고, 최종 결과는 이들의 앙상블로 결정됩니다.

화면 구성

- 프레임워크 상태 카드: 각 프레임워크의 활성화/비활성화 상태, 호출 횟수, 평균 응답 시간
- 에이전트 트리: 20 개 에이전트의 계층 구조 시각화
- 가드레일 현황: AI 안전 장치의 통과/차단 비율

용어 해설

용어	설명
멀티에이전트(Multi-Agent)	여러 AI 에이전트가 협력하여 하나의 결론을 도출하는 시스템
LangGraph	그래프 기반 AI 워크플로우 오케스트레이션 프레임워크
CrewAI	여러 AI 전문가가 역할을 분담하여 토론하는 프레임워크
가드레일(Guardrail)	AI의 출력이 안전하고 정확한지를 검증하는 안전 장치

10.2 에이전트 설정

메뉴 경로: 사이드바 > 에이전트 AI > 에이전트 설정 | 경로: /agent-settings

기능: 각 에이전트의 활성화/비활성화, 사용할 LLM 모델 할당, 주요 파라미터(Temperature, Context Window, Max Tokens 등)를 세밀하게 조정합니다. 분석 목적과 환경에 따라 특정 에이전트만 선택적으로 활성화함으로써 분석 속도와 정확도의 균형을 최적화할 수 있습니다.

10.3 AutoGen 토론

메뉴 경로: 사이드바 > 에이전트 AI > AutoGen 토론 | 경로: /autogen-debate

기능: 판정이 모호한 UNCERTAIN 케이스에 대하여, 검찰(Prosecutor), 변호인(Defender), 판사(Judge) 역할의 AI 에이전트가 적대적으로 토론하며 결론을 도출하는 과정을 실시간으로 시각화합니다. 토론 히스토리는 전문가 검토나 법적 문서화에 그대로 활용될 수 있으며, 각 에이전트의 발언이 어떤 탐지 모듈의 결과를 근거로 했는지도 함께 표시됩니다.

10.4 LLM / 추론 엔진

메뉴 경로: 사이드바 > 에이전트 AI > LLM / 추론 엔진 | 경로: /llm-settings

기능: TruthLens의 AI 추론 엔진 전반을 관리하는 인프라 설정 도구입니다. Ollama(로컬 추론)와 vLLM(고성능 서버 추론) 중 추론 엔진을 선택하고, 텍스트 이해에 사용되는 LLM, 이미지 분석에 사용되는 VLM(Vision Language Model), 의미 검색에 사용되는 임베딩 모델을 각각 독립적으로 설정합니다. 온프레미스 환경에서 완전한 데이터 주권을 유지하면서 고성능 추론을 구현하는 데 필수적인 설정 화면입니다.

주요 설정

- VLM 모델: 비전 분석에 사용되는 모델 (기본: llama3.2-vision)
- Temperature: AI 응답의 다양성 (낮을수록 일관적, 0.1 권장)
- Context Window: AI 가 한 번에 처리하는 텍스트의 최대 길이

10.5 ADAG 레드팀

메뉴 경로: 사이드바 > 에이전트 AI > ADAG 레드팀 | 경로: /adag-redteam

기능: TruthLens 자체의 탐지 시스템을 능동적으로 공격하여 취약점을 사전에 발견하는 자기 강화형 적대적 테스트 시스템입니다. BiologicalSignalInjector(생체신호 주입 공격), GANFingerprintDisruptor(GAN 지문 교란), TemporalManipulator(시간 도메인 조작), TextHumanizer(텍스트 인간화) 등 4종의 공격 에이전트가 탐지 회피를 시도합니다. BiologicalSignalInjector의 현재 회피율이 약 50%로 가장 높은 우선순위 취약점으로 식별되어 집중 개선이 진행 중입니다.

용어 해설

용어	설명
레드팀(Red Team)	시스템의 취약점을 찾기 위해 공격자 역할을 수행하는 팀
ADAG	Adaptive Defense-Attack Game — 방어와 공격이 반복적으로 학습하는

	프레임워크
DER(Detection Evasion Rate)	공격이 탐지를 회피한 비율. 낮을수록 탐지 시스템이 강건함

10.6 MC Fusion 시뮬레이션

메뉴 경로: 사이드바 > 에이전트 AI > MC Fusion 시뮬레이션 | 경로: /mc-simulation

기능: 몬테카를로 시뮬레이션을 통해 각 탐지 모달리티(시각, 오디오, 생체, A/V 동기화)의 최적 가중치 조합을 과학적으로 산출합니다. Grid Search, Optuna(베이지안 최적화), 차분 진화(Differential Evolution), 몬테카를로의 4가지 최적화 방법을 병렬로 실행하여 결과를 상호 검증합니다. 결과로 산출된 최적 가중치는 온톨로지 파이프라인의 Fusion Weights에 자동으로 반영될 수 있습니다.

화면 구성

- 최적화 방법별 결과: Grid Search, Optuna, 차분 진화, 몬테카를로의 4 가지 최적화 결과 비교
- 가중치 추천: 최적의 모달리티별 가중치 조합
- 신뢰도 등급: A~F 등급의 전체 시스템 신뢰도

용어 해설

용어	설명
몬테카를로 시뮬레이션	무작위 시행을 수만 번 반복하여 최적 해를 찾는 통계적 방법
가중치(Weight)	각 모달리티의 판정이 최종 결과에 기여하는 비율
F1 Score	정밀도와 재현율의 조화 평균. 분류 성능의 종합 지표

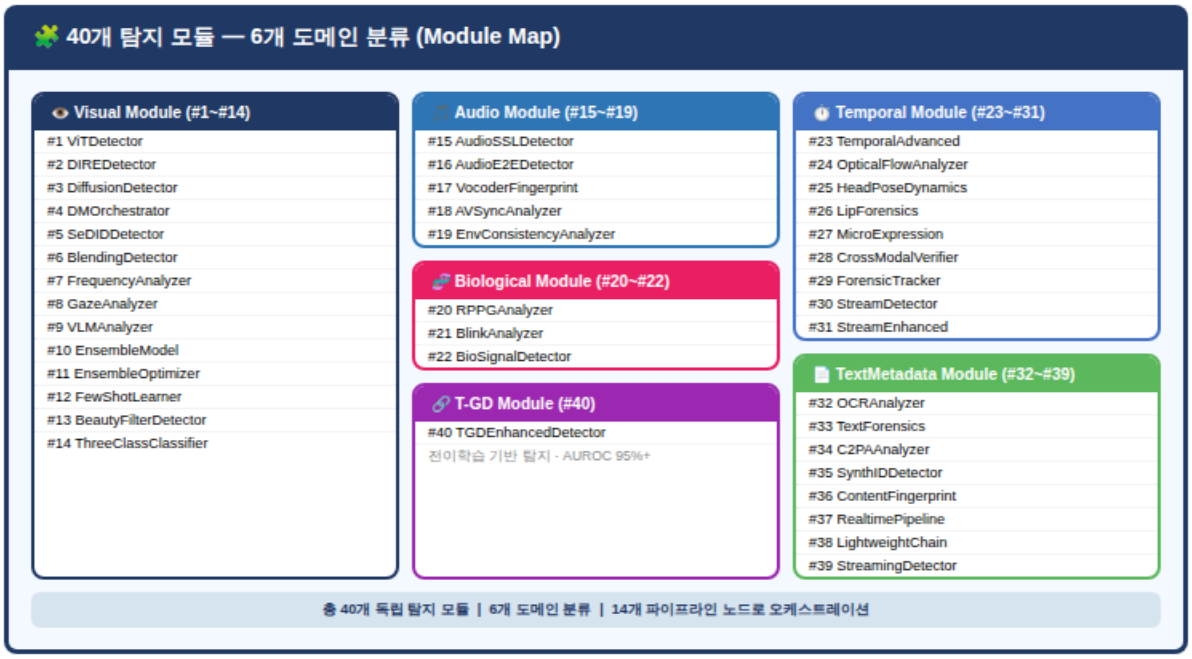
10.7 온톨로지 파이프라인

메뉴 경로: 사이드바 > 에이전트 AI > 온톨로지 파이프라인 | 경로: /ontology-pipeline

기능: TruthLens의 40개 탐지 모듈과 14개 파이프라인 노드의 전체 구조를 OWL(Web Ontology Language) 온톨로지로 형식화하고, 6개 SWRL(Semantic Web Rule Language) 추론 규칙으로 파이프라인의 실시간 무결성을 검증하는 고급 시스템 관리 도구입니다. SWRL 규칙은 "단일 모달리티만 활성화된 경우 신뢰도 60% 상한 적용"과 같은 운영 정책을 형식 언어로 표현하여, 정책 위반이 발생하면 즉각적인 알림과 자동 수정이 이루어집니다.

14개 파이프라인 노드 — 실행 순서

40개 탐지 모듈 — Module Map


40개 탐지 모듈 — 6개 도메인 분류 (Module Map)
Visual Module (#1~#14)
#1 ViTDetector
#2 DIREDetector
#3 DiffusionDetector
#4 DMOrchestrator
#5 SeDIDDetector
#6 BlendingDetector
#7 FrequencyAnalyzer
#8 GazeAnalyzer
#9 VLMAnalyzer
#10 EnsembleModel
#11 EnsembleOptimizer
#12 FewShotLearner
#13 BeautyFilterDetector
#14 ThreeClassClassifier
Audio Module (#15~#19)
#15 AudioSSLDetector
#16 AudioE2EDetector
#17 VocoderFingerprint
#18 AVSyncAnalyzer
#19 EnvConsistencyAnalyzer
Biological Module (#20~#22)
#20 RPPGAnalyzer
#21 BlinkAnalyzer
#22 BioSignalDetector
T-GD Module (#40)
#40 TGDEnhancedDetector
전이학습 기반 탐지 · AUROC 95%+
Temporal Module (#23~#31)
#23 TemporalAdvanced
#24 OpticalFlowAnalyzer
#25 HeadPoseDynamics
#26 LipForensics
#27 MicroExpression
#28 CrossModalVerifier
#29 ForensicTracker
#30 StreamDetector
#31 StreamEnhanced
TextMetadata Module (#32~#39)
#32 OCRAnalyzer
#33 TextForensics
#34 C2PAAnalyzer
#35 SynthIDDetector
#36 ContentFingerprint
#37 RealtimePipeline
#38 LightweightChain
#39 StreamingDetector
총 40개 독립 탐지 모듈 | 6개 도메인 분류 | 14개 파이프라인 노드로 오케스트레이션

SWRL 추론 규칙 현황


판정 기준 & SWRL 온톨로지 추론 규칙
판정 임계값 (Verdict Threshold)
FAKE
위조 판정
확률 ≥ 0.65
UNCERTAIN
판단 유보
0.25 ~ 0.65
REAL
진본 판정
확률 ≤ 0.25
신뢰도 상한 규칙 (Confidence Cap)
활성 모달리티 수
최대 신뢰도
의미
1개 (시각만)
60%
한 가지 관점만으로는 확신하기 어려움
2개
75%
두 관점에서 교차 검증되었으나 완전하지 않음
3개 이상
100%
다각적 교차 검증이 이루어짐
SWRL 온톨로지 추론 규칙 (6개)
SWRL-1 단일 모달리티 의존
1개 모달리티 활성 시 신뢰도 60% 상한 자동 적용
SWRL-2 데이터 순서 위반
파이프라인 데이터 전달 순서 오류 자동 감지
SWRL-3 VLM 환각 탐지
비전 모델 균일 점수 이상 감지 → sigmoid 보정
SWRL-4 순환 의존성
모듈 간 순환 참조 감지 → 무한 루프 방지
SWRL-5 모달리티 누락
필수 분석 모듈 비활성화 시 경고 발생
SWRL-6 가중치 합산 오류
가중치 합계 100% 불일치 시 자동 정규화

규칙	감지 대상	의미
SWRL-1	단일 모달리티 의존	1개 모달리티만 활성화되면 신뢰도를 60%로 제한
SWRL-2	데이터 순서 위반	파이프라인에서 데이터가 잘못된 순서로 전달되는 오류
SWRL-3	VLM 환각	비전 모델이 모든 카테고리에 동일 점수를 부여하는 이상
SWRL-4	순환 의존성	모듈 간 순환 참조 (무한 루프 위험)
SWRL-5	모달리티 누락	필수 분석 모듈이 비활성화된 상태
SWRL-6	가중치 오류	모달리티 가중치 합계가 100%가 아닌 상태

Fusion Weights 탭: 각 모달리티의 가중치를 바 차트로 표시하고, 합계가 100%인지 검증합니다.

용어 해설

용어	설명
온톨로지(Ontology)	지식을 체계적으로 분류하고 관계를 정의하는 형식 체계
OWL(Web Ontology Language)	웹 표준 온톨로지 기술 언어
SWRL	Semantic Web Rule Language — 온톨로지 위에 추론 규칙을 정의하는 언어
파이프라인 노드(Pipeline Node)	분석 과정의 각 단계를 수행하는 처리 단위

---

11. 관리 섹션

관리자(ADMIN) 역할의 사용자만 접근할 수 있습니다.

- 관리자 대시보드 (/admin): 사용자 목록 관리, 역할 변경, 비밀번호 초기화
- 조직 관리 (/admin/organizations): 조직 생성/수정, 요금제(Tier) 변경, 사용량 쿼터 설정
- API 키 관리 (/admin/api-keys): API 키 발급/폐기, 만료 일자 설정

---

12. 시스템 섹션

- 설정 (/settings): 언어 선택(한국어/영어/일본어/중국어), 사용자 프로필, 비밀번호 변경
- API 문서 (/api-docs): TruthLens REST API 의 전체 사양을 Swagger UI 로 제공합니다. 외부 시스템과의 연동 개발 시 참고합니다.

---

13. 판정 기준 및 해석 가이드

13.1 판정 임계값

조건	판정
최종 확률 ≥ 0.65	FAKE (위조)
최종 확률 ≤ 0.25	REAL (진본)
0.25 < 최종 확률 < 0.65	UNCERTAIN (판단 유보)

13.2 신뢰도 상한 규칙

활성 모달리티 수	최대 신뢰도	의미
1개 (시각만)	60%	한 가지 관점만으로는 확신하기 어려움
2개	75%	두 가지 관점에서 교차 검증되었으나 완전하지 않음
3개 이상	100%	다각적 교차 검증이 이루어짐

13.3 UNCERTAIN 판정 시 권장 조치

1. 영상 품질 확인: 원본 파일이 고해상도인지 확인합니다. 과도하게 압축된 파일은 분석 정확도가 떨어집니다.

2. 정밀 모드 재분석: 신속/표준 모드로 분석했다면 정밀(Precise) 모드로 재분석합니다.
3. 전문가 리뷰: XAI 심층 분석 결과를 포렌식 전문가에게 검토 요청합니다.
4. 원본 대조: 가능하다면 원본 파일과 대조합니다.

---

14. 부록: 딥페이크 위협과 TruthLens의 사회적 책무

14.1 딥페이크 위협의 현실

2025~2026년 현재, 딥페이크 기술의 대중화로 인한 피해가 전례 없는 규모로 확산되고 있습니다.

금융 분야

- 기업 CEO 의 화상 회의 영상을 딥페이크로 위조하여 6,200 만 달러(약 830 억 원)를 송금하도록 유도한 사건이 보고되었습니다 (2024 년 홍콩).
- 보이스피싱에 AI 음성 합성이 활용되어, 가족의 목소리를 완벽하게 모사한 사기가 급증하고 있습니다.

정치/사회

- 선거 기간 중 후보자의 딥페이크 영상이 유포되어 민주주의적 의사결정을 왜곡하는 사례가 다수 발생하였습니다.
- 딥페이크 포르노그래피 등 개인의 인격권을 침해하는 범죄가 사회적 문제로 대두되고 있습니다.

방송/미디어

- 가짜 뉴스 영상이 SNS 를 통해 급속히 확산되어 사회적 혼란을 야기하는 사례가 빈번합니다.
- 유명인을 사칭한 딥페이크 광고가 소비자 피해를 유발하고 있습니다.

14.2 대한민국의 법적 대응

- AI 기본법 제 15 조: AI 시스템의 판단에 대한 설명 의무를 규정합니다. TruthLens 는 모든 판정에 XAI 시각화와 감사 로그를 제공하여 이 의무를 충족합니다.
- 정보통신망법: 허위 영상물의 유포를 금지하고 있으며, TruthLens 의 분석 보고서는 수사기관의 증거 자료로 활용될 수 있습니다.
- 개인정보보호법: TruthLens 는 분석 대상 미디어를 분석 완료 후 즉시 삭제하며, 개인정보를 별도 저장하지 않습니다.

14.3 TruthLens의 사회적 책무

TruthLens는 단순한 기술 도구가 아닌, 사회 안전망의 핵심 인프라로서의 사명을 가지고 개발되었습니다.

진실 보호: 영상, 음성, 문서의 진위를 과학적으로 검증하여 허위 정보로부터 시민을 보호합니다.

피해 예방: 금융 사기, 보이스피싱, 여론 조작 등 딥페이크를 악용한 범죄를 사전에 탐지하여 선의의 피해자 발생을 방지합니다.

증거 보전: SHA-256 해시 체인으로 보호되는 감사 로그와 법적 증거력(LES) 점수를 제공하여, 수사와 재판에서 활용 가능한 과학적 증거를 확보합니다.

투명한 판단: 모든 판정의 근거를 XAI 시각화로 제공하여, AI의 "블랙박스" 문제를 해소하고 인간의 최종 판단을 지원합니다.

지속적 진화: ADAG 레드팀 시스템을 통해 스스로의 취약점을 발견하고 개선하는 자기강화 메커니즘을 갖추고 있어, 날로 교묘해지는 딥페이크 기술에 지속적으로 대응합니다.

> *"기술은 양날의 검입니다. AI가 만들어낸 위협에는 AI로 대응해야 합니다. TruthLens는 디지털 시대의 진실을 지키는 방패이며, 선의의 피해자가 더 이상 발생하지 않도록 하는 것이 우리의 사회적 책무입니다."*

문서 끝

본 가이드는 TruthLens v4.4.0 기준으로 작성되었습니다.

Brian Lee | AI R&D Center | A3 Security Co.,Ltd. | April 14, 2026