ANVI

- 안 봐도 비디오 -

요구사항 명세서

Software Requirements Specification

프로젝트 주제 : 모바일 온-디바이스 초 단위 Vision AI & VLM 추론 솔루션

개정 이력

버전	변경일자	개정 내용	작성자
0.0.1	2026-04-07	초안 작성 (1차 멘토링 반영)	박형주

0. 프로젝트 개요

주제 모바일 온-디바이스의 초 단위 Vision & VLM 추론 솔루션 개발

사용 대상 프록토매틱(Proctormatic) 고객 — 온라인 시험 응시자 및 감독관

개발 형태 프록토매틱 앱의 애드온(Add-on) 형식 기술 검증 단계 → 신규 앱 런칭 마인드로 개발

배경

기존 AI 서비스는 대부분 서버로 데이터를 보내 추론 결과를 받는 구조이며, 특히 영상 기반 AI 서비스는 개인정보 침해 위험이 상존한다. 이를 해결하기 위해 추론을 온-디바이스(Android 기기)에서 직접 처리하여 개인정보 노출을 최소화한다.

목표

초소형 Vision AI & VLM 모델을 Android 온-디바이스 환경에서 실행한다. 앱 실행 시 VLM이 메모리에 로딩되고, 감독 시작 시 카메라 영상을 0.5초마다 추론하여 의심 시나리오 감지 시 서버로 로그 및 이미지를 전송한다.

기술 스택

구분	내용
Frontend (Android)	Kotlin — Android Native
Backend	Spring Boot (Java)
AI Infra	TFLite (TensorFlow Lite) — On-Device 추론 프레임워크
DB	PostgreSQL — 감독 이력 / 유저 데이터 저장
이미지 저장	Amazon S3 — 감지 이미지 업로드 저장소
배포 환경	SSAFY 제공 EC2 인스턴스
서버-앱 통신	SSE (Server-Sent Events) — 헬스 체크 및 실시간 연결
AI 가중치 파일	에딘트 제공 LiteTF weight 파일 (.tflite)

1. 기능적 요구사항

1.1 요구사항 정의 규칙 (ID 규칙)

ID 접두사	적용 도메인
REQ-APP-XXX	Android 앱 / 온-디바이스 AI 실행 관련 기능 (핵심 도메인)
REQ-AI-XXX	AI 추론 모델 로딩 · 추론 · 시나리오 판단 관련 기능
REQ-SRV-XXX	백엔드 서버 API · SSE · 데이터 저장 관련 기능
REQ-MGT-XXX	시험 관리 웹 사이트 · 감독 이력 조회 관련 기능 (가능하면)

1.2 기능 요구사항 상세

1.2.1 Android 앱 / 온-디바이스 AI (App Domain)

표 1. Android 앱 기능적 요구사항

Req ID	기능명	상세 설명
REQ-APP-001	VLM 모델 로딩	• 앱 실행 시 에딘트가 제공하는 .tflite 가중치 파일을 메모리에 로딩한다. • 모델 로딩이 완료된 이후에 감독 세션 시작이 가능하다.
REQ-APP-002	실시간 카메라 영상 수신	• 감독 세션 시작 시 기기 전면 카메라로부터 실시간 영상 스트림을 수신한다. • 영상 스트림은 Vision AI 추론 모듈로 지속 전달된다.
REQ-APP-003	온-디바이스 AI 추론 (0.5초 주기)	• Vision AI는 카메라 영상 프레임을 0.5초마다 1장씩 추론한다. • 정확성이 속도보다 우선한다 — 초당 1장이어도 정확성이 가장 중요하다. • 추론 결과는 JSON 형식으로 반환된다.
REQ-APP-004	의심 물체 감지 → VLM 상황 판단	• Vision AI가 의심스러운 물체(시험 금지 품목)를 감지한 경우에만 VLM이 상황을 2차 판단한다. • VLM의 판단 결과가 지정 시나리오에 해당하는 경우에만 서버로 로그를 전송한다. • **의심 행위 감지** : 정상 행위와 부정 행위가 아닌 경우가 계속 반복될 경우를 의미한다. • **부정 행위 감지** : 명백한 시험 부정 행위를 감지하는 경우이다
REQ-APP-005	추론 결과 서버 전송	• 특정 시나리오 감지 시에만 추론 결과(JSON 로그)를 백엔드 서버로 전송한다. • 정상 상황에서는 서버 전송을 생략하여 네트워크 부하를 최소화한다.

Req ID	기능명	상세 설명
REQ-APP-006	감지 이미지 캡처 및 전송 (추가 기능)	• 특정 시나리오 감지 시 해당 프레임 이미지를 캡처하여 서버(S3)로 업로드한다. • 이미지 전송은 시나리오 감지 시점에만 수행하며, 항상 전송하지 않는다.
REQ-APP-007	얼굴 블러 처리 (심화 기능)	• 서버로 전송하는 이미지에서 얼굴 영역을 블러 처리하여 개인정보 위험을 최소화한다. • 블러 처리는 기본 기능 완성 후 심화 구현 항목이다.
REQ-APP-008	환경 점검 (Pre-check)	• 시험 시작 전, 카메라 각도와 조명이 AI 추론에 적합한지 확인하는 가이드라인(UI 오버레이)을 제공한다.
REQ-APP-009	실시간 상태 인디케이터	• 화면 구석에 작은 인디케이터를 배치하여, 현재 AI 모니터링이 정상 작동 중인지(On-Air) 응시자가 인지하게 한다.
REQ-APP-010	시험 일시 정지/재개 대응	• 전화 수신이나 앱 이탈 등 예외 상황 발생 시 감독관에게 즉시 알림을 보내고, 해당 시점의 추론 상태를 보존한다.
REQ-APP-011	네트워크 불안정 알림	• 서버와의 SSE 연결이 끊기거나 이미지 업로드 실패 시, 응시자에게 오프라인 상태임을 알리고 로컬 저장을 시도한다 ?

1.2.2 AI 추론 관련 (AI Domain)

표 2. AI 추론 기능적 요구사항

Req ID	기능명	상세 설명
REQ-AI-001	TFLite Vision AI 추론	• TFLite 프레임워크를 사용하여 Android 온-디바이스에서 Vision AI 추론을 실행한다. • Vision AI는 카메라 프레임에서 시험 금지 품목 등 의심 물체를 파악한다.
REQ-AI-002	VLM 상황 판단	• Vision AI의 의심 물체 감지 결과를 입력받아 VLM이 시나리오를 2차 판단한다. • VLM 판단 결과를 JSON 형식으로 앱 상위 레이어에 반환한다.
REQ-AI-003	NPU / GPU / CPU 자동 선택 (추가 기능)	• 기기 상태에 따라 NPU → GPU → CPU 순서로 추론 가속 자원을 자동 선택한다. • TFLite Delegate API를 활용하여 최적의 추론 자원을 동적으로 선택한다.

1.2.3 백엔드 서버 관련 (Server Domain)

표 3. 백엔드 서버 기능적 요구사항

Req ID	기능명	상세 설명
REQ-SRV-001	SSE 헬스 체크	• SSE(Server-Sent Events)를 통해 백엔드 서버와 Android 앱 간의 연결 상태를 실시간으로 확인한다. • 앱은 감독 세션 시작 전 서버 연결 정상 여부를 SSE로 검증한다.
REQ-SRV-002	시험 접속 권한 관리	• 백엔드는 응시자의 시험 접속 권한(인증)을 관리한다. • 권한이 없는 기기/사용자의 접근을 차단한다.
REQ-SRV-003	시험 접속 경로(URL) 관리	• 각 시험 세션에 대한 고유 접속 URL을 생성 및 관리한다. • URL 만료 처리 및 유효성 검증 기능을 포함한다.
REQ-SRV-004	추론 결과 로그 수신 및 저장	• Android 앱에서 전송한 추론 결과(JSON)를 수신하여 PostgreSQL에 저장한다. • 감독 이력(시험 ID, 응시자, 감지 일시, 감지 내용)을 포함하여 저장한다.
REQ-SRV-005	이미지 수신 및 S3 저장	• 앱에서 전송한 감지 이미지를 수신하여 Amazon S3에 업로드하고 저장 경로를 DB에 기록한다. • 이미지 업로드는 감지 이벤트와 연결하여 추론 로그와 함께 관리한다.
REQ-SRV-006	사용자 인증 (간소화)	• 로그인 기능을 구현하되, 과도한 시간 투자를 피하고 핵심 감독 기능에 집중한다. • JWT 기반 간단한 토큰 인증으로 구현한다.

1.2.4 시험 관리 웹 (Management Domain)

표 4. 시험 관리 기능적 요구사항 (선택 구현)

Req ID	기능명	상세 설명
REQ-MGT-001	시험 관리 웹 사이트	• 감독관이 시험 생성, 응시자 목록, 감지 이력을 조회할 수 있는 웹 인터페이스를 제공한다.
REQ-MGT-002	감지 이력 조회	• 감독 세션별로 의심 감지 이벤트 목록(일시, 종류, 이미지 링크)을 조회할 수 있다.
REQ-MGT-003	실시간 이상 탐지 대시보드	• 여러 응시자 중 VLM이 '부정 행위' 또는 '의심 행위'로 판단한 로그가 올라온 사용자를 리스트 최상단에 배치한다.
REQ-MGT-004	의심 구간 타임라인 뷰	• 특정 응시자의 상세 페이지에서 AI가 감지한 시점의 로그와 이미지를 시간 순서대로 나열하여 전후 사정을 파악하게 한다.
REQ-MGT-005	부정 행위 확정/반려 처리	• AI가 탐지한 이벤트에 대해 감독관이 직접 '부정 행위 확정' 또는 '오탐(False Positive)'으로 라벨링할 수 있는 기능을 제공한다.
REQ-MGT-006	감독 리포트 생성	• 시험 종료 후, 전체 응시자의 탐지 횟수 및 주요 의심 사례를 요약한 통계 리포트를 PDF 또는 엑셀로 추출한다.

2. 비기능적 요구사항

2.1 요구사항 정의 규칙 (ID 규칙)

ID 접두사	적용 도메인
NFR-PER-XXX	Performance / 성능 : 추론 정확성, 추론 주기, 서버 응답 속도 등
NFR-SEC-XXX	Security / 보안 : 시험 접속 권한, 이미지 개인정보 보호, 통신 암호화 등
NFR-REL-XXX	Reliability / 신뢰성 : 감독 세션 안정성, 로그 무결성, 모델 로딩 안정성 등
NFR-MNT-XXX	Maintainability / 유지보수 : CI/CD, 코드 컨벤션, 로깅 및 모니터링 등
NFR-CMP-XXX	Compatibility / 호환성 : Android 버전, 기기 NPU/GPU/CPU 호환성 등

2.2 비기능적 요구사항 상세

2.2.1 성능 (Performance & Accuracy)

표 5. 성능 요구사항

Req ID	분류	설명
NFR-PER-001	추론 정확성 우선	정확성이 속도보다 반드시 우선한다. 추론 주기가 초당 1장(1fps)이더라도 정확한 감지 결과를 반환해야 한다. (멘토 전달사항)
NFR-PER-002	추론 주기 목표	목표 추론 주기는 0.5초(2fps)이며, 기기 성능에 따라 최소 1초(1fps) 이상을 보장해야 한다.
NFR-PER-003	서버 API 응답 속도	백엔드 REST API는 95% 이상의 요청에 대해 500ms 이내에 응답해야 한다.
NFR-PER-004	이미지 전송 최소화	이미지 전송은 시나리오 감지 시점에만 수행하여 네트워크 대역폭 사용을 최소화한다.

2.2.2 보안 (Security & Privacy)

표 6. 보안 요구사항

Req ID	분류	설명
NFR-SEC-001	시험 접속 권한 제어	각 시험 세션에 고유 URL을 발급하고, 권한 없는 접근을 서버에서 차단한다.
NFR-SEC-002	통신 보안	Android 앱과 백엔드 서버 간의 모든 통신은 HTTPS/TLS를 의무 사용하여 패킷 스니핑을 방지한다.
NFR-SEC-003	이미지 개인정보 보호	서버로 전송하는 이미지에는 얼굴 블러 처리를 심화 목표로 적용하여 개인정보 노출을 최소화한다. 블러 미적용 시에도 S3 접근 권한을 제한하여 보호한다.
NFR-SEC-004	시크릿 키 관리	S3 키, DB 접속 정보 등 민감 정보는 환경 변수(.env) 또는 외부 비밀 관리 도구를 통해 주입하며 소스코드에 하드코딩하지 않는다.

2.2.3 신뢰성 (Reliability & Availability)

표 7. 신뢰성 요구사항

Req ID	분류	설명
NFR-REL-001	감독 세션 연결 안정성	SSE를 통해 서버와 앱의 연결 상태를 지속 모니터링하며, 연결이 끊길 경우 자동 재연결을 시도한다.
NFR-REL-002	로그 무결성	감지 이벤트 로그 저장은 트랜잭션으로 처리하여, 중간 오류 발생 시 롤백(Rollback)되어 데이터 손실이 없도록 한다.
NFR-REL-003	모델 로딩 안정성	TFLite 모델 로딩 실패 시 사용자에게 명확한 오류 메시지를 표시하며, 앱이 크래시되지 않도록 예외 처리한다.
NFR-REL-004	시스템 복구	Docker 컨테이너 환경에서 서버 프로세스 비정상 종료 시 자동 재시작(Restart Policy)하여 다운타임을 최소화한다.

2.2.4 유지보수 및 배포 (Maintainability & Deployability)

표 8. 유지보수 및 배포 요구사항

Req ID	분류	설명
NFR-MNT-001	CI/CD 자동화	Jenkins 또는 GitHub Actions를 활용하여 소스코드 병합 시 자동 빌드, 테스트, Docker 이미지 생성 및 EC2 배포 파이프라인을 구축한다.
NFR-MNT-002	코드 컨벤션	Kotlin/Java(Android), 백엔드 각 언어별 정적 분석 도구를 도입하여 일관된 코드 품질을 유지한다.

Req ID	분류	설명
NFR-MNT-003	로깅 및 모니터링	API 호출 로그, 에러 로그, AI 추론 결과 전송 로그를 KST 기준 타임스탬프와 함께 기록하며, 2~3일 이상 이슈 발생 시 즉시 멘토에게 보고한다. (멘토 전달사항)
NFR-MNT-004	DB 쿼리 최적화	N+1 문제를 지양하고 JOIN, 인덱스, 페이징 등을 활용하여 쿼리를 최적화한다. (멘토링 Q6 반영)

2.2.5 호환성 (Compatibility)

표 9. 호환성 요구사항

Req ID	분류	설명
NFR-CMP-001	Android 버전 호환성	Android 10(API 29) 이상의 기기에서 앱이 정상 구동되어야 한다. TFLite Delegate 사용 시 기기별 호환성을 검증한다.
NFR-CMP-002	NPU/GPU/CPU 자동 선택	추론 실행 시 기기 상태에 따라 NPU → GPU → CPU 순으로 Delegate를 자동 선택하여 최적 성능을 보장한다. (추가 기능)
NFR-CMP-003	TFLite 모델 포맷	에딘트가 제공하는 .tflite(LITE 변환) 가중치 파일 및 오픈소스 모델 링크를 사용하며, 모델 포맷 변경 시 팀 내 공유한다.