캡스톤디자인 과제 결과보고서

과제명	엣지 디바이스 환경에서의 이상 상황 실시간 탐지				
과제기간	2024.11.25. ~ 2025.04.20.				
지도교수	성명	최아영	학과	AI.소프트웨어학부	
	e-mail	aychoi@gachon.ac.kr	연락처	031-750-8656	
참여학생	학과	학번	학년	성명	
	AI.소프트웨어학부 인공지능전공	202235128	4	조재현	
	AI.소프트웨어학부 인공지능전공	202034104	4	최경민	
	AI.소프트웨어학부 인공지능전공	202235148	4	한해빈	
	AI.소프트웨어학부 인공지능전공	202235152	4	홍지민	

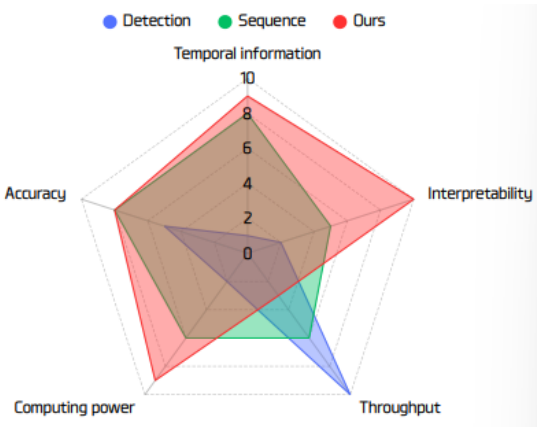
1. 과제 개요

최근 통계청의 2024년 사회조사에 따르면, 우리 국민 중 단지 28.9%만이 '안전하다'고 느끼고 있어 2년 전 대비 4.4%p 감소한 것으로 나타났습니다. 이는 범죄 및 화재 등 이상 상황 발생 시 즉각적인 대응 체계가 미흡하여 피해가 확대되고 있음을 시사합니다.

본 과제에서는 이러한 사회적 불만을 해소하기 위해 엣지 디바이스 기반 실시간 이상 상황 탐지 시스템을 개발하고, 감지된 위험 상황을 모바일 앱을 통해 사용자에게 즉시 알림으로 전달하는 통합 플랫폼을 제안합니다. 특히, 칼 사용처럼 일상과 위협이 공존하는 행위도 AI 모델이 맥락을 이해하여 오탐지를 최소화하고, 화재,폭행,침입 등 다양한 보안 이벤트를 정확히 구분할 수 있도록 설계하였습니다.

핵심 기술로는 기존의 객체 탐지 및 시퀀스 모델의 장점을 결합한 Vision-Language 기반 상황 인지형 탐지 모델을 도입하였으며, 이를 통해 단순 존재 확인을 넘어 '왜 위험한가'를 설명하는 자연어 리포팅 기능을 제공합니다. 실제 구현된 시스템은 Jetson Orin Nano와 같은 엣지 보드에서 네트워크 의존 없이 모든 연산을 수행하며, 이상 탐지 -> 상황 설명 -> 앱 알림으로 이어지는 End-to-End 워크플로우를 완성하였습니다.

아래 레이더 차트는 전통적 탐지 모델, 시퀀스 모델과 비교한 본 시스템의 성능을 시각화한 것으로, 해석 가능성, 이상 감지 정밀도, 시간 정보 반영 등 여러 항목에서 우수함을 확인할 수 있습니다. 이를 바탕으로 국민들이 체감할 수 있는 신속,정확,맥락,인지형 대응 체계를 구축함으로써, 보다 안전한 사회 실현을 목표로 합니다.

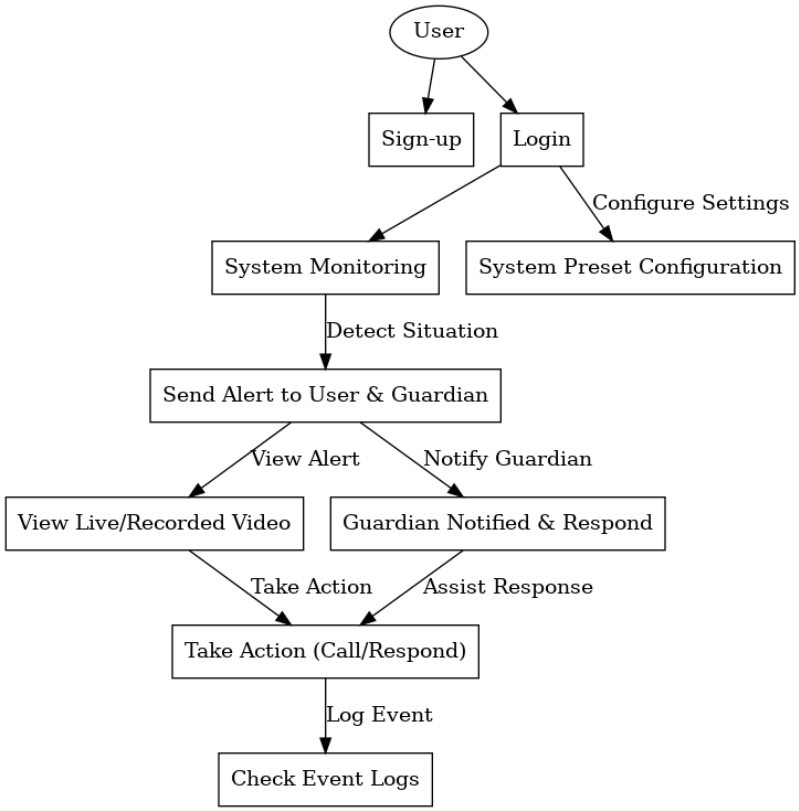


성능 비교 레이더 차트

2. 과제 진행 내용

2-1. USER SCENARIO

해당 과제를 시작하기 이전에, 시나리오를 구체적으로 설계해보았다.



유저 시나리오

2-2 Visual Language Model

이상 상황 탐지에서 Viusal Language Model은 영상 데이터를 분석해 상황을 묘사하고 분류하는 역할을 한다. 기존 객체 탐지 모델이 객체만 인식하는데 집중했다면, Visual Language Model은 상황의 맥락을 이해해 정상 상황과 비정상 상황을 구분한다.

예를 들어, 칼이 화면에 나타나더라도 마트에서 구매하는 장면은 정상으로, 위협하는 장면은 이상 상황으로 판단한다. Visual Language Model은 이렇게 이상 상황을 판단한 후 객체 탐지 모델에 원인 객체를 탐지하도록 지원하며, 시스템의 실시간성과 정확성을 높이는 핵심 역할을 수행한다.

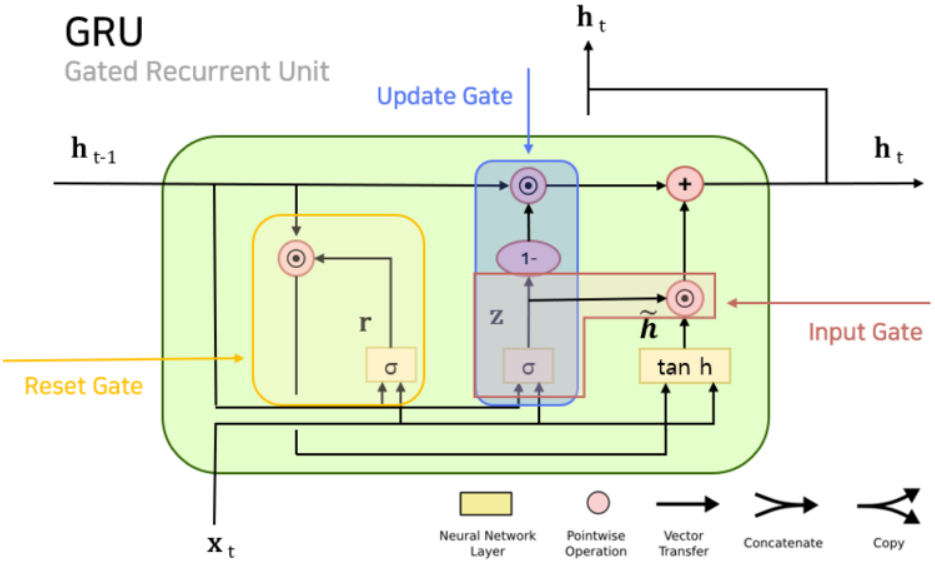
2-3 Risk Detection Model (Vision Encoder + GRU)l

엣지 디바이스에서의 실시간성을 고려하여 고성능 Vision-Language Model의 연속 사용을 최소화하고, 경량화된 시퀀스 기반 이상 탐지 모델을 먼저 적용함으로써 시스템의 효율성을 극대화하였다. 경량화된 시퀀스 기반 이상 탐지 모델을 먼저 적용함으로써 시스템의 효율성을 극대화하였다.

이상 상황 판단의 1차 필터링은 Vision Encoder와 GRU로 구성된 위험 탐지 모델이 담당하며, 이 모델은 연속된 프레임의 흐름 속에서 시간적 패턴의 이상 여부를 탐지한다. 구체적으로, MobileNetV2 기반 Vision Encoder가 각 프레임의 시각적 특징을 추출하고, 이를 GRU(Gated Recurrent Unit)가 시퀀스 형태로 받아들여 행동 패턴 및 맥락의 변화를 학습한다. 모델의 출력은 Anomaly Score(0~1)로 정량화되어, 일정 임계값을 초과할 경우 해당 구간을 이상 상황으로 로깅한다.

예를 들어, 반복적으로 쓰러지거나 급격한 움직임이 있는 프레임들이 연속적으로 감지될 경우 이를 폭행이나 기절 등 위험 신호로 판단한다.

이처럼 경량화된 이상 탐지 모델이 1차적으로 판단을 수행하고, 이후 대표 프레임만 Vision-Language Model에 전달함으로써 전체 시스템의 추론 속도와 자원 효율성을 동시에 확보하였다.

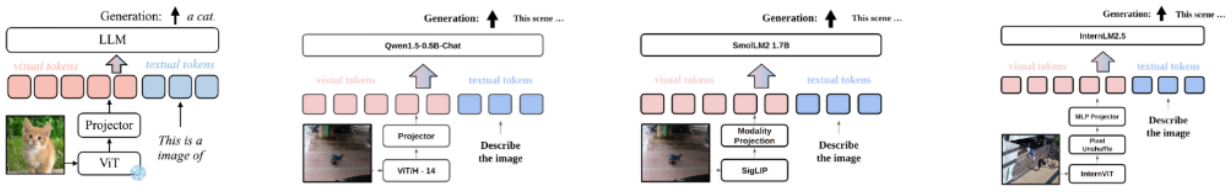


GRU 모델 구조도

2-4 Application Development

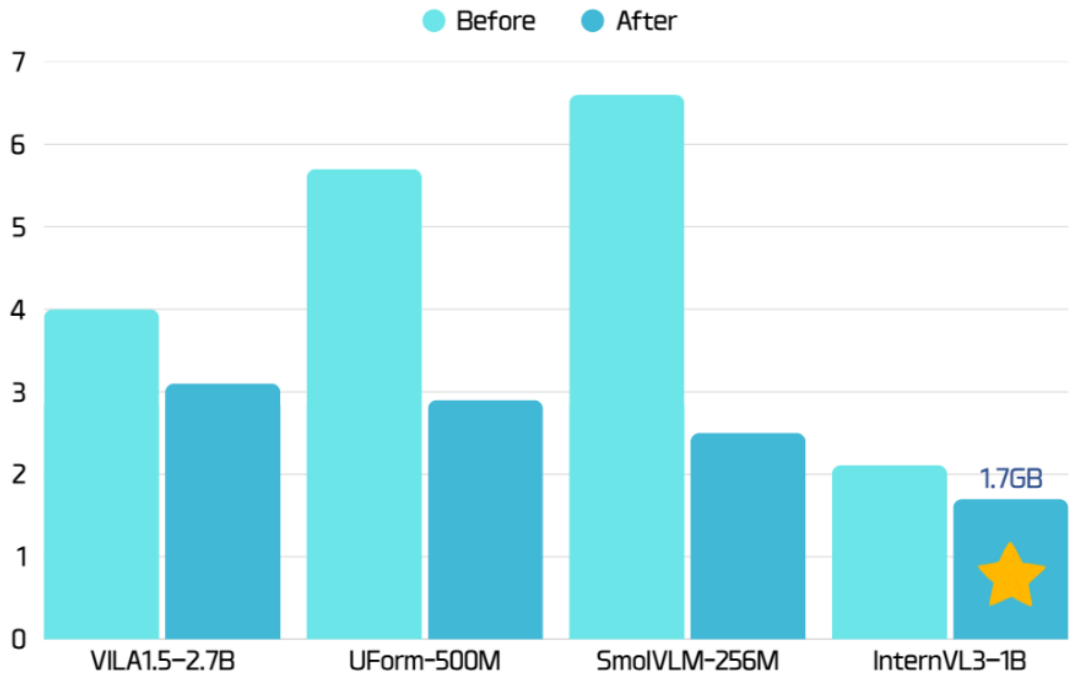
본 프로젝트는 엣지 디바이스 상에서 실시간 이상 상황 탐지 및 설명이 가능한 전체 파이프라인을 구성하였으며, 각 모듈은 Jetson Orin Nano 환경에 맞추어 최적화되었다. Vision Encoder 기반의 위험 탐지 모델과 경량 VLM을 연동하여, 효율적인 처리와 높은 설명력을 동시에 구현하였다.

a. InternVL3-1B to Jetson Orin Nano



후보 VLM 모델

엣지 디바이스인 Jetson Orin Nano는 메모리와 연산 성능이 제한적이기 때문에, 대형 VLM을 구동하기 위해 다양한 최적화 전략이 적용되었다.



VRAM 비교 표

최종적으로 선정된 InternVL3-1B 모델은 다양한 VLM 후보군(SmolVLM, UForm 등) 중에서 경량성, 실행 가능성, 설명 성능 측면에서 가장 우수한 모델로 평가되었다.

Jetson 보드의 메모리 제약을 해결하기 위해 다음과 같은 작업이 수행되었다 :

- Swap Memory 설정 : 일부 데이터를 디스크로 이동하여 메모리 공간을 확보하고, VLM 추론 안정성 확보
- Docker 환경 구축 : 다양한 AI 라이브러리 및 의존성 충돌을 방지하고, 독립된 어플리케이션 실행 환경 구성

이와 같은 최적화 과정을 통해, InternVL3-1B 모델이 Jetson Orin Nano 상에서 실시간 추론

가능하도록 성공적으로 배포되었다.

b. InternVL3 + Adpater

InternVL3는 사전 학습된 대형 멀티모달 vision-language 모델로, 일반적인 이미지-텍스트 매칭, 설명 등의 task에 우수한 성능을 보여주는 모델이다. 본 시스템에서는 이를 CCTV 기반의 이상 상황 탐지 및 설명이라는 도메인 특화 목적에 맞게 활용하기 위해, 기존 InternVL3 구조를 그대로 유지하되, Adapter 기반의 미세 조정 전략을 적용하였다. 이를 통해 모델 전체를 다시 학습하지 않고도 도메인 특화 설명 능력을 강화할 수 있었으며, 프레임 내 객체의 상태, 행동 맥락 등을 종합적으로 설명할 수 있는 기능이 향상되었다.

Adapter는 InternVL3의 Attention Layer 내부에 삽입되었으며, 기존 모델의 파라미터는 고정(Freeze)된 상태로 유지되어 연산량과 메모리 사용을 최소화할 수 있었다. 결과적으로 InternVL3는 단순한 객체 설명을 넘어, "사람이 위협적인 자세로 서 있음", "칼을 들고 특정 방향을 응시함"과 같은 복합적인 이상 상황에 대한 설명도 가능해졌다.

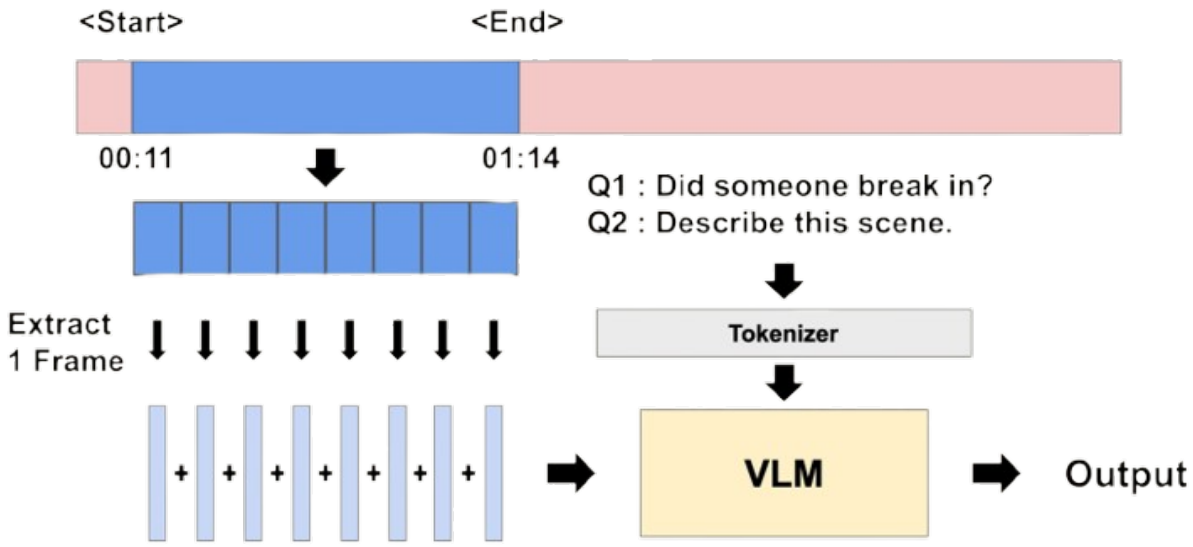
c. Pipeline

본 장에서는 학습된 이상 탐지 모델을 활용해 영상에서 이상 상황을 자동으로 식별 및 로깅하고, Vision-Lanuguage Model(VLM)을 이용해 해당 구간에 대한 자연어 설명을 생성하는 파이프라인을 상세히 다룬다. 먼저, 연속된 8프레임 단위로 이상 점수를 계산하고, 그 결과가 레이블과의 IoU기준 50% 이상을 만족하는 시퀀스를 이상 구간으로 판정한다. 각 이상 구간은 타임스탬프 형태로 저장되며, 이후 처리의 정확성을 위해 시작과 종료 지점을 명확히 기록한다.

이어서 판정된 이상 구간은 시간적으로 균등하게 8개의 파트로 분할되며, 각 파트에서 대표성이 높은 1개의 키 프레임을 추출한다. 이렇게 확보된 총 8장의 키 프레임은 VLM 입력을 위한 시각 정보로 활용된다. VLM에는 단순히 프레임만 전달되지 않고, 두 단계의 질의-응답 절차가 함께 이루어진다. 첫 번째 질문 “Did someone break in?”은 오탐 가능성을 낮추기 위해 이상 구간을 재검토하는 필터링 역할을 수행하며, 두 번째 질문 “Describe this scene.”은 8장의 키 프레임이 담고 있는 행동과 상황 전반을 자연어로 상세히 기술하도록 유도한다.

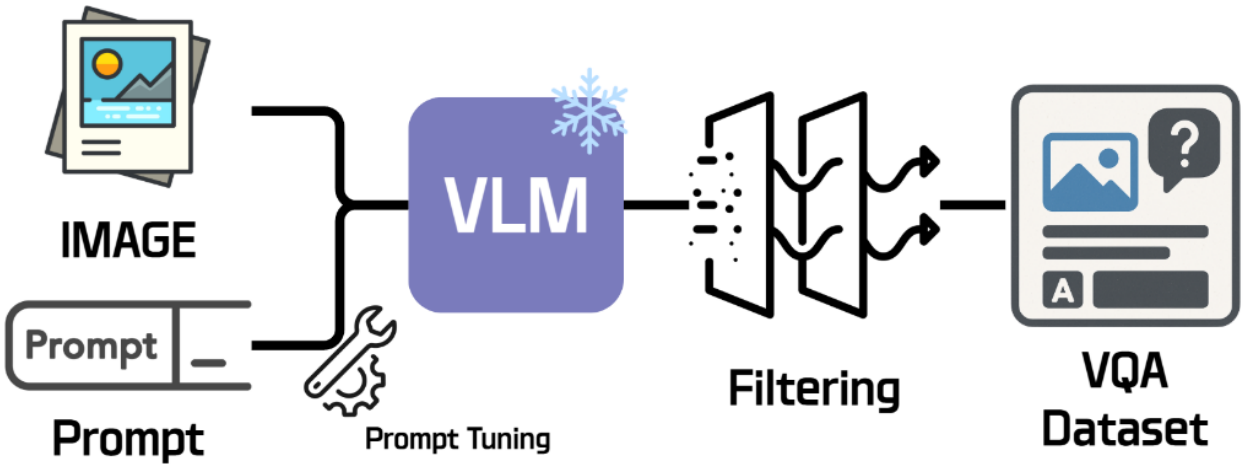
이 과정에서 이미지와 텍스트는 각각 비주얼 토큰과 텍스트 토큰으로 변환된 뒤 하나의 멀티모달 입력으로 VLM에 전송된다. 필터링 단계의 답변으로 이상 여부가 ‘Yes’로 확인되면 곧바로 설명 단계로 넘어가며, VLM은 “The children break in over the fence”와 같이 사람 이해에 용이한 문장을 생성한다. 생성된 결과는 JSON 형식 또는 데이터베이스 레코드로 저장되어, 관리자 알림 시스템이나 모바일 앱 대시보드와 연동되어 실시간으로 제공된다.

이와 같은 파이프라인은 전체 영상이 아닌 실제 이상 구간만 처리함으로써 연산 비용과 네트워크 전송량을 크게 절감할 뿐 아니라, 2단계 질의 설계를 통해 오탐을 효과적으로 걸러낸 경고 신뢰도를 높인다. 또한, 단순 객체 탐지를 넘어 사건의 시간적, 공간적 맥락을 파악하여 자연어로 설명함으로써 보안 담당자가 긴급 상황을 더욱 신속,정확하게 이해하고 대응할 수 있도록 지원한다.



d. AutoLabeling

본 프로젝트에서는 대형 VLM과 Decision Tree 형식의 프롬프트 전략을 적용하여 학습 데이터의 AutoLabeling 과정을 수행하였다. 각 이미지에 대해 "이 이미지는 위험 상황인가? 위험하다면 객체는 어떤 위험 상태인가?"와 같은 트리 구조의 질문들을 하나의 프롬프트 안에 계층적으로 배치하고 프롬프트로 통합하여 모델에 전달했고, 계별 분류 기준에 따라 답변을 유도하도록 구성하였다. 이 방식은 별도 후속 호출 없이도 초기 객체 검출부터 세부 상태 및 클래스 식별까지의 논리 흐름을 유지하여, 프롬프트 호출 횟수를 최소화하면서도 오류 누적을 효과적으로 방지할 수 있었다. 또한 객체 탐지, 상태 분류, 위험 등급 판단의 흐름을 일관되게 유지하여 높은 품질의 자동 레이블링 결과를 확보할 수 있었다.

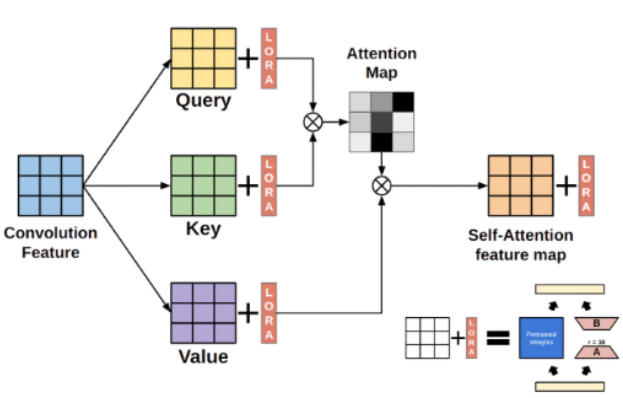


AutoLabeling Framework

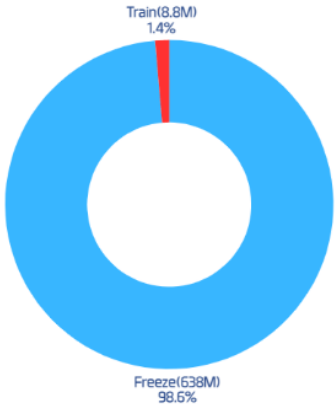
e. LoRA (Low-Rank Adaptation)

LoRA는 대형 사전학습 모델의 효율적인 도메인 적응을 위한 방법으로, Attention Layer 내부에 Low-Rank 형태의 Adapter를 삽입하여 미세 조정을 수행한다. 본 프로젝트에서는 InternVL3 모델에 LoRA 기반 Adapter를 적용하여, 전체 파라미터의 약 1.4%만을 학습 대상으로 설정하였다.

각 Attention Layer의 Querry, Key, Value, Output에 rank 16이 Adapter를 삽입함으로써, 적은 연산량으로도 도메인 특화 데이터에 대한 적응이 가능하도록 하였다.



(a) Adapter details location

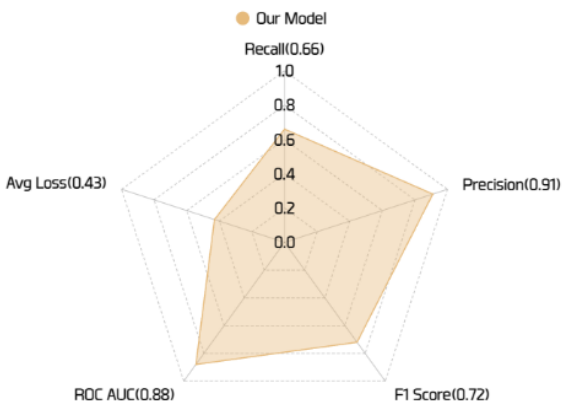


(b) LoRA trainable parameters

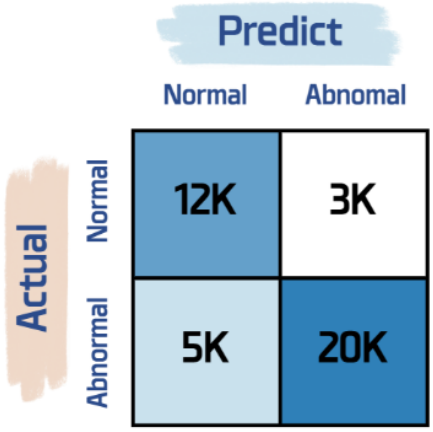
이를 통해 InternVL3는 특정 도메인에 최적화된 표현 학습 능력을 갖추게 되었으며, 실제 환경에 배포가능한 실용적인 경량 미세 조정 전략으로 작용했다.

f. 모델 성능 평가

본 시스템의 성능은 도메인 특화된 이상 상황 설명 및 위험 탐지 정확도 측면에서 평가되었다. 평가에는 수동 라벨링된 테스트 셋과 AutoLabeling 결과 비교를 통해 수집된 Ground Truth 데이터를 활용했다. 주요 평가 지표로는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score 등이 사용되었으며, 성능 결과는 아래 도표로 확인할 수 있다.



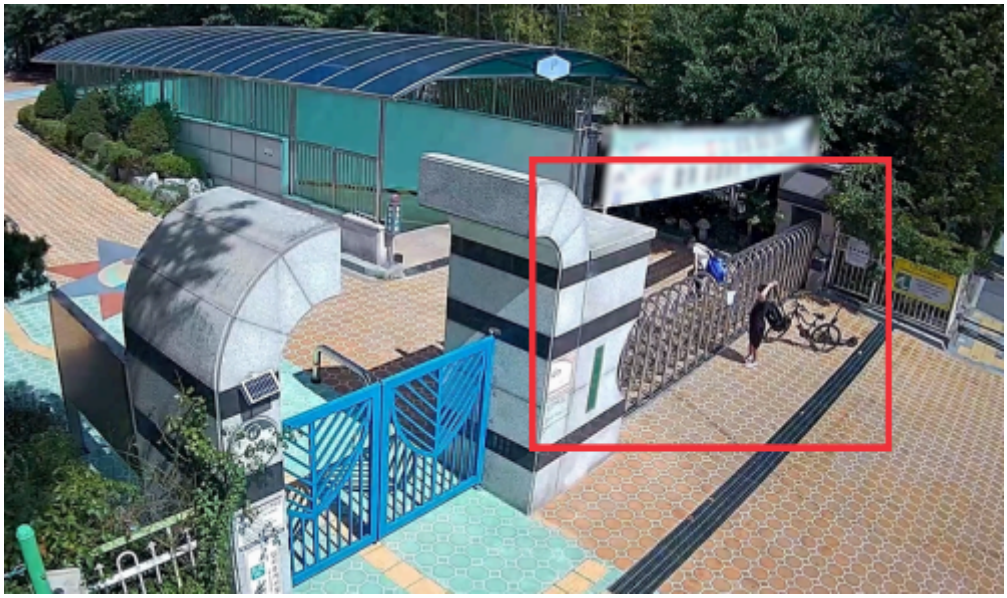
(a) Evaluation Indicators



(b) Confusion Matrix

3. 결과

본 프로젝트는 Jetson Orin Nano에서 터미널 기반으로 실시간 위험 상황 감지,알림,영상 전송 기능이 최적화된 상태로 구현되어 있으며, 현재 이를 전용 애플리케이션으로 확장하기 위해 서버 구축이 진행 중이다. 앞으로 앱과 서버 연동이 완료되면 사용자는 터미널 대신 직관적인 GUI를 통해 상황 설명과 영상을 실시간 확인하고 즉각 대응할 수 있게 될 것이며, 모듈화된 설계를 바탕으로 추가 기능 개발 및 실제 환경 적용도 용이하다.



Example of Autolabeling and Description

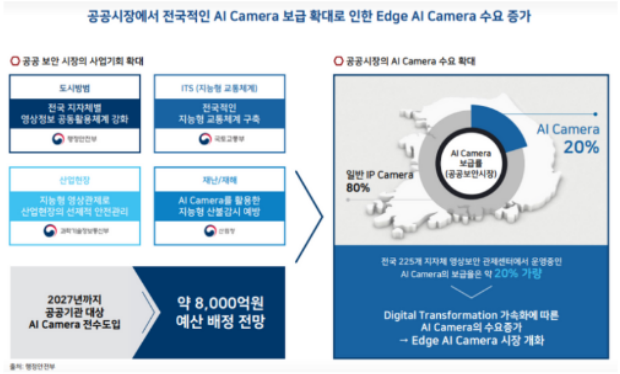
- Start Anomaly Frame : 1826
- End Anomaly Frame : 2854
- Event Caption : “Two people arrive at the same time in front of the school, and the person carrying the blue bag first breaks into the school, followed by the other person wearing the black top and bottom”
- Region of Interest : (1150,300,1600,700)

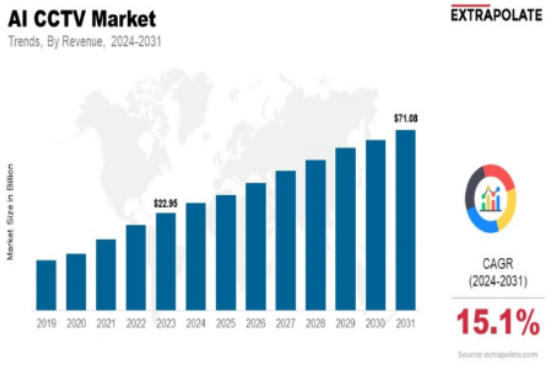
Example of application

4. 창업 플랜

a. 3C 분석

1) Customer





시장 특징 : 산업 현장, 공공기관, 학교 등에서 실시간 안전 모니터링 수요 증가

주요 니즈

- 단순 객체 감지가 아닌 행위 흐름 파악과 설명 기능
- 실시간성, 낮은 오탐율, 맥락 기반 고도화된 분석 시스템

주요 타겟

- 산업 현장/공장: 중대재해 예방
- 대형 유통시설/마트: 폭행·강도 대응
- 학교/공공기관: 폭력·침입 감지
- 도시 안전망: 무인 감시 및 위협 분석

2) Competitor

하이앤텍

CSB2-8M5C-1MQARK

- Object detection
- Thermal imaging
- Fixed-area access detection
- Cloud-based

kt Enterprise

KTT-5MIPSAI

- Object detection
- Fixed-area access detection
- Cloud-based

한화비전

QNO-C8023R

- Object detection
- Fixed-area access detection
- NPU-based edge board

업체	주요 기능	취약점 및 한계점

하이앤텍	다양한 안전 관련 감지 기능 (작업복 미착용 등)	룰 기반 감지, 프레임 단위 분석 한계, 맥락 파악 불가
BIGEYE	상태 변화 모니터링	객체 중심 판단, 행위 해석 불가능, 설명 기능 없음
Shield365	AI 관제	상세 기능 및 동작 방식 미흡, 시퀀스 기반이 아님
KT CCTV	다양한 도시형 안전 감지	고정 범주에 한정, 장면 설명 및 상황 해석 불가
에스원	고도화된 위험 감지 제공	텍스트 설명 부재, 시간 흐름 반영 불가

차별점 요약 : 시퀀스 단위 분석, VLM 기반 설명 생성, Edge 최적화, 설명 가능한 AI

3) Company

- 기술력
  : MobileNetV2 + GRU 기반 시퀀스 분석
  : VLM으로 자연어 설명 생성
  : Edge 디바이스에 최적화된 구조
- 혁신성
  : 기존 CCTV의 한계를 넘어선 맥락 기반 위험 인식 실현
  : 행동의 의도를 팡가할 수 있는 구조
- 시장 적합성
  : 중대재해처벌법 이후 설명 가능한 AI 보안 시스템 수요 급증
  : 현재 시장은 객체인식 중심에 머물러있기 때문에 고도화 필요

b. SWOT 분석

Strength **시퀀스 기반**으로 장면의 흐름 인식 가능 -VLM을 통해 **상황에 대한 자연어 설명 출력** **-엣지 디바이스 환경에 최적화** -기존 대비 **오탐율 낮고 맥락 인식이 정확**	**Weakness** 시퀀스 모델은 **학습량이 많고**, 학습 시간이 상대적으로 더 필요함 다양한 상황 데이터 확보가 필수적 (데이터 구축 비용 존재)
Opportunity 중대재해처벌법 등 **안전규제 강화로 수요 증가** "AI가 설명해주는 CCTV"에 대한 **수요와 사회적 관심 증가**	**Threat** **경쟁사 제품 다수 출시로 차별화 어려움** 영상 기반 AI 분석의 **개인정보 이슈 및 규제 가능성**

c. STP 분석

1) Segmentation

	주요 니즈	특징
산업 현장	중대재해 예방	보안보다 안전 위주 시스템
유통/마트	절도, 강도 감지, 비정상 행동 탐지	고객/직원 모두 감시대상으로 두고 보안, 안전 모두 고려, 즉시성 중요
학교/교육기관	침입, 학교폭력, 흡연 등 감지	
지자체/공공기관	군중, 화재, 범죄 탐지	24시간 무인 감시 필요

2) Targeting

: 중대재해 대응이 필수인 산업 현장, 강도폭행 등 대응이 필요한 대형 유통시설, 사람의 의도를 인식할 수 있는 CCTV 시스템의 니즈가 클 것으로 예상되고 VLM의 강점이 직접적으로 반영될 수 있을 것으로 예상됨

3) Positioning

: 맥락을 이해하는 CCTV로, 단순히 보는 것이 아닌, 이해하고 설명하는 보안시스템이다.
기존 시스템은 단순 객체 감지에 머무리고, 판단이 미흡할 수 있는데, 우리의 시스템은
예를 들어, 화염을 감지하고 추가적으로 위험 상황 전체를 설명한다.