HarassWatch: 소셜 VR 플랫폼에서의 피해자 관점 괴롭힘 행위 탐지

이준희[1], 김진우[2]
[1,2]광운대학교 (대학원생, 교수)

Virtual Reality 기술의 발전과 문제점

오늘날 VR기술은 소셜 플랫폼으로 발전하여 사용자들간 몰입감 있는 상호작용을 경험

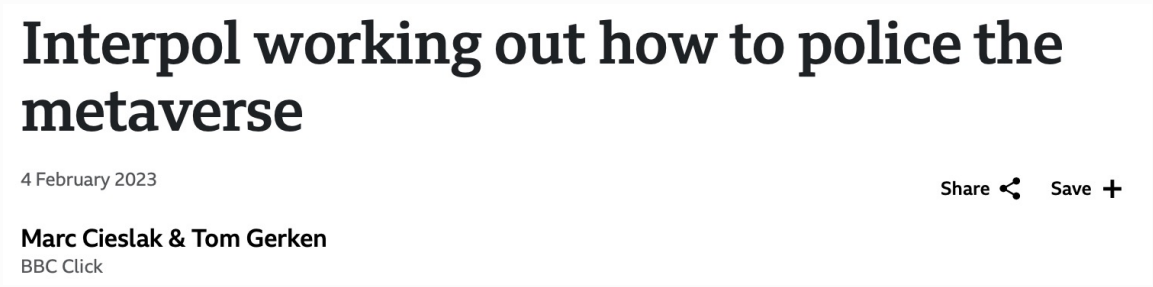
소셜 VR 플랫폼의 괴롭힘 문제

- 소셜 VR 플랫폼에서 괴롭힘 실태
  - 소셜 플랫폼의 온라인 괴롭힘 (공격적인 메시지, 공격적인 목소리)
  - VR 월드 내의 아바타 성추행 [1]
  - 온라인 내 범죄에 대한 수사 방법, 범죄에 대한 채증 방법에 대한 부재 [2]

Police investigate virtual sex assault on girl's avatar

3 January 2024

Chris Vallance
Technology reporter, BBC News



Interpol working out how to police the metaverse

4 February 2023

Share Save

Marc Cieslak & Tom Gerken
BBC Click

[1] "VRChat - Steam Charts", https://steamcharts.com/app/438100
[2] "Police investigate virtual sex assault on girl's avatar", https://www.bbc.com/ne ws/technology-67865327

기존의 연구

□ 실제 VR 소셜 플랫폼에서의 피해자 인터뷰 [3]

■ 대부분의 이용자가 10대인 플랫폼에서 성희롱을 포함한 괴롭힘 문제가 빈번히 보고

■ VR 소셜 플랫폼에서 이런 괴롭힘을 막을 방법을 고지하지 않음

Group	ID	Gender	Age	Occupation	Location	Num. Kids	Usage Experience	Used Social VR Platforms
Teenager	T1	Female	17	Student	USA	0	2 years	VRChat, Rec Room Horizon Worlds
	T2	Male	14	Student	USA	0	2 years	VRChat, Rec Room
	T3	Male	17	Student	USA	0	1 year	VRChat, Rec Room
	T4	Male	14	Student	USA	0	2 years	Rec Room
	T5	Male	15	Student	Lithuania	0	2 years	Rec Room, EchoVR
	T6	Male	13	Student	USA	0	1 year	Rec Room
	T7	Male	13	Student	USA	0	1.5 years	VRChat
	T8	Female	17	Student	Belgium	0	2 years	VRChat, Rec Room

Table 1: Participants' demographics and social VR experience [3]

[3] Delda ri, Elmira, et al. "An investigation of teenager experiences in social virtual reality from teenage rs', parents', and by standers' perspectives." Symposium on Usable Privacy and Security (SOUPS). 2023.

기존의 연구

□ VR 소셜 플랫폼의 개발자 인터뷰 [4]

- 괴롭힘 문제를 인지하고 있음
- 그러나 기술이 부족하고, 인력이 부족함

"Companies are very money-first, fix later. [Maybe] this is why it's getting pushed off, and not many people are talking about it or fixing it." (D9)

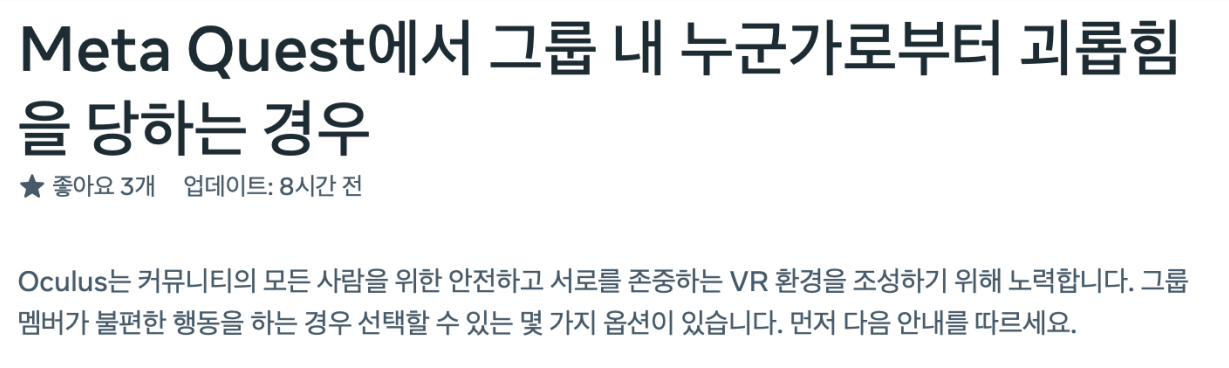
개발자 인터뷰 중... [4]

[4] Abhinaya, S. B., Aafaq Sabir, and AnupamDas. "Enabling Developers, Protecting Users: Inverstigating Harassment and Safety in VR. " USENIX Security Symposium, 2024.

기존의 연구

□ Meta Horizon World 내의 해결방안 [5]

- 월드내의 행동을 모니터링하고 녹화하는 안전 전문가 배치
- 다른 플레이어가 접근하지 못하게 하는 개인 경계 기능 도입



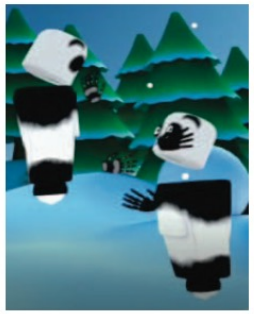
[5] "Meta Horizon World의 안전 및 개인정보 보호", https://www.meta.com/ko-kr/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/personal-boundary-horizon-worlds/

기존의 연구

☐ HardenVR의 괴롭힘 감지 연구 [6]

- 유저의 VR 기기, 컨트롤러의 좌표 정보와 컨트롤러 버튼 입력 정보를 통해 괴롭힘 감지
- 괴롭힘을 정확도 98% 이상으로 감지하고 있음

(a) OK gesture



(b) Slapping

Figure 3: Screenshots of two studied social behaviors in a virtual room in social VR [6]

[6] Wang, Na, et al. "HardenVR: Harassment detection in social virtual reality." 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, 2024.

연구 목표

□ 기존 연구에서는 좌표 및 입력 정보를 통해 괴롭힘 탐지하나 시각적인 증거가 부족

➔ 괴롭힘에 대한 명확한 증거 제작

□ 기존 연구는 모든 유저에 대한 좌표와 입력 값을 수집해야 함

➔ 유저가 늘어나도 탐지에 문제가 없어야 함

□ 기존 연구는 VR기기에서의 모델 추론을 고려하지 않음

➔ VR기기의 제한된 자원으로 탐지 및 증거 제작이 가능해야 함

연구 목표

□ 기존 연구에서는 좌표 및 입력 정보를 통해 괴롭힘 탐지하나 시각적인 증거가 부족

➔ 괴롭힘에 대한 명확한 증거 제작

□

VR기기에서 작동하는 1인칭 시점 괴롭힘 탐지 및 증거 제작 시스템 제작

□ 기존 연구는 VR기기에서의 모델 추론을 고려하지 않음

➔ VR기기의 제한된 자원으로 탐지 및 증거 제작이 가능해야 함

실험 환경

□ 머신 : CPU – Intel i7 14세대
메모리 – 32GB
GPU – GeForce RTX 4080

□ VR 기기 :

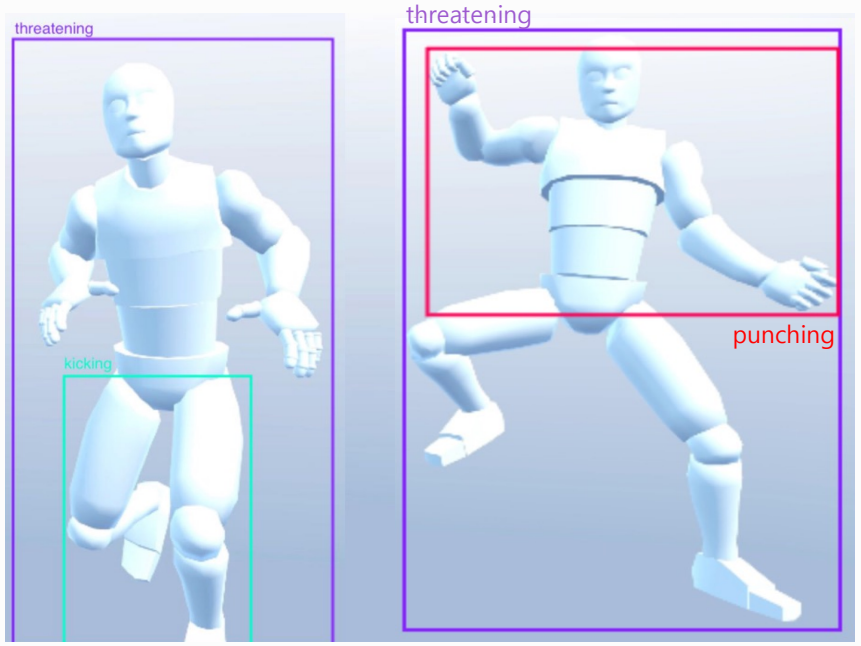
메타 퀘스트 2 : 2 대

메타 퀘스트 프로 : 1 대

□ 개발 환경 : Unity Engine 6

데이터셋

- Punching, Kicking, Threatening 총 3개의 클래스를 수집
- VR을 장착한 유저 앞에서 위협 행위를 녹화하여 수집
- 녹화된 영상을 Roboflow를 통해 라벨링
- 약 500개의 데이터셋 확보

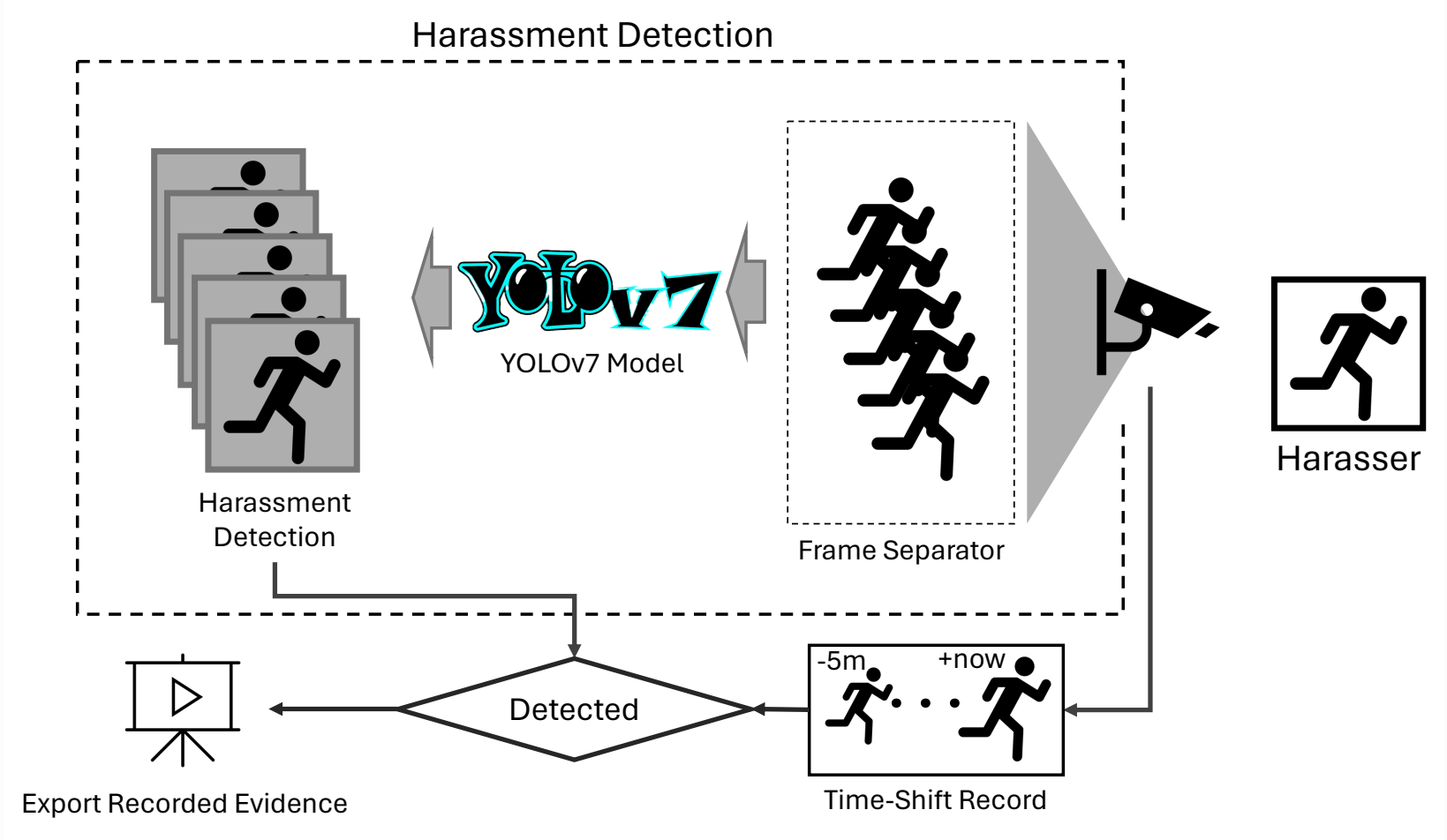


학습 모델

- YOLOv7-tiny 모델을 사용
  - Unity 엔진에서 제공하는 AI 추론 엔진인 Sentis는 ONNX 형식을 지원해야 함
  - VR 기기에서 사용하기 위해 용량이 작고 사용하는 자원이 적어야 함

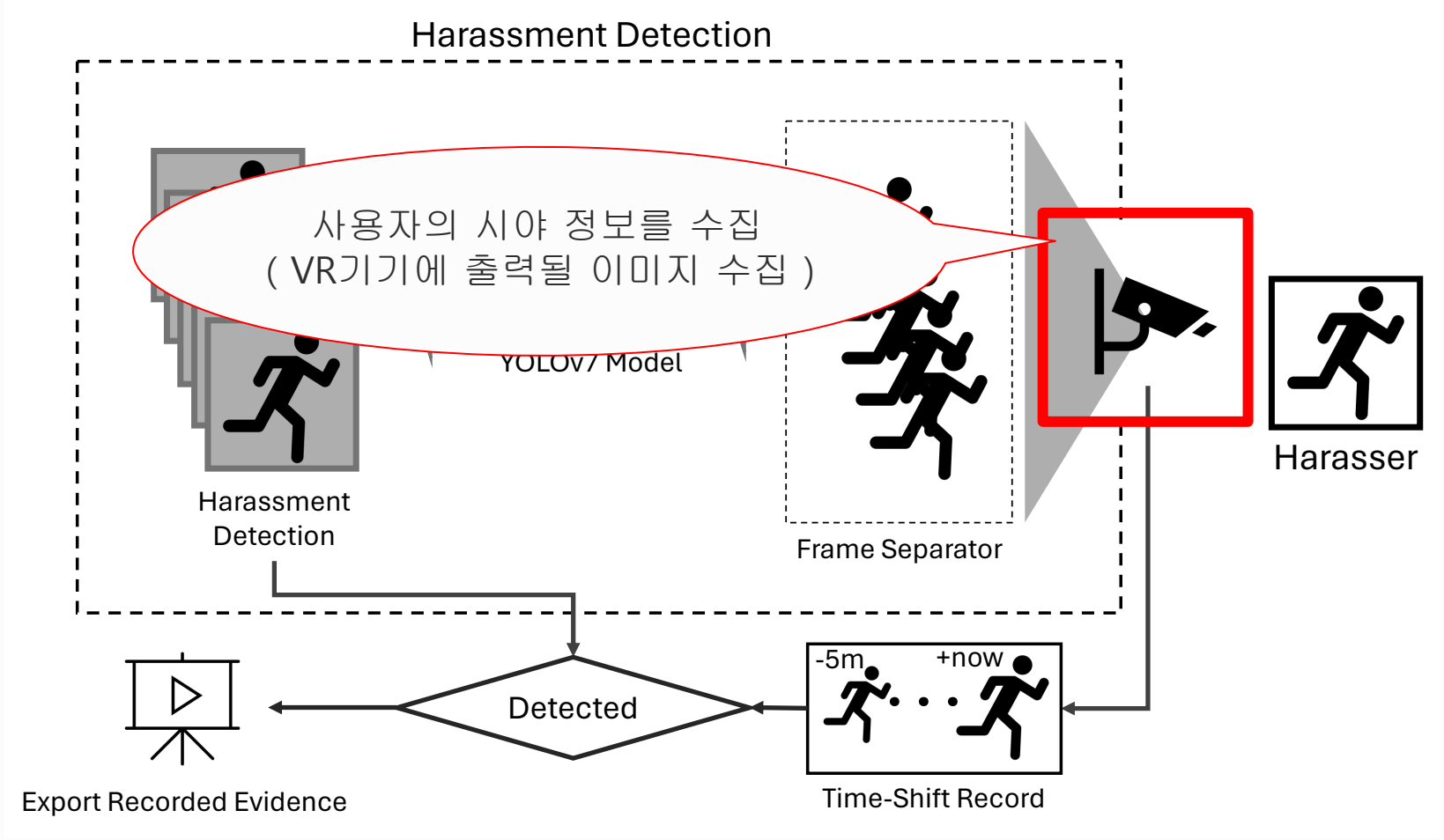
HarassWatch

☐ HarassWatch 시스템도



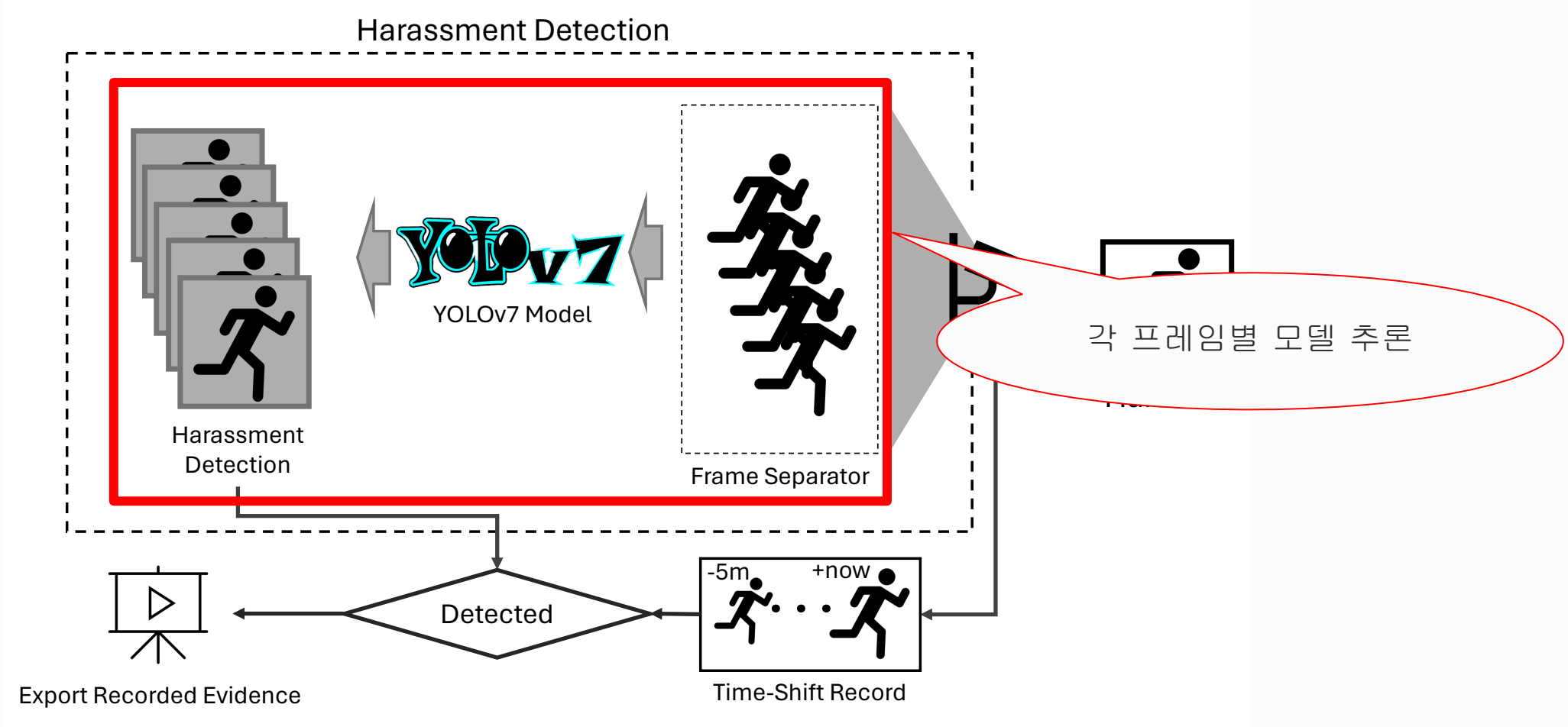
HarassWatch

☐ HarassWatch 시스템도



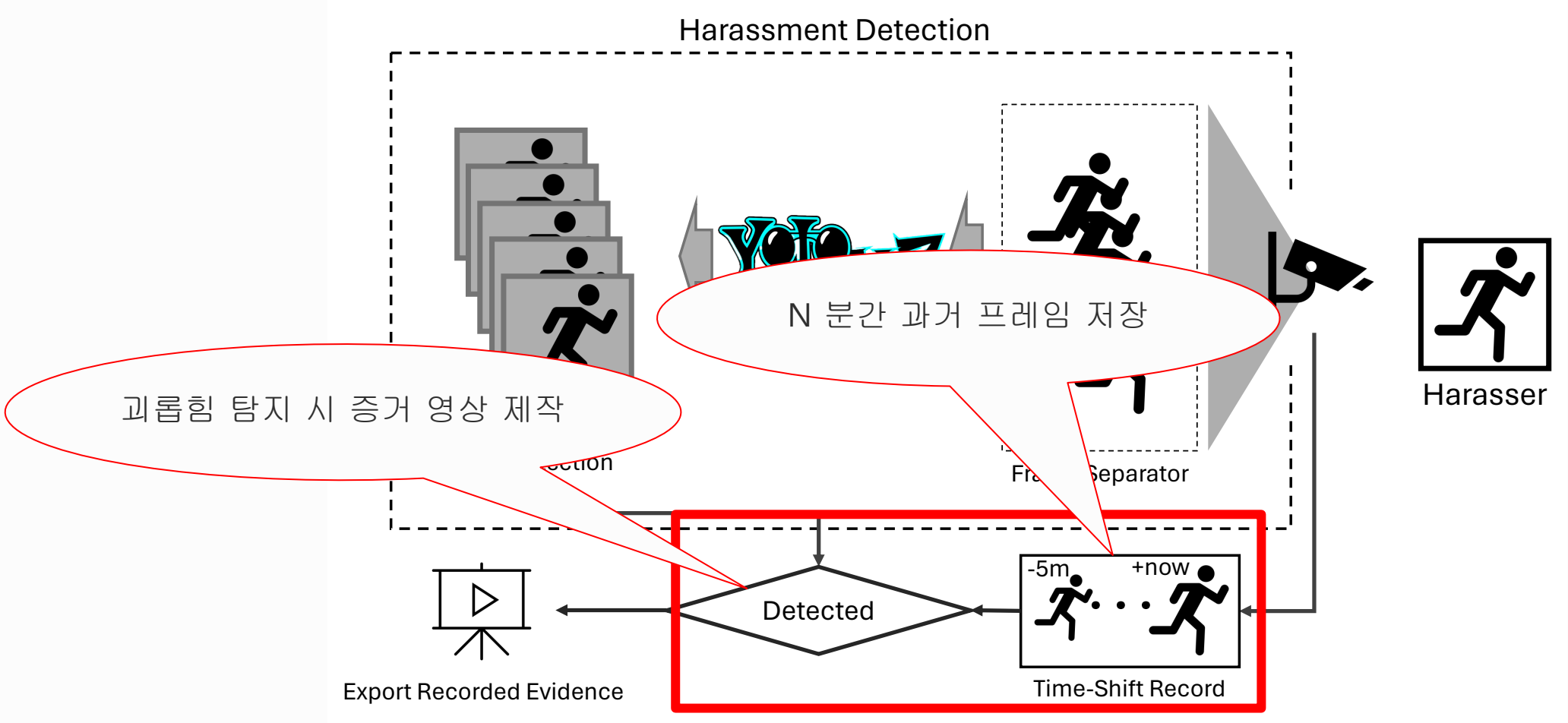
HarassWatch

☐ HarassWatch 시스템도



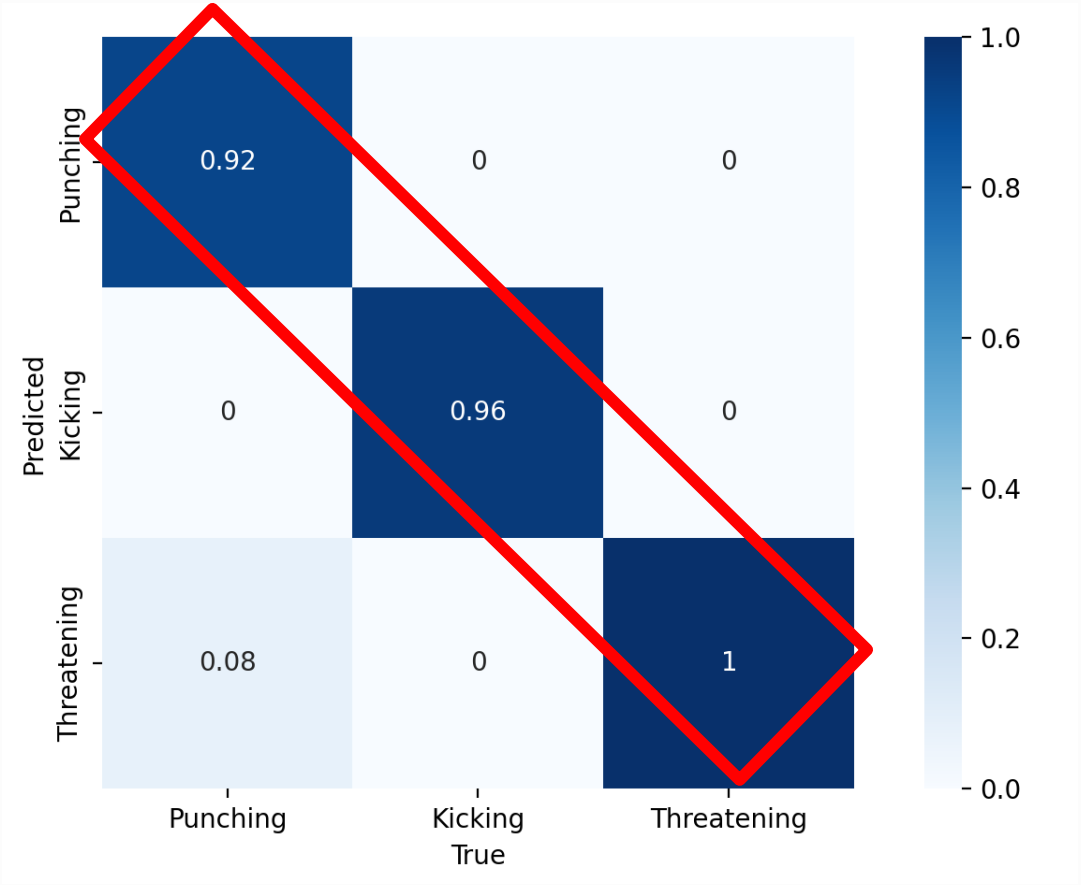
HarassWatch

☐ HarassWatch 시스템도



실험 결과

☐ 모델 학습 결과 – Confusion Matrix



모델이 예측한 클래스

실제 클래스

☐ 각 클래스가 준수하게 학습됨

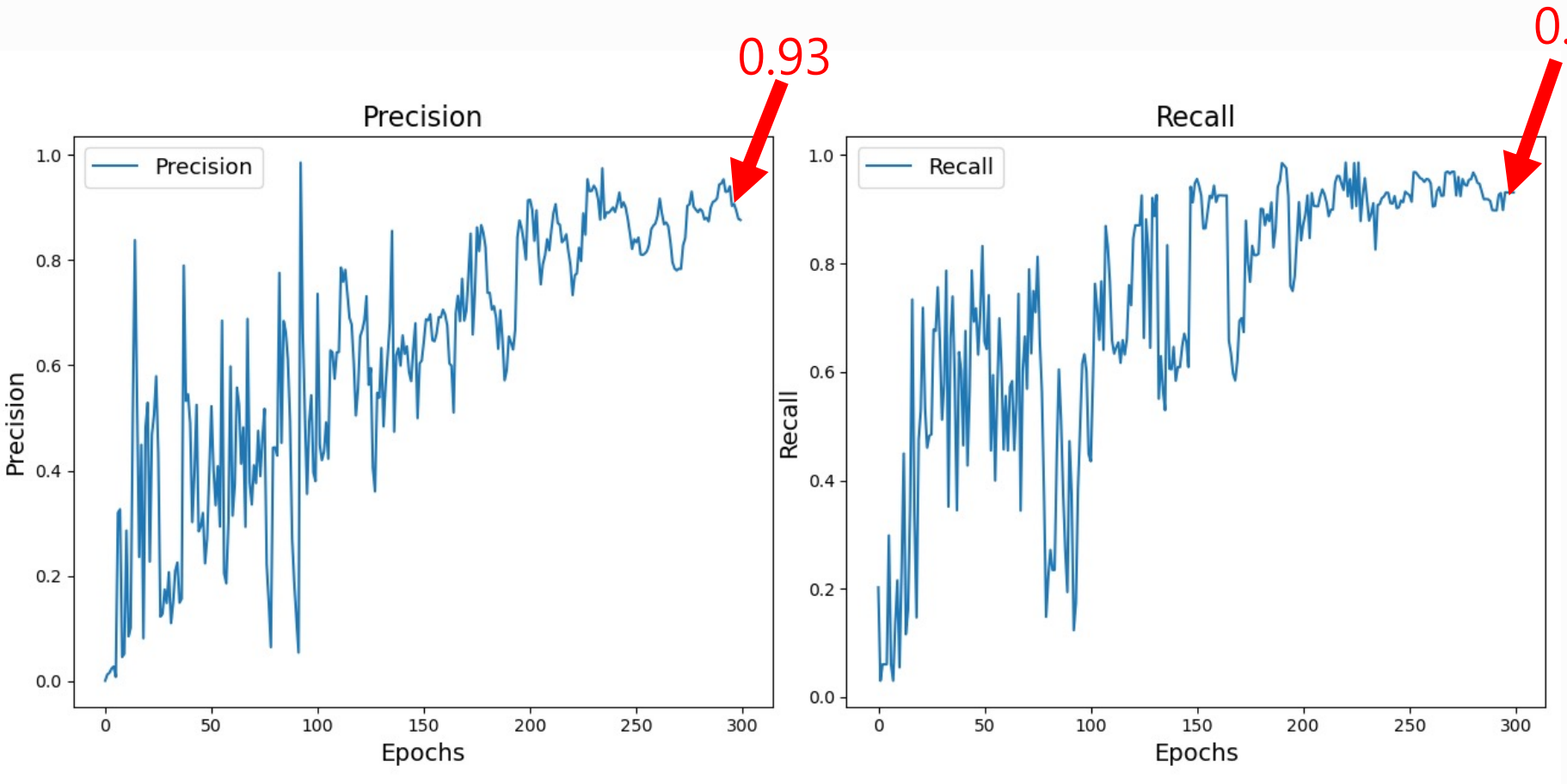
☐ Punching – 0.92

☐ Kicking – 0.96

☐ Threatening – 1

실험 결과

□ 모델 학습 결과 – Precision & Recall

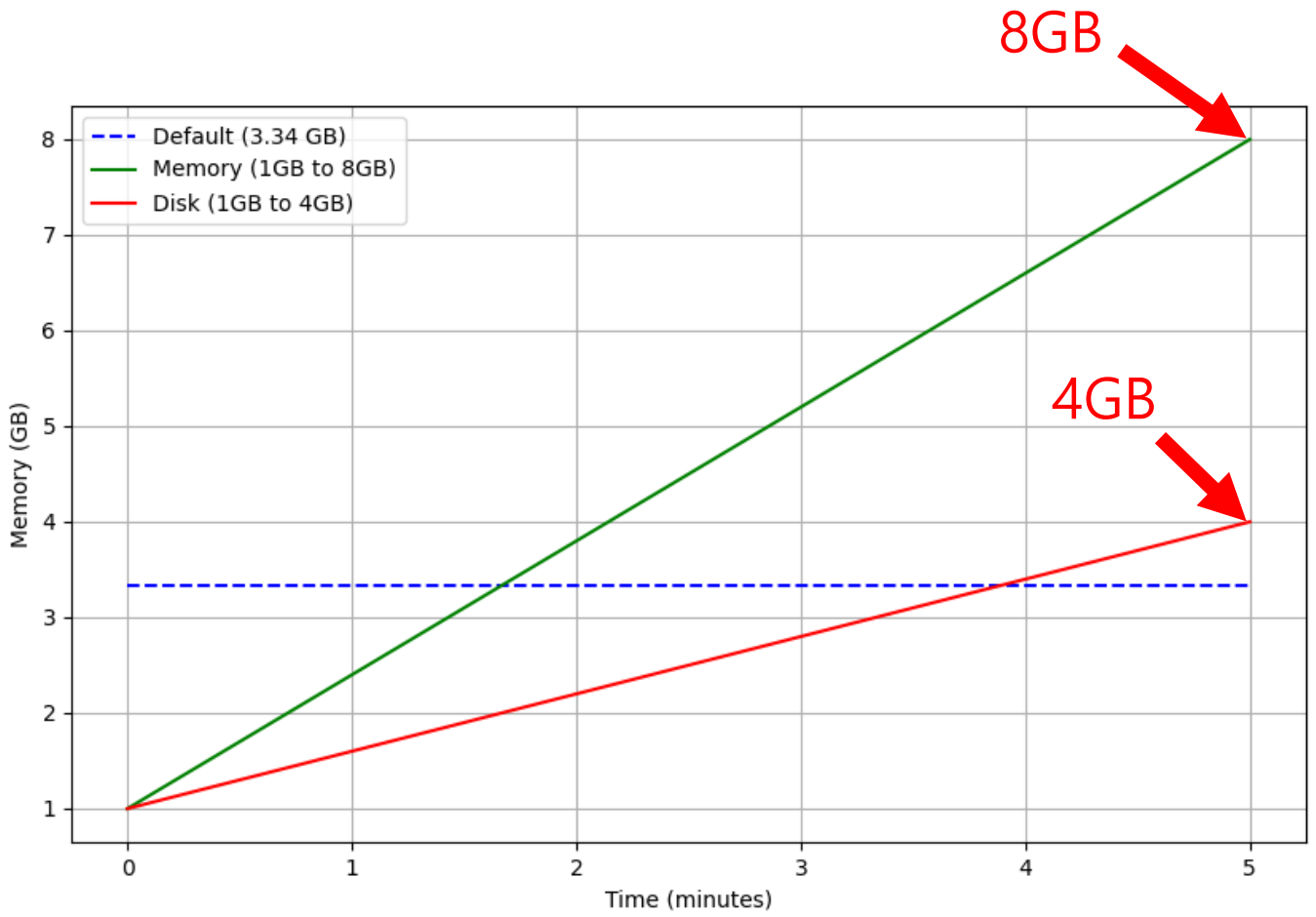


□ 정밀도 : 0.93

□ 재현율 : 0.92

실험 결과

□ 영상 녹화 기능 메모리 사용량



□ 메모리 사용 시
8 GB – 3.34 GB
= 4.66 GB

□ 디스크 공간 사용 시
4 GB – 3.34 GB
= 0.66 GB

실험 결과

- 정밀도, 재현율 각각 93%, 92%
- Punching, Kicking, Threatening 클래스의 정확도 각각 92%, 96%, 100%
- 영상 녹화 기능은 녹화 시간에 따라 메모리, 디스크 점유율이 달라짐
  - 5분 기준으로 약 0.7GB 사용

결론

- 현재 제안한 HarassWatch는 괴롭힘을 감지하고 증거를 제작하는 시스템
- 향후 연구에서는 부족한 데이터셋을 보안
  - 괴롭힘 행동 클래스 추가
  - 풀 트래킹 장비를 이용해 현실적인 괴롭힘 행동 수행
  - 실험을 통해 실제 유저들이 괴롭힘을 당했을 때의 상황을 인터뷰
- YOLOv7-tiny가 아닌 수집한 데이터에 최적화된 모델을 제시

KWANGWOON UNIVERSITY

감사합니다