Social VR Administrator: 소셜 VR내 괴롭힘 방지를 위한 상황 인지형 모델 에이전트 디자인*

이준희[1] 김민석[2] 허환조[3] 우승원[3] 김진우[4†]

[1,2,4]광운대학교 (대학원생, 학부생, 교수) [3]ETRI (연구원)

Social VR Administrator: A Context-Aware Model Agent Design for Harassment Prevention in Social VR

Junhee Lee[1], Minseok Kim[2], Hwanjo Heo[3], Seungwon Woo[3], Jinwoo Kim[4†]

[1,2,4]Kwangwoon University(Graduate student, Student, Professor)
[3]ETRI(Researcher)

요 약

소셜 가상현실(VR) 플랫폼은 사용자에게 몰입감 높은 경험을 제공하지만, 동시에 심각한 온라인 괴롭힘 문제에 노출되어 있다. 기존의 괴롭힘 탐지 시스템은 단일 모델에 의존하여 모든 상황에 대응하므로 유연성과 효율성에 한계가 있었다. 본 논문에서는 Vision-Language Model(VLM) 기반 탐지 시스템을 확장한 에이전트 프레임워크인 SVRA를 제안한다. SVRA는 소셜 VR 환경의 동적인 상황과 맥락을 분석하여 VLM, LSTM, Transformer 등 다양한 탐지 모델 풀에서 최적의 모델을 동적으로 선택하고 조합한다. 이는 각 모델이 가진 고유의 장점(예: LSTM의 효율성, Transformer의 정확성, VLM의 맥락 이해 능력)을 극대화하는 접근법이다. 또한 탐지 결과의 심각도와 확실성에 따라 경고, 보고, 차단의 3단계로 차등 대응하는 행동 모듈을 통합하여, 보다 정교하고 실용적인 괴롭힘 방지 조치를 가능하게 한다. 본 논문은 SVRA의 설계 철학과 아키텍처를 제시하며, 동적 모델 선택과 다단계 대응 시스템이 지능적인 소셜 VR 안전 관리 패러다임을 구축할 수 있음을 논의한다.

I. 서론

소셜 가상현실(VR) 플랫폼은 전례 없는 수준의 사회적 상호작용과 몰입감을 제공하며 빠르게 성장하고 있다. 대표적인 소셜 VR 플랫폼인 VRChat은 동시 접속자 수 6만 명을 기록하는 등 대중적인 인기를 얻고 있다. 하지만 이러한 성장의 이면에는 스토킹, 신체적 침해, 성희롱 등 심각한 온라인 괴롭힘이라는 어두운 단면이 존재한다. VR 환경의 높은 몰입감은 장소 환상과 사실성을 유발하여 가상 세계에서의 경험을 실제처럼 느끼게 만든다. 이로 인해 소셜 VR에서의 괴롭힘은 사용자에게 증폭된 정신적 피해를 주며, 이는 단순한 가상 세계의 문제를 넘어 현실 세계의 법적 조치로 이어질 만큼 심각한 사회 문제로 대두되고 있다.

현재 대부분의 플랫폼은 개인 경계(버블) 설정이나 사용자 차단/신고 기능과 같은 사후 대응책에 의존하고 있다[1]. 그러나 이러한 방식은 괴롭힘이 이미 발생한 후에야 작동하며, 갑작스러운 공격에 피해자가 즉각적으로 대응하기 어렵다는 근본적인 한계를 가진다. 사용자들은 종종 신고 절차의 불투명성과 불충분한 조치에 불만을 표하며, 괴롭힘의 증거를 수집하고 제출하는 과정 자체가 또 다른 부담으로 작용하기도 한다. 이를 극복하기 위해 사용자의 컨트롤러 입력이나 아바타 위치 같은 생체 정보를 활용한 딥러닝 기반의 사전 탐지 연구가 시도되었으나, 이는 민감한 개인정보 수집에 대한 프라이버시 침해 우려를 낳는다[4].

이러한 점을 고려하여 우리는 사용자의 시각

* 본 연구 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(No. RS-2023-00215700, 트러스트 메타버스 실현을 위한 블록체인 융합기술)과 한국연구재단(No. RS-2024-00457937, 안전한 웹어셈블리 기반 서버리스 환경을 위한 보안 계층 설계 및 구현)의 지원을 받아 수행된 연구임.

† 교신저자(jinwookim@kw.ac.kr)

적 데이터만을 활용하여 개인정보 침해 없이 괴롭힘을 탐지하는 Vision-Language Model (VLM) 기반의 에이전트 시스템인 SVRA를 제안한다. VLM은 예외적인 상황(예: 격투기 체험 가상 세계에서의 폭력 행동)의 문맥을 이해하고 괴롭힘 행동을 분류하는데 유리하지만, 하나의 강력한 모델에 의존하는 방식은 몇가지 한계를 가진다. 첫째로 모든 상호작용을 고비용의 VLM으로 분석하는 것은 계산적으로 비효율적이다. 둘째로 다중 레이블 분류 문제에서는 VLM보다 더 가볍고 빠른 모델(예: LSTM, Transformer)이 더 적합할 수 있다.

본 논문에서는 VLM의 강력한 문맥 파악 능력과 가볍고 빠른 모델인 LSTM, Transformer을 상황에 따라 적용하고, 실제 관리자의 역할을 대신하는 경고, 보고, 차단의 3단계로 차등 대응하는 지능형 에이전트 프레임워크 SVRA를 제안한다. 이는 비용과 성능 사이의 균형을 맞추는 시스템을 지향하며, 탐지 결과에 따라 경고, 보고, 차단이라는 세 단계로 차등화된 조치를 실행하여 유연하고 실용적인 대응을 가능하게 한다.

II. 관련 연구 및 배경 지식

2.1 소셜 VR에서의 괴롭힘 탐지

소셜 VR 내 괴롭힘 문제에 대응하기 위해 여러 해결책들이 논의되어 왔다. 사용자 연구에 따르면 사적 공간에서의 성적 괴롭힘 등 다양한 괴롭힘이 빈번하며 이를 예방하기 위한 관리자의 필요성이 강조되었으나[3], 대부분의 플랫폼은 자원과 인력 부족으로 괴롭힘 대응을 우선수위로 다루지 못하고 있다.

기술적 접근 방법으로는 괴롭힘 탐지 연구인 HardenVR이 있다. HardenVR은 사용자의 컨트롤러 입력, VR 기기 위치, 컨트롤러의 위치 정보 등 사용자의 생체 정보를 활용한 딥러닝 모델로 주먹질, 뺨 때리기 등의 행동을 98% 이상의 높은 정확도로 탐지했다[2]. 그러나 민감한 개인정보 수집으로 인해 심각한 프라이버시 우려를 낳는다.

2.2 괴롭힘 탐지 모델별 성능 분석

우리는 사용자 연구를 통해 구축한 우리의 시각 기반 데이터셋을 기반으로 LSTM, Transformer, VLM (GPT-4o, Chain of Thought) 모델을 사전에 학습 및 파인튜닝(fine-tuning) 시켜 성능을 평가하였다. 데이터셋은 8명에게서 수집한 825개의 비디오 클립을 10초 단위로 전처리하여 3,408개의 비디오 클립을 확보했다. 데이터는 정상 행동, 비정상 행동을 분류하는 Stage 1과 비정상 행동 중 공격적인 행동, 개인 공간 침해, 방해 행동, 정상 행동으로 분류하는 Stage 2로 각각 학습시켰다. 표1은 각각의 모델을 Stage 1, Stage 2 데이터로 학습 및 파인튜닝한 결과이다.

모델	Stage	정확도	평균 F1-Score
LSTM/CNN	1	0.77	0.63
	2	0.43	0.23
Transformer	1	0.88	0.68
	2	0.70	0.67
GPT-4o, CoT	1	0.88	0.84
	2	0.64	0.62

표 1 모델별 학습 및 파인튜닝 결과

LSTM/CNN 기반 모델: 이 모델은 비교적 가볍고 추론 속도가 빠르다는 장점이 있다. 그러나 전반적인 정확도와 F1-Score는 다른 모델에 비해 낮다. 이는 복잡한 맥락 이해 능력이 부족하기 때문이다. 따라서 1차적인 빠른 판단 역할에 가장 적합하다.

Transformer 기반 모델: 비디오 분류에서 뛰어난 성능을 보이는 모델로 우리의 데이터셋에서도 가장 높은 수준의 정확도를 기록했다. 다양한 유형의 괴롭힘을 탐지할 수 있다.

VLM (GPT-4o, CoT): VLM의 가장 큰 장점은 맥락 이해 능력이다. Transformer와 유사한 수준의 높은 정확도를 보이면서도 소셜 VR내의 친구 관계, 가상 세계 정보와 같은 텍스트 정보를 함께 처리하여 행동의 의도를 파악할 수 있다. 특히 Chain-of-Thought(CoT) 프롬프팅을 통해 단계적 추론을 유도하여 괴롭힘과 정상적인 플레이를 구분해야 하는 모호한 상황에서 결정적인 역할을 한다. 그러나 API 호출 비용이 가장 높고 응답 지연 시간이 길다는 단점이 있다.

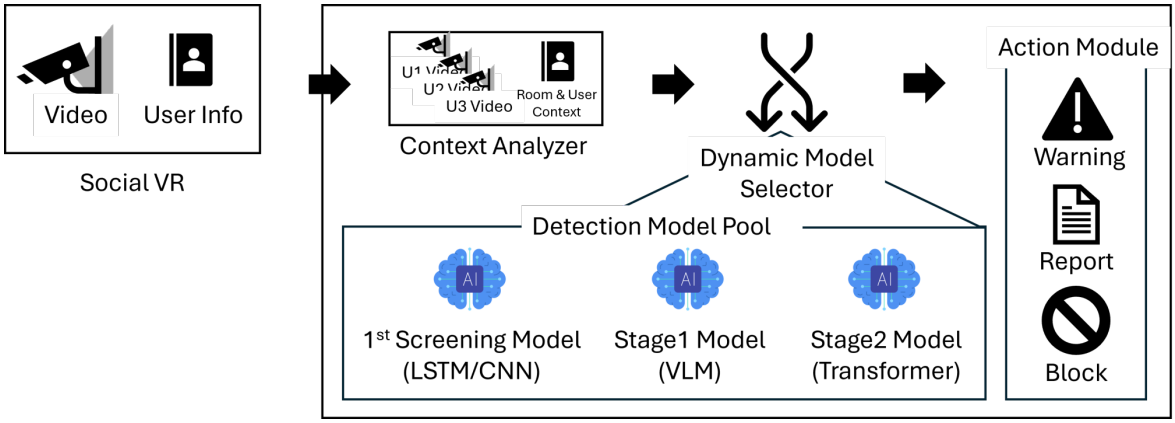

그림 1 SVRA 디자인 설계도

이러한 분석은 어떤 단일 모델도 효율성, 정확성, 맥락 이해 능력 세 가지 요소를 모두 만족시키지 못한다는 것을 보여준다. 따라서 이러한 장단점이 다른 모델들을 도구로 간주하고, 주어진 상황을 지능적으로 분석하여 최적의 도구를 선택하고 활용하는 에이전트의 필요성이 대두된다.

III. SVRA 프레임워크 설계

3.1 SVRA 구조

그림 1은 SVRA의 디자인 설계에 대한 그림이다. SVRA의 목표는 앞서 분석한 모델들의 장점을 결합하여 효율성, 정확성, 유연성을 갖춘 에이전트 시스템을 구축하는 것이다. 이를 위해 네 가지 핵심 모듈로 구성된 에이전트 기반 디자인을 설계했다.

컨텍스트 분석기(Context Analyzer): 소셜 VR 공간의 종류(대화방, 마피아 게임 방, 격투기 방 등), 사용자 간의 관계(친구, 낯선 사람), 과거 경고 여부, 세션 시간 등 현재 상황에 대한 메타데이터와 소셜 VR 공간내의 유저들의 시각 데이터를 수집하여 분석한다.

동적 모델 선택기(Dynamic Model Selector): 컨텍스트 분석기의 정보를 바탕으로, 탐지 모델 풀에서 현재 상황에 가장 적합한 모델 또는 모델의 조합을 선택하여 실시간으로 탐지 파이프라인을 구성한다. 여기서 선택기는 규칙 기반 정책과 학습된 정책 선택 방법 모두 적용 가능하다.

탐지 모델 풀(Detection Model Pool): 각기 다른 강점을 가진 여러 괴롭힘 모델의 집합으로 모델에 대한 간단한 설명이 포함된다.

행동 모듈(Action Module): 탐지 파이프라인의 결과를 종합하고, 사전에 정의된 경고, 보고, 차단 중 하나의 행동을 최종적으로 결정하고 실행한다.

소셜 VR내의 유저 메타데이터와 시각 정보는 컨텍스트 분석기를 거쳐 컨텍스트화 되어 모델 선택기로 전달된다. 모델 선택기는 전달받은 컨텍스트 정보를 분석하여 최적의 모델을 선택하거나, 여러 모델을 선택하는 다단계 파이프라인을 구성하고, 결과를 종합하여 행동 모듈을 통해 최종적으로 조치를 실행한다.

3.2 동적 모델 선택 및 다단계 탐지

SVRA은 동적 모델 선택 과정에서 총 1-3단계까지 거치는 다단계 파이프라인을 구성할 수 있다.

1단계는 스크리닝(Screening) 과정으로 계산 비용이 낮은 LSTM/CNN 모델을 이용하여 명백하게 공격적인 움직임 패턴을 빠르게 감지하여 1차적으로 걸러내는 역할이다. 대부분의 정상적인 상호작용은 이 단계에서 필터링되어 시스템의 전체적인 부하를 줄일 수 있다.

2단계는 잠재적인 괴롭힘으로 분류되거나, 모호한 상황이라고 판단된 경우 사용할 수 있다. 정상, 비정상을 나누는 Stage 1, 이보다 더 자세한 비정상 행동 패턴을 나누는 Stage 2를 검사하는 방식으로 각 Stage를 Transformer, VLM 등 원하는 모델을 선택할 수 있다. 다만 우리의 모델 성능 결과에 따라 Stage 1은

VLM, Stage 2는 Transformer를 사용하는 것으로 가정하였다.

3단계는 1, 2단계를 종합적으로 사용하여 판단하는 단계이다. 각 단계에서 나온 결과와 신뢰도 점수를 종합하여 판단하는 것이다. 예를 들어 1단계 LSTM과 2단계 VLM 모두 높은 신뢰도로 공격적 행동으로 판단했다면 최종 결정은 공격적 행동으로 확실해지나, 모델 간 판단이 엇갈릴 경우 이를 모호한 상황으로 분류하여 차후 행동 모듈 선택에 영향을 줄 수 있다.

3.3 행동 모듈

SVRA의 최종 판단은 행동 모듈로 전달되어 구체적인 조치로 이어진다. 이는 단순한 판단에 그치지않고 실제 대응하는 기능을 추가하여 현실적인 운영을 가능하게 한다.

경고(Warning)의 경우 괴롭힘의 강도가 약하거나(예: 일시적인 개인 공간 침범), 시스템의 판단 신뢰도가 낮을 경우 발동된다. 가해자로 추정되는 사용자에게 “당신의 행동이 다른 사용자에게 불편을 줄 수 있습니다.”와 같은 경고 메시지를 노출하여, 사용자 스스로의 행동을 교정할 기회를 제공하고 시스템이 이를 감지하였음을 알려 괴롭힘 행동을 위축시키는 역할을 한다. 반복적으로 경고가 누적된 경우 보고, 차단 단계로 상향될 수 있다.

보고(Report)의 경우 시스템이 판단을 내리기 어려운 모호한 상황에서 발동된다. 해당 상호작용의 영상, 각 모델의 분석 결과와 신뢰도 점수, 컨텍스트 데이터 등이 포함된 상세 보고서가 생성되어 관리자에게 전송된다. 이는 AI의 판단을 보완하고 오탐으로 인한 부당한 제재를 방지하기 위한 안전장치이다.

차단(Block)의 경우 명백하고 심각한 괴롭힘 행위가 높은 신뢰도로 탐지되었을 때 발동된다. 가해 사용자에 대해 아바타 전환, 강제 퇴장과 같은 즉각적인 조치가 이루어진다.

IV. 논의 및 한계점

본 연구는 소셜 VR 괴롭힘 문제에 대한 새로운 접근법을 제시하지만 이를 위한 평가 지표 및 평가 방법은 제시되지 못했다. 동시에 몇 가지 논의점과 기술적 과제를 안고 있다.

기술적 과제: 현재 SVRA의 성능은 각 개별 모델의 성능과 에이전트의 모델 선택 정책에 의존한다. 실시간으로 수많은 상호작용을 처리하기 위해 높은 수준의 최적화가 요구되며, 이를 모두 처리하기 위한 견고한 모델을 구축하는 것은 어려운 과제이다.

윤리적 고려사항: 자동화된 괴롭힘 탐지 방법은 특정 제스처를 공격적인 행동으로 오인하거나, 문화별 다른 행동 패턴에 대해 학습용 데이터가 부족하여 자칫 편향된 판단을 내릴 수 있다. 이러한 위험을 최소화하기 위한 검사 방법과 작동 방식을 투명하게 공개하여 사용자가 납득할 수 있는 절차를 구성하는 것이 중요하다.

V. 결론 및 향후 연구

본 논문에서는 기존의 소셜 VR 괴롭힘 탐지 시스템의 한계를 극복하고, 소셜 VR 플랫폼의 현실적인 문제를 해결하기 위한 지능형 에이전트 SVRA를 제안했다. SVRA는 각기 다른 장단점을 가진 탐지 모델들을 상황에 맞게 선택하고, 행동 모듈이 처리하여 실제 운영 환경을 크게 향상시킬 수 있다. 향후 연구로는 제안된 SVRA 디자인을 프로토타입으로 구현하고, 이를 평가하기 위한 평가 도구를 마련하여 공정한 평가를 진행할 것이다. 또한 대규모 사용자 연구를 통해 더욱 공정하고 효과적인 괴롭힘 탐지 시스템으로 구축할 예정이다.

[참고문헌]

[1] “Learn about community guides in Worlds”, https://www.meta.com/help/quest/2259751227488308/?srsltid=AfmBOorJtoG-th6rkKHCKde41hhRHGv_Am7X9kA7nGftQn66Lde3Am9G

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

[3] Abhinaya, S. B., Aafaq Sabir, and Anupam Das. "Enabling Developers, Protecting Users: Investigating Harassment and Safety in {VR}." 33rd USENIX Security Symposium (USENIX Security 24). 2024.

[4] Meng, Yan, et al. "De-anonymization attacks on metaverse." IEEE INFOCOM 2023-IEEE Conference on Computer Communications. IEEE, 2023.