Journal of KIIT. Vol. 24, No. 8, pp. 171-182, Aug. 31, 2026. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2026.24.8.171

# 시계열 연속성 검증과 VLM 자연어 리포트를 결합한 관광지 군중 병목 탐지 모델

서태영*, 유현**, 정경용***

# Tourist-Site Crowd Bottleneck Detection using Temporal Verification and VLM Reporting

Taeyoung Seo*, Hyun Yoo**, and Kyungyong Chung***

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2020-NR049579).

## 요 약

본 논문은 관광지 CCTV 영상에서 군중 병목 현상을 탐지하고 자연어 리포트를 생성하는 모델을 제안한다. 제안 모델은 사람 객체를 탐지·추적한 뒤 공간 그리드 기반 밀집도, 평균 이동속도, 유입-유출 비율을 산출한다. 병목 여부는 이들 조건이 일정 시간 이상 지속될 때에만 판정된다. 또한 검증된 이벤트에 한해 VLM을 활용하여 상황 요약, 위험 등급, 권고 조치를 포함한 리포트를 생성한다. 이러한 구조는 단일 프레임 기반 탐지에서 발생하는 일시적 혼잡과 실제 위험 상황을 보다 안정적으로 구분하게 한다. 정량 메타데이터와 시각 정보를 함께 활용함으로써 현장 관리자가 즉시 이해하고 대응할 수 있는 행동 지향적 관제 정보를 제공한다. 실험 결과, 제안 모델은 F1-score 0.84와 AUC 0.91을 달성하였으며 false positive rate를 최대 71% 감소시켰다.

## Abstract

This paper proposes a model for detecting crowd bottlenecks in tourist-site CCTV videos and generating natural language reports. The model detects and tracks pedestrians, computes grid-based crowd features including density, average speed, and inflow-outflow ratio, and confirms bottlenecks only when these conditions persist over time. For validated events, a VLM generates reports containing a situation summary, risk level, and recommended actions. By combining temporal verification with quantitative metadata, the proposed model improves the reliability of bottleneck detection and provides actionable information for on-site crowd management. Experimental results show that the model achieves an F1-score of 0.84 and an AUC of 0.91, while reducing the false positive rate by up to 71%.

Keywords

crowd bottleneck detection, CCTV analysis, temporal verification, VLM reporting

* 경기대학교 컴퓨터과학과 석사과정
- ORCID: https://orcid.org/0009-0001-3981-3880
** 경기대학교 콘텐츠융합소프트웨어 연구교수
- ORCID: https://orcid.org/0000-0001-9289-2170
*** 경기대학교 AI컴퓨터공학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-6439-9992

· Received: May 06, 2026, Revised: Aug. 09, 2026, Accepted: Aug. 12, 2026
· Corresponding Author: Kyungyong Chung
Dept. of Division of AI Computer Science and Engineering, Kyonggi University, South Korea
Tel.: +82-31-249-1382, Email: dragonhci@gmail.com

# Ⅰ. 서 론

코로나19 이후 외래관광객의 국내 유입이 회복되면서 관광 흐름은 대도시 중심에서 전통시장·한옥마을·해안 산책로 등 로컬 관광지로 확산되고 있으며, 이는 좁은 보행 공간에 인파가 집중되는 군중 병목 위험을 동반한다. 2022년 이태원 참사 이후 군중 안전 관리는 실시간 감지 기술과 지능형 관제 시스템을 필요로 하는 학술적·기술적 의제로 확장되었으며, 공공 통합관제 환경에서 AI 도입이 안전 관리 역량에 미치는 영향에 대한 연구도 활발히 진행되고 있다[1]. 그러나 기존 방범용 CCTV는 사후 녹화 매체로 활용되거나 다수의 카메라를 제한된 인력이 동시 모니터링하는 구조에 의존하여 순간적 군중 병목을 선제적으로 인지하기 어렵다.

이러한 사후적 구조는 두 가지 문제를 낳는다. 첫째, 단일 프레임 기반 탐지는 일시적 occlusion이나 카메라 흔들림을 위험 신호로 오인할 수 있으며, 잦은 오알람은 관제 인력의 알람 피로를 유발하여 실제 위험에 대한 반응성을 둔화시킨다. 둘째, 기존 모델의 출력은 밀집도 수치나 위험 점수와 같은 정량 지표에 머물러, 현장 관리자가 필요로 하는 "무슨 상황인가, 어떤 조치를 취해야 하는가" 와 같은 서술적·행동 지향적 정보로 매개되지 못한다. 이 두 한계는 독립적이지 않다. 오검출이 잦으면 후속 자연어 리포트의 신뢰도도 낮아지며, 탐지 정확도가 높더라도 결과가 행동 가능한 형태로 제공되지 않으면 실제 안전 관리에 활용되기 어렵다.

따라서 본 연구는 관광지 CCTV 영상에서 군중 병목 현상을 시계열적 맥락 속에서 탐지하고, 검증된 위험 이벤트에 한해 영상 프레임과 정량 메타데이터를 VLM(Vision-Language Model)에 함께 입력하여 현장 관리자가 즉각 이해하고 대응할 수 있는 자연어 리포트를 생성하는 통합 모델을 제안한다.

본 연구의 기여는 다음 세 가지이다. 첫째, 셀 단위 밀집도·이동속도·유입유출비를 결합한 다변수 조건을 시계열 연속성 위에서 평가하는 시공간적 통합 병목 판정 알고리즘을 수식 수준에서 정식화한다. 둘째, 영상 프레임과 정량 메타데이터를 함께 활용하는 하이브리드 입력 구조를 통해 VLM의 환각 가능성을 정량 정보로 제약함으로써 고신뢰 도메인에서 VLM을 책임 있게 활용하는 설계 패턴을 제시한다. 셋째, 공개 벤치마크와 한국 로컬 관광지 자체 수집 데이터를 결합한 이중 평가 체계를 통해 모델의 일반화 성능과 도메인 적합성을 동시에 검증하는 평가 프로토콜을 제안한다.

본 논문은 2장에서 관련 선행연구를, 3장에서 제안 모델을, 4장에서 실험 결과를, 5장에서 한계와 결론을 논의한다.

# Ⅱ. 관련 연구

본 장은 본 연구와 관련된 세 갈래의 선행연구—군중 밀집도 추정, 군중 이상행동 및 병목 탐지, Vision-Language Model의 영상 이해 응용—를 검토하고, 각 갈래에서 본 연구의 차별점을 제시한다.

## 2.1 군중 밀집도 추정

군중 밀집도 추정은 단일 영상으로부터 인원수 또는 밀집도 분포를 추정하는 과제로, 초기의 검출 기반 접근에서 회귀 기반 접근, 밀도맵 기반 접근으로 발전해 왔으며, 최근에는 attention 및 transformer 계열 모델이 도입되며 정확도가 지속적으로 개선되어 왔다[2]. 그러나 이들 연구의 출력은 "어디에 얼마나 모여 있는가"라는 정적 분포에 머물며, 군중이 어떤 방향으로 흐르고 있고 어디에서 흐름이 정체되어 위험으로 발전하고 있는가라는 동적 흐름의 진단은 직접적 관심사가 아니다. 본 연구는 카운팅을 출발점으로 삼되, 셀 단위의 시계열적 동역학을 결합하여 정적 분포에서 동적 위험 진단으로 분석 대상을 확장한다.

## 2.2 군중 병목 탐지 및 시계열 검증 접근

군중 병목 탐지는 특정 구역에서 보행 흐름이 정체되고 압력이 누적되는 위험 상태를 조기에 식별하는 과제로, 군중 안전 관리 분야에서 중요하게 다뤄져 왔다. 기존 연구들은 광학흐름, 밀집도 변화, 이동속도 저하와 같은 수작업 지표를 활용하는 방

식에서 시계열 특징을 학습 기반으로 모델링하는 방식으로 발전해 왔으나[3], 많은 경우 단일 프레임 또는 짧은 구간의 즉시적 판정에 의존하거나 이벤트의 지속성 검증 없이 프레임 단위 이상 점수에 기반하여 판정하기 때문에 일시적 가림, 카메라 흔들림, 순간적 인파 집중에 의한 오검출에 취약하다. 특히 관광지 CCTV 환경에서는 보행자의 정지, 방향 전환, 군집 이동이 빈번하므로, 단일 시점의 밀집 상태만으로 실제 병목 위험을 안정적으로 판단하기 어렵다. 본 연구는 이러한 한계를 보완하기 위해 셀 단위 밀집도, 평균 이동속도, 유입유출비를 결합하고, 해당 조건이 일정 시간 이상 지속되는지를 검증하여 병목 여부를 판단한다.

### 2.3 Vision-Language Model의 영상 이해 응용

최근 VLM은 영상 및 이미지에 대한 자연어 설명과 추론이 가능해지면서 감시 영상 해석에도 활용 가능성이 주목받고 있다[4]. 그러나 안전 관제 환경에서는 두 가지 한계가 존재한다. 첫째, 시각 정보만으로 과도한 상황 해석을 생성하는 환각 문제이며, 이를 완화하기 위해 외부 데이터를 참조하여 생성 결과를 정박(grounding)시키는 검색 증강 생성 방식이 안전 정보 도메인에서 활용되고 있다[5]. 둘째, 매 프레임 VLM을 적용할 경우 실시간 처리 비용이 커진다는 점이다. 본 연구는 이 문제를 해결하기 위해 시계열 검증을 통과한 이벤트에 대해서만 VLM을 호출하고, 영상 프레임과 함께 정량 메타데이터를 입력하여 자연어 리포트의 신뢰성과 활용성을 높이고자 한다.

## III. 제안 모델

본 연구가 제안하는 시스템은 관광지 CCTV 영상 스트림을 입력으로 받아, 자연어 기반의 관제 리포트를 출력으로 생성하는 end-to-end 파이프라인이다. 시스템은 네 개의 모듈로 구성된다. 모듈 A는 객체 탐지 및 위치 추정을 담당하며, 모듈 B는 추출된 위치 정보를 공간 그리드에 매핑하여 셀 단위 정량 지표를 산출한다. 모듈 C는 셀 단위 지표의 시계열적 변화를 분석하여 병목 후보 이벤트를 탐지하고 연속성 검증을 수행한다. 모듈 D는 검증을 통과한 이벤트를 시각 프레임 및 정량 메타데이터와 함께 VLM에 전달하여 자연어 리포트를 생성한다.

이 구성은 세 가지 설계 원칙을 따른다. 첫째, 오검출 최소화 우선 원칙(Precision-first)에 따라 시계열 검증을 거친 이벤트만 알람으로 연결한다. 둘째, 계산 비용의 계층화 원칙에 따라 VLM 추론을 검증된 이벤트에만 트리거하여 실시간성과 모델 품질의 균형을 도모한다. 셋째, 정량과 자연어의 결합 원칙에 따라 VLM의 환각 위험을 줄이기 위해 영상 프레임과 정량 메타데이터를 구조화된 형식으로 함께 입력한다.

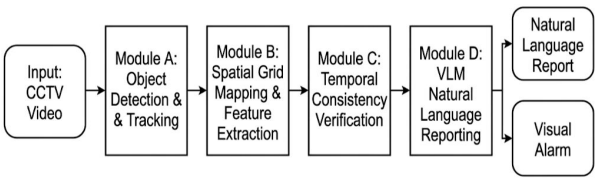


그림 1. 전체 시스템 아키텍처
Fig. 1. Overall system architecture

이하 3.1절에서는 객체 탐지 및 공간 그리드 매핑(모듈 A·B)을, 3.2절에서는 시계열 연속성 검증을 결합한 병목 판정 로직(모듈 C)을, 3.3절에서는 VLM 기반 자연어 리포트 생성(모듈 D)을 차례로 상술한다.

### 3.1 객체 탐지 및 공간 그리드 매핑

본 시스템은 사람 객체의 실시간 탐지를 위해 YOLO 계열을 채택한다. 관광지 CCTV는 광각으로 설치되어 개별 객체가 작게 포착되는 경우가 많으므로 소형 객체 탐지 성능이 중요하며, 동시에 실시간 관제 환경에서는 추론 속도가 보장되어야 한다. 본 연구는 두 요인의 균형점에서 YOLOv8s를 기본 모델로 채택하되 다른 변형은 4장에서 비교 평가한다[6]. 객체의 위치 좌표로는 경계상자의 하단 중심점을 사용하는데, 이는 머리 위치보다 실제 점유 공간을 더 정확히 반영하기 때문이다.

탐지된 객체의 화면 좌표는 사전에 정의된 공간

그리드에 매핑된다. 본 연구는 최근 군중 동역학 연구에서 congestion과 위험 임계 조건으로 보고된 밀집도 수준(약 4 persons/m² 이상)과 정합하는 1m × 1m를 셀 크기로 설정하며[7] 이에 대한 ablation은 4장에서 수행한다. 화면 좌표에서 실세계 평면 좌표로의 변환은 호모그래피 변환으로 수행하며, 카메라가 고정된 관광지 CCTV 환경에서는 일회성 캘리브레이션으로 충분하다.

정량 지표 산출에 앞서, 프레임 간 객체의 연속적 식별과 좌표 정합이 선행되어야 한다. 본 시스템은 YOLOv8s로 탐지된 객체에 대해 ByteTrack 기반의 다중 객체 추적(Multi-object tracking)을 적용하여 각 객체에 고유 ID를 부여하고 프레임 간 매칭을 수행하며, 추적된 각 객체의 하단 중심점은 호모그래피 변환을 통해 실세계 평면 좌표로 매핑되어 셀 소속과 이동 벡터의 기준이 된다. 짧은 occlusion으로 인해 프레임에서 일시적으로 탐지되지 않는 객체는 추적기의 예측 상태를 일정 시간까지 유지하여 ID 불연속을 완화한다.

이 위에서 각 셀 c, 시점 t의 세 가지 정량 지표가 다음과 같이 산출된다. 밀집도 D(c, t)는 셀 c 내 객체 수를 셀 면적으로 나눈 값(persons/m²)으로, 객체의 하단 중심점이 셀 경계선 위에 놓이는 경우 반개구간 규칙 $[x_{\min}, x_{\max}) \times [y_{\min}, y_{\max})$을 적용하여 이중 카운트를 방지하고, 인접 셀 경계 근처에 위치한 객체가 계산에서 배제되지 않도록 셀 경계에 폭 δ의 완충 영역을 두어 후술할 방향성 분석 시 양측 셀에서 참조되도록 한다.

평균 이동속도 V(c, t)는 셀 내 객체들의 이동 벡터로부터 산정되는데, 사람들이 서로 다른 방향으로 이동하거나 복잡한 패턴을 보이는 밀집 상황에서는 단순 벡터 합의 크기만으로는 흐름 정체 여부를 판단하기 어렵다. 예컨대 서로 반대 방향으로 이동하는 두 군집은 개별 속도가 큼에도 순 이동이 0에 가까워 흐름이 과소평가된다. 이를 보완하기 위해 본 시스템은 속도 크기 평균 $V_{mag}(c, t) = (1/N) \Sigma_i \|v_i\|$와 흐름 결맞음도 $\phi(c, t) = \|\Sigma_i v_i\| / \Sigma_i \|v_i\| \in [0, 1]$을 함께 산출하고, 최종 이동속도를 $V(c, t) = V_{mag}(c, t) \cdot \phi(c, t)$로 정의한다. $\phi$가 1에 가까울수록 단일 방향의 정렬된 흐름을, 0에 가까울수록 다방향의 상쇄 상태를 의미하므로, 이 정의는 개별 속도는 크더라도 방향이 상쇄되어 실제 순 이동이 미미한 정체 구간의 눈치 이동이나 밀집 상태의 제자리 진동을 낮은 속도로 반영할 수 있다. 순간 속도의 잡음을 완화하기 위해 개별 이동 벡터 $v_i$는 최근 $T_w$초 창의 이동평균으로 산출한다.

유입-유출 비율 R(c, t)는 시간창 Δt 내 셀 c로 진입한 객체 수 $N_{in}$과 이탈한 객체 수 $N_{out}$의 비 $R(c, t) = N_{in} / \max(N_{out}, 1)$로 정의되며, 카운트는 앞 단계에서 부여된 객체 ID의 셀 소속 변화(시점 t-Δt에는 다른 셀, 시점 t에는 c에 있음, 또는 그 반대)로부터 결정된다. 그러나 관광지 CCTV 환경에서는 사거리 형태의 셀, 병렬 진입 통로, 다방향 관광객 흐름 등으로 인해 하나의 셀에 대해 여러 방향의 유입·유출이 동시에 발생할 수 있으며, 이때 유입과 유출을 단순 스칼라 합으로 처리하면 서로 다른 방향의 흐름이 상쇄되어 특정 방향의 정체 압력이 은폐될 수 있다. 이를 반영하기 위해 본 시스템은 셀의 네 변(동·서·남·북)을 독립적인 경계 세그먼트로 취급하여 각 세그먼트별 유입·유출을 별도 카운트하고, 셀 단위 값은 $N_{in} = \Sigma_e N_{in}$^(e), $N_{out} = \Sigma_e N_{out}$^(e)로 집계하되, 임의의 경계 e에서 $N_{in}$^(e) / max($N_{out}$^(e), 1)이 셀 단위 R에 비해 유의하게 큰 경우 해당 방향의 압력 편중을 함께 기록하여 3.3절 VLM 프롬프트의 인접 셀 상태 요약에 반영한다.

한편 경계선을 따라 평행하게 이동하는 객체는 실제로는 셀을 통과하지 않음에도 위치 흔들림이나 궤적 지터로 인해 순간적으로 소속 셀이 뒤바뀔 수 있고, 경계선 부근을 왕복하는 객체는 짧은 시간에 유입·유출로 반복 카운트될 수 있다. 이를 억제하기 위해 본 시스템은 경계 통과 이벤트에 세 가지 조건을 함께 요구한다. 첫째, 경계 통과 시점의 이동 벡터가 해당 경계 세그먼트에 대해 유효한 수직 성분을 가질 것(경계에 평행한 이동은 카운트에서 배제). 둘째, 셀 소속 변화가 완충 영역 폭 δ를 완전히 넘어서는 실질적 이동일 것(궤적 지터에 의한 미세 위치 이동 배제). 셋째, 동일 객체가 최근 $T_{hys}$초 이내에 같은 셀·같은 방향으로 유입 또는 유출로 이미 카운트된 이력이 없을 것(왕복 재카운트 억제). 이 세 조건에 의해 정지 상태의 자세 조정,

경계선을 따라가는 통행, 짧은 왕복 이동이 유입·유출 이벤트로 오카운트되는 것이 방지된다.

이상 세 지표는 셀 c의 상태 s(c, t) = (D, V, R)로 통합되어 다음 절의 병목 판정 로직의 입력으로 전달된다.

### 3.2 병목 판정 로직

제안 시스템은 병목 현상을 단일 지표의 임계값 초과로 판정하지 않고, 세 지표의 결합 조건이 시계열적으로 지속될 때에만 병목으로 판정한다. 셀 c, 시점 t의 밀집도 D, 평균 이동속도 V, 유입유출비 R이 각 임계값 $\theta_D$, $\theta_V$, $\theta_R$ 조건을 만족할 때 다음과 같이 병목 후보로 정의한다.

$$Candidate(c,t) = D > \theta_D \wedge V < \theta_V \wedge R > \theta_R \quad (1)$$

식 (1)은 밀집도가 임계치를 초과(공간이 차 있음)하는 동시에 평균 속도가 떨어지고(흐름이 정체) 유입이 유출을 초과(압력이 누적)하는 상태이며, 단일 지표 기반 판정이 놓치는 압력 누적 상태를 포착한다. 각 임계값은 두 방법으로 결정될 수 있으며 본 연구는 이를 병행 비교한다. 도메인 기반 결정은 최근 군중 동역학 empirical 연구에서 보고된 congestion 임계 조건－보행 가능 한계 약 4 persons/m², 평균 보행 속도 1.2~1.4m/s, 정체 시 절반 이하 수준－을 근거로 $\theta_D$ = 4.0, $\theta_V$ = 0.6, $\theta_R$ = 1.5로 설정한다[7]. 데이터 기반 결정은 검증 데이터셋의 ground truth 라벨에 대한 ROC 분석에서 Youden's J statistic을 최대화하는 임계값을 도출하며, 데이터 분포에 적응적이지만 학습 데이터 편향에 영향받을 수 있다. 4장에서는 두 방법의 성능을 비교한다.

단일 시점 병목 후보가 즉시 알람으로 이어지면 일시적 occlusion이나 카메라 흔들림에 의한 오검출이 발생하므로, 본 연구는 시계열 연속성 검증을 도입한다. 확정 병목 판정 조건은 식 (2)와 같으며, 셀 c에서 연속된 K 프레임 동안 Candidate(c, t) = 1이 유지될 때에만 시점 t + K - 1에서 c를 확정 병목으로 판정한다.

$$Bottleneck(c,t) = \prod_{k=0}^{K-1} Candidate(c,t-k) \quad (2)$$

식 (2)의 K-연속 검증 규칙을 베이스라인 검증 기법(옵션 A)으로 두고, 본 연구는 두 가지 대안을 함께 평가한다. 옵션 B는 셀의 (D, V, R) 시계열을 입력으로 하는 경량 시퀀스 모델(1D-TCN 또는 LSTM)을 학습하여 시간 패턴의 비선형적 결합을 학습하는 방법이고, 옵션 C는 인접 셀의 동시 정체 등 셀 간 공간적 상관을 함께 고려하는 ST-GCN 계열의 그래프 기반 검증이다[8]. 구체적으로 각 그리드 셀을 그래프의 노드로 정의하고, 노드 특징은 해당 셀의 (D, V, R) 삼차원 지표 벡터로 구성한다. 공간 엣지는 셀 격자에서 상하좌우로 인접한 셀(4-이웃) 간에 연결하되, 대각 방향 인접까지 포함하는 8-이웃 구성은 4장 예비 실험에서 비교하였다. 엣지 가중치는 이진 인접 여부로 고정하지 않고, 두 인접 셀 간 유입-유출 흐름의 연속성을 반영하도록 3.1절에서 산출한 경계 세그먼트별 통과량에 비례하여 부여함으로써, 실제 보행 흐름으로 연결된 셀 간 상관이 더 강하게 전파되도록 한다. 시간 축으로는 동일 셀의 연속된 프레임 노드를 시간 엣지로 연결하여 시공간 그래프를 구성하며, 이 위에서 그래프 합성곱과 시간 합성곱을 교대로 적용하여 인접 셀의 동시 정체 패턴과 시계열 지속성을 함께 학습한다. K가 작으면 응답성은 빠르나 오검출이 늘고 크면 안정적이나 알람 지연이 증가하므로, 본 연구는 K ∈ {3, 5, 7, 10, 15}에 대한 grid search와 함께 세 옵션의 비교를 4장 ablation에서 수행한다.

### 3.3 VLM 기반 자연어 리포트 생성

확정 병목이 검출된 시점에 한해 VLM이 호출되어 자연어 리포트가 생성된다. 영상 직접 처리 모델(Video-LLaVA 등)은 시간적 맥락을 내부 통합할 수 있으나 추론 비용과 한국어 출력 편차가 크므로, 본 연구는 정적 이미지 VLM에 핵심 프레임 샘플링을 결합하고 시간적 맥락은 정량 메타데이터로 매개하는 방식을 채택한다[9]. 영상 직접 처리 모델은 비교 베이스라인으로 4장에서 평가하며, 정적 이미지

VLM으로는 Qwen2-VL과 InternVL을 예비 실험으로 비교한 뒤, 최종적으로 Qwen2-VL을 리포트 생성 모델로 선정하였다[10].

본 연구의 핵심 설계 결정은 VLM에 영상 프레임만 제공하지 않고 모듈 B·C에서 산출한 정량 메타데이터를 구조화된 형식으로 함께 제공한다는 점이다. 프롬프트의 일반 형식은 다음과 같다.

[입력 프레임]
[메타데이터]
- 위치: {camera_id}, {zone_label}
- 밀집도: {D} persons/m² (임계 {Θ_D} 초과)
- 평균 이동속도: {V} m/s (임계 {Θ_V} 미만)
- 유입유출비: {R} (임계 {Θ_R} 초과)
- 지속 시간: {duration} 초
- 인접 셀 상태: {neighbor_summary}
[지시]
위 영상과 메타데이터를 바탕으로 다음을 생성하라.
1) 상황 요약 (한국어 1문장, 25자 이내)
2) 위험 등급 (주의 / 경고 / 긴급 중 택일)
3) 권고 조치 (한국어, 행동 지향적 1~2개)

이 설계는 두 효과를 의도한다. 환각 억제 측면에서, VLM이 시각적으로만 추론할 경우 군중 패닉 등 과대 해석이 발생할 수 있으나 정량 메타데이터가 사실적 정박점을 제공함으로써 출력이 데이터에 기반하도록 유도된다. 출력 구조화 측면에서, 자유 형식 캡션 대신 세 항목으로 구조화된 출력을 요구함으로써 후속 시스템(대시보드, 모바일 알림)에서의 자동 파싱이 용이해진다. 이러한 계층화된 호출 구조는 평균 계산 비용을 매 프레임 VLM 호출 대비 크게 절감할 것으로 기대되며, 정확한 절감률과 자연어 출력의 품질 평가는 4장에서 보고한다.

## Ⅳ. 실험 결과 및 분석

본 장에서는 3장에서 제안한 모델의 성능을 검증하기 위한 실험 설계와 결과를 보고한다. 4.1절에서 데이터셋과 평가 프로토콜을, 4.2절에서 평가 지표를, 4.3절에서 베이스라인과 구현 세부사항을, 4.4절에서 주요 결과와 정성 분석을 함께 제시한다.

### 4.1 데이터셋 및 평가 프로토콜

본 연구는 제안 모델의 일반화 성능과 관광지 환경에 대한 적용 가능성을 함께 검증하기 위해 공개 벤치마크와 자체 수집 데이터를 병행하여 사용하였다. 공개 데이터셋으로는 군중 이상행동 및 보행자 영상 분석에 활용되는 UCF-Crime과 ShanghaiTech Campus 데이터셋을 사용하였으며, 이 중 군중 밀집, 흐름 정체, 국소 혼잡, 병목 발생 전후 구간을 선별하였다[11]. 자체 수집 데이터는 전통시장, 한옥마을, 해안 산책로 등 국내 로컬 관광지 CCTV 영상을 대상으로 구성하였으며, 공개 데이터셋에 부족한 좁은 보행 공간, 불규칙한 이동 흐름, 관광객 정지 행동 등을 반영하기 위해 활용하였다.

구체적으로, 공개 데이터셋에서는 총 96개 영상, 총 18.5시간 규모의 데이터를 사용하였고, 자체 수집 데이터는 8대의 CCTV로부터 확보한 72개 영상, 총 21.8시간 규모로 구성하였다. 자체 수집에 사용된 CCTV는 지면으로부터 약 3~5m 높이에 하향 약 15~30°의 부감 각도로 설치되었으며, 영상 해상도는 대체로 FHD(1920×1080)급, 프레임률은 약 25~30fps 수준이다. 촬영 시간대는 주간과 야간을 함께 포함하되, 야간·저조도 구간은 가로등 및 상가 조명에 의존하는 조건에서 수집되었다. 이러한 설치 조건은 관광지 CCTV의 일반적 운용 환경을 반영하며, 호모그래피 캘리브레이션과 셀 그리드 매핑의 기준이 된다. 전체 데이터에서 병목 이벤트 구간은 214개, 정상 구간은 436개로 구성하였다. 전 병목 라벨은 두 명의 라벨러가 독립적으로 부여하였으며, 불일치 구간은 제3자의 검토를 통해 조정하였다. 병목은 일정 구역에서 밀집도 증가, 평균 이동속도 저하, 유입량이 유출량을 초과하는 상태가 일정 시간 이상 지속되는 경우로 정의하였다. 데이터는 영상 단위로 학습 60%, 검증 20%, 테스트 20%로 분할하였으며, 동일 카메라에서 촬영된 영상이 서로 다른 분할에 동시에 포함되지 않도록 구성하여 카메라 종속적 편향을 최소화하였다. 또한 개인정보 보호를 위해 얼굴 등 식별 가능한 영역은 비식별화 처리하였다.

공개 데이터셋의 클립은 화면 내 동시 등장 인원, 부감 각도, 해상도, 연속 길이의 네 조건을 사

전에 정의된 최소 기준을 모두 충족하는 후보에서 선별하였으며, 병목 라벨은 밀집도 증가·평균 보행 속도 저하·특정 방향 압력 축적의 세 조건이 일정 시간 이상 함께 지속되는 구간에 대해서만 부여하고, 세 조건 중 일부만 관찰되는 경계 사례는 별도로 분리하여 학습에서는 제외하였다. 두 라벨러의 프레임 단위 이진 라벨에 대한 Cohen's κ는 substantial agreement 수준이었으며, 불일치 구간은 제3자 재검토를 거쳐 합의 기반으로 확정하여 개별 라벨러의 편향이 최종 라벨에 단독으로 반영되지 않도록 하였다.

### 4.2 평가 지표

본 연구는 병목 탐지 성능, 자연어 리포트 품질, 실시간 처리 성능의 세 측면에서 모델을 평가하였다. 병목 탐지 성능은 Precision, Recall, F1-score, AUC를 사용하여 측정하였으며, 시계열 연속성 검증의 효과를 확인하기 위해 단일 프레임 판정 대비 false positive rate 감소율을 함께 비교하였다. 자연어 리포트 품질은 자동 평가와 인간 평가를 병행하였다. 자동 평가에는 BLEU-4, ROUGE-L, BERTScore를 사용하였고, 인간 평가는 정확성, 명료성, 행동가능성의 세 항목을 5점 척도로 평가하였다. 실시간성은 모듈별 처리 시간, 종단간 평균 FPS, 병목 발생부터 리포트 생성까지의 알람 지연시간을 기준으로 측정하였다.

인간 평가에는 CCTV 관제 실무 경력자와 안전관리 전공 대학원생으로 구성된 복수의 평가자가 참여하였고, 각 평가자는 모델 식별 정보가 제거된 세 모델의 리포트를 무작위 순서로 제시받아 동일 이벤트 집합에 대해 평정을 수행하였다. 평가자 간 신뢰도는 ordinal Krippendorff's α로 산출하였으며, 세 평가 항목 모두 문헌에서 acceptable로 보고되는 기준을 만족하는 수준이었다.

### 4.3 베이스라인 및 구현 세부사항

본 연구는 제안 모델의 효과를 검증하기 위해 병목 탐지 성능과 자연어 리포트 품질 두 측면에서 베이스라인을 구성하였다. 병목 탐지 비교에는 단순 임계값 기반(Naive Threshold), 광학흐름 기반(Farneback), 약지도학습 기반 video-level 이상 탐지 모델을 사용하였으며, 자연어 리포트 비교에는 메타데이터 없이 영상 프레임만 입력하는 Frame-only VLM과 영상 시퀀스를 직접 입력하는 Video VLM을 사용하였다.

구현 세부사항은 다음과 같다. 객체 탐지 모듈은 COCO 사전학습 가중치를 사용한 YOLOv8s를 기반으로 하였으며, 광학흐름 기반 베이스라인에는 OpenCV의 Farneback 알고리즘을 적용하였다[12]. 시계열 검증의 옵션 B는 1D-TCN 구조를, 옵션 C는 셀을 노드로 인접 셀 관계를 엣지로 구성한 ST-GCN 기반 검증 모델을 사용하였다. 모든 실험은 Windows 11 기반 워크스테이션에서 수행되었으며, VLM 모듈은 예비 실험에서 한국어 출력의 일관성과 구조화된 메타데이터 반영 성능을 비교한 결과 Qwen2-VL을 최종 선정하였다[13]. 프롬프트는 3.3절에서 제시한 구조화 형식을 따르며, 출력 일관성을 위해 낮은 temperature 값을 사용하였다.

객체 추적에는 ByteTrack을 사용하였고, 추적 상태 유지 시간은 $\tau_{miss}$ = 0.5초로 설정하였다. 3.1절에서 도입한 셀 경계 완충 폭은 $\delta$ = 0.1m, 이동 벡터의 시간평균 창은 $T_w$ = 1초, 유입-유출 시간창은 $\Delta t$ = 2초, 유입-유출 히스테리시스 창은 $T_{hys}$ = 1초로 설정하였으며, 경계 통과의 유효 수직 성분 기준은 0.05 m/frame으로 두었다.

재현성 확보를 위해 각 모듈의 학습 및 추론 설정을 별도로 기술한다. YOLOv8s는 COCO 사전학습 가중치를 자체 수집 라벨 서브셋으로 소수 epoch 추가 fine-tuning하여 사용하였고, 옵션 B의 1D-TCN은 소수의 residual block과 dilation 조합으로, 옵션 C의 ST-GCN은 공간-시간 합성곱 블록의 stack으로 각각 경량 구조를 채택하였으며, 두 시계열 검증 모델 모두 AdamW optimizer와 focal loss, 클래스 불균형 대응을 위한 weighted sampler를 적용하고 validation F1 기준의 early stopping으로 최종 가중치를 선택하였다. Qwen2-VL은 로컬 FP16 추론으로 운용하되, 확정 시점 전후의 소수 프레임을 shot으로 입력하고 생성 파라미터는 결정론성을 확보하는

방향으로 낮게 설정하였으며, 전체 프롬프트는 부록에 수록하였다. 모듈별 처리시간은 warm-up 이후 반복 실행의 평균으로 측정하였고, 학습 기반 모듈이 포함된 실험은 서로 다른 random seed와 카메라 단위 교차검증을 결합한 복수의 독립 실행으로 수행하여 이하 4.4절의 성능값을 평균 기준으로 보고한다. 사용한 하드웨어, optimizer 및 loss 하이퍼파라미터, VLM 생성 파라미터의 구체값은 재현성 확보를 위해 부록에 표로 정리하였다.

## 4.4 주요 결과

본 절에서는 병목 탐지 성능, 자연어 리포트 품질, 실시간 처리 성능, 그리고 ablation study 결과를 차례로 보고한다. 병목 탐지 성능 비교 결과는 표 1과 같다.

표 1. K=7 기준 병목 탐지 성능 및 FPR 감소율
Table 1. Bottleneck detection performance and FPR reduction at K=7

| Model | Precision | Recall | F1 | AUC | FPR reduction |
|---|---|---|---|---|---|
| Naive threshold | 0.58 | 0.81 | 0.68 | 0.74 | - |
| Optical flow baseline | 0.64 | 0.72 | 0.68 | 0.77 | - |
| Deep learning baseline | 0.76 | 0.79 | 0.77 | 0.84 | - |
| Proposed (Option A) | 0.84 | 0.78 | 0.81 | 0.88 | 64% |
| Proposed (Option B) | 0.87 | 0.80 | 0.83 | 0.90 | 69% |
| Proposed (Option C) | 0.86 | 0.82 | 0.84 | 0.91 | 71% |

옵션 C는 F1 0.84, AUC 0.91로 가장 높은 성능을 보였으며, 옵션 B와 옵션 C는 단일 프레임 판정 대비 각각 69%, 71%의 FPR 감소를 달성하였다. 옵션 간 F1 차이의 통계적 유의성을 paired bootstrap test와 다중 비교 보정을 함께 적용하여 검증한 결과, 옵션 C는 딥러닝 베이스라인 및 옵션 A 대비 유의한 성능 향상을 보인 반면 옵션 B와의 차이는 유의성이 확인되지 않아, 뒤이어 서술하는 옵션 B의 실용적 대안 가능성이 통계적으로도 뒷받침됨을 알 수 있다. 이는 시계열 연속성 검증이 오검출 억제에 효과적으로 작용함을 의미하며, 옵션 B와 C의 근소한 차이는 계산 비용을 고려할 경우 경량 시계열 모델 역시 실용적 대안이 될 수 있음을 시사한다. 자연어 리포트 품질 평가 결과는 표 2와 같다.

표 2. 자연어 리포트 품질 평가
Table 2. Natural language report quality evaluation

| Model | BLEU-4 | ROUGE-L | BERT score | Accuracy | Clarity | Actionability |
|---|---|---|---|---|---|---|
| Frame only VLM | 0.21 | 0.38 | 0.79 | 3.2 | 3.6 | 2.9 |
| Video VLM | 0.26 | 0.43 | 0.82 | 3.5 | 3.7 | 3.4 |
| Proposed model | 0.34 | 0.51 | 0.87 | 4.3 | 4.1 | 4.2 |

제안 모델은 자동 평가 지표와 인간 평가 모두에서 가장 높은 점수를 보였으며, 특히 행동가능성 점수에서 Frame-only VLM 대비 1.3점의 격차를 보였다. 이는 정량 메타데이터가 VLM 출력을 단순 상황 설명이 아니라 현장 관리자가 즉시 수행할 수 있는 조치 중심의 리포트로 유도하고, 동시에 시각 정보만으로 인한 환각 사례를 줄이는 효과를 함께 가져왔기 때문으로 해석된다. 실시간 처리 성능 측면에서, 모듈 A·B·C의 매 프레임 처리 시간은 각각 18ms, 3ms, 5ms로 총 26ms 수준이었다. VLM 추론은 호출 1회당 약 1.4초가 소요되었으나 시계열 검증을 통과한 이벤트에 대해서만 호출하므로 전체 프레임 대비 호출률은 약 1.2%로 제한되었고, 종단간 평균 처리 속도는 약 27 FPS, 평균 알람 지연시간은 2.8초로, 일반적인 CCTV 관제 환경에서 요구되는 실시간 조건을 충족한다. 다만 위 수치는 단일 병목 이벤트를 기준으로 한 것으로, 관제 구역 내 다수의 셀에서 병목이 동시에 확정될 경우 VLM 호출이 직렬 큐에 누적되어 개별 리포트의 지연이 증가할 수 있다. 본 실험 환경에서는 시계열 검증에 의해 순간적 오검출이 걸러지면서 동시 확정 이벤트 수가 대체로 소수에 머물렀고, 위험 등급이 높은 이벤트를 우선 처리하는 단순 우선순위 큐를 적용하여 평균 지연 증가를 억제하였다. 그러나 대규모 행사 등으로 동시 병목이 급증하는 상황에서는 이

러한 단일 인스턴스 직렬 처리만으로는 한계가 있으며, 이에 대한 구조적 대응은 5장에서 논의한다. 각 구성 요소의 독립적 효과를 확인하기 위한 ablation study 결과는 표 3과 같다.

표 3. 셀 크기 및 검증 옵션에 따른 성능 비교
Table 3. Performance by cell size and verification option

| Cell size | Option | F1 | FPR | Actionability score |
|---|---|---|---|---|
| 0.5m × 0.5m | A | 0.68 | 0.24 | 3.9 |
| 0.5m × 0.5m | B | 0.75 | 0.19 | 4.0 |
| 0.5m × 0.5m | C | 0.81 | 0.14 | 4.1 |
| 1m × 1m | A | 0.81 | 0.13 | 4.1 |
| 1m × 1m | B | 0.83 | 0.11 | 4.2 |
| 1m × 1m | C | 0.84 | 0.09 | 4.2 |
| 2m × 2m | A | 0.74 | 0.15 | 3.7 |
| 2m × 2m | B | 0.75 | 0.14 | 3.8 |
| 2m × 2m | C | 0.76 | 0.13 | 3.8 |

셀 크기와 검증 옵션의 상호작용을 살펴보면, 셀 크기가 0.5m × 0.5m로 작을 때 옵션 A의 F1은 0.68에 그친 반면 옵션 C는 0.81로 상승하여 두 옵션 간 격차가 0.13에 이르렀다. 이는 셀이 작을수록 개별 객체 위치 흔들림과 경계 왕복이 지표에 미치는 영향이 커지므로, 셀 자체의 시계열만 참조하는 옵션 A로는 잡음을 충분히 억제하기 어렵고, 인접 셀 간 상관을 활용하는 옵션 C가 상대적 이득을 크게 얻기 때문으로 해석된다. 반대로 셀 크기가 2m × 2m로 커지면 세 옵션의 F1 차이가 0.02 이내로 축소되었으며, 이는 큰 셀에서는 이미 셀 내부의 공간 평균화가 잡음을 흡수하여 인접 셀 정보의 추가 이득이 제한적임을 의미한다. 1m × 1m 셀에서는 세 옵션 모두 F1 0.81 이상을 유지하며 옵션 C가 F1 0.84, FPR 0.09로 가장 높은 성능을 보였고, 셀 크기와 옵션 간 상호작용을 고려할 때 1m × 1m 셀은 응답성·정확도·계산 비용의 균형점에 해당한다.

한편 검증 축을 제외하고 다른 구성 요소를 개별적으로 제거한 결과에서도 각 요소의 기여가 확인되었다. Temporal Verification을 제거한 조건에서는 F1이 0.84에서 0.71로 감소하고 FPR이 0.09에서 0.27로 증가하여, 시계열 연속성 검증이 본 모델에서 가장 중요한 안정화 요소임이 확인되었다. Multivariate Combination을 제거하고 밀집도만 사용할 경우에도 F1이 0.74로 감소하여 병목이 단순 밀집이 아니라 밀집도·속도·유입유출의 복합 상태로 판단되어야 함을 보여준다. 반면 Metadata를 제거한 조건은 F1에는 큰 영향을 주지 않았으나 Actionability Score를 4.2에서 2.9로 낮추었으며, 이는 메타데이터 결합의 효과가 탐지 정확도가 아닌 자연어 리포트의 실용성에 직접적으로 작용함을 의미한다. 마지막으로 모델의 작동 양상을 사례 기반으로 검토한다. 성공 사례 시각화는 그림 2와 같다.

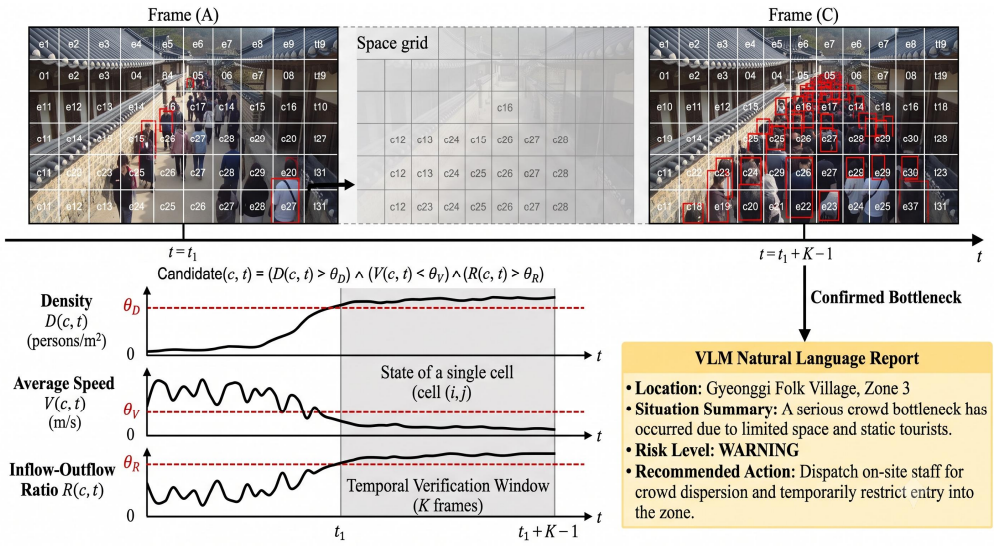


그림 2. 성공 사례 시점별 (D, V, R)변화와 시스템 출력
Fig. 2. Temporal changes in (D, V, R) and system output for a successful case

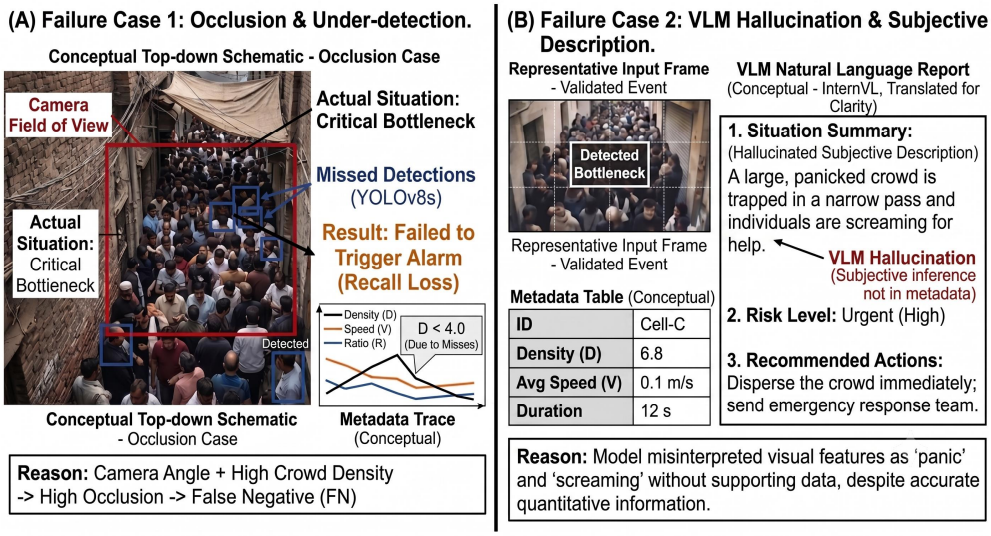


그림 3. 실패 사례(occlusion 사례 및 VLM 환각 사례)
Fig. 3. Failure cases (Occlusion and VLM hallucination)

성공 사례에서는 단일 임계값 베이스라인이 일시적 인파 변동으로 인해 알람이 해제되었다가 재발생하는 불안정한 패턴을 보인 반면, 제안 모델은 시계열 검증을 통해 압력 누적이 본격화되는 시점에 안정적으로 알람을 생성하였다. Qwen2-VL이 생성한 자연어 리포트는 위치, 밀집도, 권고 조치를 구조화된 형식으로 제시하였으며, 인간 평가자는 이를 현장 관리자가 즉시 행동에 옮길 수 있는 수준으로 평가하였다. 실패 사례 시각화는 그림 3과 같다.

실패 사례는 두 유형으로 나타났다. 첫째, 좁은 골목에서 군중이 카메라에 가까이 접근할 때 객체 탐지 성능이 저하되어 실제 밀집도가 임계를 초과함에도 모델이 이를 포착하지 못하는 occlusion 사례가 관찰되었다. 둘째, 메타데이터를 결합했음에도 일부 사례(전체 출력의 약 3%)에서 VLM이 메타데이터에 명시되지 않은 군중의 감정 상태에 대한 단정적 추론을 출력하는 환각 사례가 발견되었다. 이러한 환각은 주로 세 가지 요인에서 유발되었다. 첫째, 밀집도가 임계를 크게 상회하는 고밀집 프레임에서 VLM이 시각적 혼잡도를 "패닉", "공포" 등 감정 서술로 과대 해석하는 경향이 관찰되었다. 둘째, 제공된 메타데이터가 밀집도·속도·유입유출비 등 물리적 지표에 한정되어 군중의 정서 상태에 관한 근거를 담지 않음에도, VLM이 학습 과정에서 획득한 언어적 사전 지식에 의존해 명시되지 않은 정보를 보충 생성하는 경향이 확인되었다. 셋째, 상황 요약을 한 문장으로 압축하도록 요구한 프롬프트 제약이 오히려 인과적 단정을 촉진하는 경우가 있었다. 즉 환각은 시각 정보의 과대 해석과 언어 모델의 사전 편향이 결합하여 발생하며, 정량 메타데이터의 정박 효과가 이를 상당 부분 억제하나 완전히 제거하지는 못함을 보여준다. 전자는 다중 카메라 융합으로, 후자는 출력 후처리 필터로 완화 가능하다.

후자의 완화 가능성을 사전적으로 확인하기 위해 5장에서 제안하는 어휘 필터를 파일럿으로 적용한 결과, 감정·심리 상태를 단정하는 어휘를 중립적 서술로 치환하는 것만으로 잔존 환각 사례의 비율이 유의하게 감소하였고, 자동 지표와 인간 평가로 측정한 리포트 품질에는 통계적으로 유의한 저하가 관찰되지 않아 향후 후처리 계층 도입의 실현 가능성을 시사한다.

## V. 결 론

본 연구는 시계열 연속성 검증과 VLM 기반 자연어 리포트 생성을 결합한 군중 병목 탐지 모델을 제안하였다. 실험 결과, 제안 모델은 단일 프레임 기반 판정 대비 false positive rate를 최대 71% 감소시키면서도 행동 지향적 자연어 리포트 생성에서 우수한 성능을 보였으며, 계층화된 호출 구조를 통해 실시간 관제 환경에 적합한 처리 효율을 확보하였다.

다만 occlusion 상황에서의 객체 탐지 성능 저하와 일부 잔존 환각 가능성, 다수 병목이 동시다발적으로 발생할 때의 VLM 추론 지연, 그리고 야간·우천·축제 등 특수 환경에서의 추가 검증 필요성은 본 연구의 한계이다.

향후 연구에서는 다중 카메라 융합으로 occlusion 문제를 완화하고, 다양한 환경 조건의 데이터셋으로 확장하는 방향이 요구된다. 환각 억제 측면에서는 VLM 출력 후처리 검증 모듈을 세 층위로 설계할 수 있다. 첫째, 감정·심리 상태를 단정하는 어휘("패닉", "공포", "폭동" 등)를 사전으로 관리하여 해당 표현이 포함된 출력을 차단하거나 중립적 서술로 대체하는 규칙 기반 어휘 필터이다. 둘째, VLM이 생성한 서술을 입력 메타데이터와 대조하여 정량 근거가 없는 주장을 걸러내는 사실 정합성 검증으로, 출력의 각 진술이 제공된 지표 범위 안에서 지지되는지를 확인한다. 셋째, 위험 등급과 밀집도·속도 등 정량값의 대응 관계를 규칙으로 정의하여 메타데이터와 상충하는 등급 판정을 교정하는 일관성 검증이다. 이러한 후처리 계층은 VLM 재호출 없이 경량 규칙과 소규모 검증 모델만으로 동작하므로 3.3절의 계층화된 호출 구조와 정합하며, 실시간성을 훼손하지 않고 잔존 환각을 억제할 수 있을 것으로 기대된다.

처리 지연 측면에서는, 동시다발적 병목 상황에 대응하기 위해 위험 등급 기반 우선순위 스케줄링을 정교화하고, 다수 이벤트를 하나의 프롬프트로 묶는 배치 추론과 VLM 인스턴스의 병렬·분산 처리를 결합하며, 저위험 이벤트에 대해서는 VLM 호출 없이 정량 메타데이터만으로 템플릿 기반 리포트를 생성하는 경량 경로를 두어 호출 부하를 분산하는 방안이 요구된다.

본 연구의 결과는 안전 관리 인프라의 자동화가 단순한 탐지 정확도 향상의 문제가 아니라, 탐지된 위험을 현장 관리자의 행동으로 매개하는 정보 설계의 문제임을 보여준다.

## References

[1] S. Yeo, "Effects of Perceived Data Quality on Safety Management Capability in AI Adoption for Public Integrated Control Centers: Structural Equation Modeling with Acceptance Intention as a Mediator", Journal of Artificial Intelligence Convergence Technology, Vol. 6, No. 1, pp. 25-31, Mar. 2026. https://doi.org/10.23374/jaict.2026.6.1.004.

[2] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 1091-1100, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00120.

[3] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning", Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, pp. 4975-4986, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00493.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision", Proc. International Conference on Machine Learning (ICML), Online, pp. 8748-8763, Jul. 2021. https://doi.org/10.48550/arXiv.2103.00020.

[5] S. Kim, Y. Kim, T. Lee, R. Kim, and H. An, "Design and Implementation of a Public Data-Based Diplomatic and Safety Information Platform Utilizing Retrieval-Augmented Generation", Journal of Artificial Intelligence Convergence Technology, Vol. 6, No. 1, pp. 38-48, Mar. 2026. https://doi.org/10.23374/jaict.2026.6.1.006.

[6] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From

YOLOv1 to YOLOv8 and YOLO-NAS", Machine Learning and Knowledge Extraction, Vol. 5, No. 4, pp. 1680-1716, Nov. 2023. https://doi.org/10.3390/make5040083.

[7] C. Feliciani and K. Nishinari, "Measurement of Congestion and Intrinsic Risk in Pedestrian Crowds", Transportation Research Part C: Emerging Technologies, Vol. 91, pp. 124-155, Jun. 2018. https://doi.org/10.1016/j.trc.2018.03.027.

[8] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", Proc. AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, Vol. 32, No. 1, pp. 7444-7452, Feb. 2018. https://doi.org/10.1609/aaai.v32i1.12328.

[9] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning United Visual Representation by Alignment Before Projection", Proc. 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, FL, USA, pp. 5971-5984, Nov. 2024. https://doi.org/10.18653/v1/2024.emnlp-main.342.

[10] P. Wang, et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution", arXiv preprint arXiv:2409.12191, Sep. 2024. https://doi.org/10.48550/arXiv.2409.12191.

[11] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 6479-6488, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00678.

[12] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion", Proc. 13th Scandinavian Conference on Image Analysis (SCIA), Gothenburg, Sweden, Lecture Notes in Computer Science, Vol. 2749, pp. 363-370, Jun. 2003. https://doi.org/10.1007/3-540-45103-X_50.

[13] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 24185-24198, Jun. 2024. https://doi.org/10.48550/arXiv.2312.14238.

## 저자소개

서 태 영 (Taeyoung Seo)

2026년 3월 ~ 현재 : 경기대학교 컴퓨터과학과 석사과정
관심분야 : 데이터마이닝, 딥러닝, 컴퓨터비전, 인공지능, 빅데이터

유 현 (Hyun Yoo)

1999년 2월 : 상지대학교 전산학과(이학사)
2011년 8월 : 상지대학교 컴퓨터교육학과(교육학석사)
2019년 8월 : 상지대학교 컴퓨터정보공학과(공학박사)
2020년 7월 ~ 현재 : 경기대학교 콘텐츠융합소프트웨어 연구소 연구교수
관심분야 : 딥러닝, 인공지능, 빅데이터 마이닝

정 경 용 (Kyungyong Chung)

2000년 2월 : 인하대학교 전자계산공학과 (공학사)
2002년 2월 : 인하대학교 전자계산공학과 (공학석사)
2005년 8월 : 인하대학교 컴퓨터정보공학부 (공학박사)
2006년 3월 ~ 2017년 2월 : 상지대학교 컴퓨터정보공학부 교수
2017년 3월 ~ 현재 : 경기대학교 AI컴퓨터공학부 교수
관심분야 : 데이터 마이닝, 헬스케어, 빅데이터, HCI, 지능시스템, 인공지능, 추천 시스템