논문 2026-2-6 http://dx.doi.org/10.29056/jsf.2026.06.06

VLM 기반 CCTV 관제 영상 캡션의 개인정보 유출 평가 및 접근 권한 기반 추상화 효과 분석

한승완*, 강승호*†

Privacy Leakage Evaluation and the Effect of Access-Controlled Abstraction for VLM-based CCTV Surveillance Captions

Seungwan Han*, Seung-Ho Kang*†

요 약

CCTV 관제 시스템에 VLM이 도입되면서 영상 내용을 자연어 캡션으로 생성·검색할 수 있게 되었으나, 캡션 자체가 새로운 텍스트 기반 개인정보 유출 채널이 될 수 있다. 본 연구는 VLM 기반 CCTV 캡션의 의미론적 개인정보 유출을 정량화하기 위해 Semantic Leakage Score(SLS), 8개 민감 의미 속성, A0~A3 접근 권한 기반 추상화 정책을 제안한다. UCA(UCF-Crime Annotation) 영문 캡션 300개 실험에서 A2는 SLS를 0.118에서 0.013으로 89.0% 감소시키면서도 약한 키워드 분류 기준선의 유용성을 A0와 유사하게 유지하였다. 반면 A3는 SLS를 0으로 낮추지만 분류 성능이 12개 클래스 무작위 추측 수준으로 하락하였다. AI Hub 보조 사례에서는 정보 밀도가 높은 한국어 캡션에서 A2 이후에도 잔여 노출이 발생할 수 있음을 확인하였다. 본 연구는 접근 권한 기반 추상화가 VLM 기반 CCTV 관제에서 유용성과 프라이버시를 균형 있게 조정하는 실용적 방법임을 보인다.

Abstract

As Vision-Language Models (VLMs) are adopted in CCTV surveillance systems, video content can be generated and retrieved as natural-language captions. However, such captions may also become a new text-based privacy leakage channel. This study proposes the Semantic Leakage Score (SLS), an eight-attribute sensitive semantic taxonomy, and a four-level access-controlled abstraction policy from A0 to A3 to quantify and mitigate semantic privacy leakage in VLM-based CCTV captions. Experiments on 300 English UCA(UCF-Crime Annotation) captions show that A2 reduces SLS from 0.118 to 0.013, an 89.0% reduction, while maintaining a weak keyword-matching utility baseline comparable to A0. In contrast, A3 reduces SLS to zero but lowers classification accuracy to the 12-class random-guessing level. An auxiliary AI Hub case study further shows that residual exposure may remain in information-dense Korean captions even after A2 abstraction. These results indicate that access-controlled abstraction is a practical method for balancing surveillance utility and privacy in VLM-based CCTV systems.

한글키워드 : CCTV 관제, 시각-언어 모델, 의미론적 개인정보 유출, 텍스트 비식별화, 접근 권한 기반 추상화

keywords : CCTV surveillance, Vision-Language Model, semantic privacy leakage, textual de-identification, access-controlled abstraction

* 국립목포대학교 컴퓨터학부
† 교신저자: 강승호(email: shkang@mnu.ac.kr)

접수일자: 2026.05.18. 심사완료: 2026.06.09.
게재확정: 2026.06.20.

1. 서 론

CCTV 관제 시스템은 공공안전, 시설보안, 재난·범죄 예방을 위해 광범위하게 활용되며, 최근에는 시각-언어 모델(Vision-Language Model, VLM)의 발전에 따라 영상 내용을 자연어 캡션으로 요약하거나 자연어 질의로 특정 장면·사건을 검색하는 지능형 관제 시스템으로 확장되고 있다[1]. CCTV 영상 이상행동 탐지는 UCF-Crime (University of Central Florida-Crime) 기반 실세계 이상행동 탐지 연구를 통해 대표적인 벤치마크 문제로 정립되었고[2], 최근에는 UCA(UCF-Crime Annotation)와 같이 감시 영상-언어 이해(Surveillance Video-and-Language Understanding, Surveillance VALU), 영상 캡션 생성, 멀티모달 이상행동 탐지를 포괄하는 영상-언어 어노테이션 및 벤치마크가 제안되었다[3].

그러나 VLM 기반 관제 시스템은 기존 영상 기반 시스템과 다른 새로운 개인정보 유출 경로를 만든다. 기존 비식별화 연구는 원본 영상의 얼굴, 차량번호, 인물 영역을 마스킹·블러 처리하는 데 주로 집중되어 왔다. 하지만 영상이 비식별화되더라도 VLM이 생성한 캡션이 “30대 여성”, “검은색 후드”, “편의점 입구”, “고령층 남성”, “실신” 등의 의미 정보를 텍스트로 명시한다면, 해당 텍스트는 별도의 유출 채널로 작용할 수 있다.

본 연구는 CCTV 관제 VLM 캡션의 의미론적 개인정보 유출을 정량화하고, 접근 권한 기반 추상화 정책이 VLM 기반 CCTV 관제 유용성과 프라이버시 사이의 균형에 미치는 영향을 분석한다. 특히 UCA와 AI Hub는 사용 언어뿐만 아니라 어노테이션 목적, 문장 길이, 정보 밀도, 표현 세분화 정도가 상이하므로, 단순한 언어 간 비교보다는 UCA를 주 실험으로, AI Hub를 보조 사례로 활용한다.

본 연구의 기여는 다음과 같다. 첫째, CCTV 관제 캡션의 의미론적 개인정보 유출을 정량화하는 Semantic Leakage Score(SLS)를 제안한다. 둘째, CCTV 도메인에 특화된 8개 민감 의미 속성 분류체계를 정의한다. 셋째, 사용자 접근 권한에 대응하는 A0~A3 4단계 추상화 정책을 제시하고 각 단계가 SLS와 유용성에 미치는 영향을 분석한다. 넷째, UCA 주 실험을 통해 A2가 준식별자를 효과적으로 제거하면서도 관제 업무에 필요한 핵심 사건 의미를 유지하는 실용적 균형점임을 보인다.

논문의 구성은 다음과 같다. 2장에서는 Surveillance VALU, UCA 및 UCF-Crime 기반 이상행동 탐지, 영상정보 보호 제도, LLM/VLM 기반 개인정보 유출 관련 연구를 살펴본다. 3장에서는 SLS를 중심으로 CCTV 관제 캡션의 개인정보 노출 평가와 접근 권한 기반 추상화 방법을 제안한다. 4장에서는 UCA 캡션 실험과 AI Hub 보조 사례를 통해 추상화 레벨별 유출 감소와 유용성 변화를 분석한다. 마지막으로 5장에서는 주요 결과를 종합하고, 연구의 한계와 향후 연구 방향을 논의한다.

2. 관련 연구

Surveillance VALU는 감시 영상의 사건을 자연어로 설명하고 자연어 질의로 검색·시점 정렬하는 분야이다. UCA(UCF-Crime Annotation)는 문장 단위 어노테이션을 포함한 대표적 Surveillance VALU 데이터셋으로, 긴 영상 길이, 낮은 해상도, 복잡한 배경, 작은 행동 단서 등 일반 웹 영상과 다른 도메인 특성을 다룬다[3]. UCA의 기반이 되는 UCF-Crime 계열 연구는 실제 감시 영상에서의 이상행동 탐지 문제를 정립한 대표 연구이다[2]. 한편 국내 제도 측면에서 「고정형 영상정보처리기기 설치·운영 안내서」

는 영상정보 처리와 접근권한 관리를 강조하고 [4], 「개인정보 보호법」은 민감정보 처리 제한과 그 범위를 규정한다[5]. 그러나 이러한 연구와 제도는 VLM이 생성한 텍스트 캡션의 유출 채널을 별도 평가 대상으로 충분히 다루지 않는다.

LLM/VLM 개인정보 유출 연구는 입력 텍스트로부터 사용자 속성을 추론하는 속성 추론 공격, 멀티모달 모델의 임베딩·특징에서 의미 정보를 복원하는 공격, 그리고 LLM-as-a-Judge 기반 프라이버시 평가로 구분된다. Staab 등은 LLM이 텍스트 입력으로부터 위치·성별·소득 등의 개인 속성을 추론할 수 있음을 보였다[6]. M4I (Multi-modal Models Membership Inference)는 멀티모달 모델에 대한 멤버십 추론 가능성을 분석하였다[7]. CapRecover는 VLM 특징으로부터 라벨 또는 캡션 수준의 의미를 복원할 수 있음을 보였다[8]. 한편 Meisenbacher 등은 LLM 평가자가 텍스트 프라이버시 인식에서 인간 평가와 약 0.54~0.58 수준의 합의도(Krippendorff's α)에 그쳐 프롬프트 구성과 모델 선택에 민감함을 보였으며, 이는 운영 정책 기반의 명시적 속성 평가가 보완적으로 필요함을 시사한다[9]. 그러나 이들 연구는 주로 입력·임베딩 단계의 유출을 다루며, 제안 방법은 VLM 출력 캡션 자체의 의미론적 개인정보 유출에 초점을 둔다.

차분 프라이버시(Differential Privacy, DP)는 정형 데이터에서 프라이버시 보장을 수학적으로 정의하는 대표적 체계를 제공한다[10]. k-anonymity는 준식별자 결합으로 인한 재식별 위험을 완화하기 위한 고전적 익명화 모델이다[11]. 그러나 두 접근 모두 자유 형식 자연어 캡션에 직접 적용하기는 어렵다. 캡션은 속성 경계가 불명확하고, “고령층 사람”처럼 하나의 표현이 연령과 취약성을 동시에 암시할 수 있다. 이러한 특성을 고려하여 속성별 노출과 가중치를 결합한 SLS를 정의하고, 관제 핵심 의미는 보존하되 준식별자는 제거·일반화하는 정책적 추상화를 제안한다.

3. SLS 기반 캡션 추상화

CCTV 영상 클립을 v, VLM이 생성한 자연어 캡션을 c라 할 때, 두 요소의 관계는 식 (1)과 같이 정의한다. 본 연구의 목표는 캡션 c로부터 영상 속 인물의 민감 또는 재식별 가능 속성이 노출되는 정도를 정량화하고, 접근 권한 기반 추상화로 이를 완화하는 것이다. 공격자는 원본 영상이나 VLM 내부 특징에는 접근할 수 없고 캡션만 관찰하는 블랙박스 텍스트 공격자로 가정한다.

$$c = VLM(v) \tag{1}$$

CCTV 관제 캡션에서 탐지 가능한 민감 의미 속성을 8개로 정의한다(표 1). S5와 S6은 건강·신체 상태와 연결되므로 높은 가중치를 부여하며, S7은 법령 상 민감정보 그 자체라기보다는 사건·범죄·위험 상황을 나타내는 운영상 민감 속성으로 간주한다. 속성별 가중치 서열은 임의로 설정된 것이 아니라 정보 민감도 실증 연구와 준식별자 재식별 위험 연구에 근거한다. Schomakers 등은 건강·위치 정보가 높은 민감도로, 복장 등 단일 시각 속성은 낮게 인식됨을 보였으며[12], 이는 S5·S6에 높은 가중치를 부여한 근거가 된다. 또한 El Emam 등은 연령·장소·성별이 준식별자로서 동치류 크기를 통해 재식별 위험을 결정함을 보였고[13], 이는 S2·S4(가중치 3)가 S1(가중치 2)보다 높은 재식별력을 갖는다는 설정을 뒷받침한다. 본 가중치는 예시적 설정이며 운영 환경에 따라 재조정될 수 있고, 그 강건성은 그림 4의 민감도 분석으로 확인하였다.

이 분류체계를 기반으로 캡션 c의 의미론적

표 1. 민감 의미 속성 분류체계, 가중치 및 식별 키워드 예시
Table 1. Sensitive semantic attributes, weights, and example keywords

코드	속성	가중치	성격	식별 키워드(예시)
S1	성별(Gender)	2	준식별자	남성/여성, man/woman, 30대 여성
S2	연령(Age)	3	강한 준식별자	고령층, 청소년, 20대, elderly, child
S3	복장(Clothing)	2	시각 기반 준식별자	검은색 후드, 흰색 셔츠, red jacket
S4	장소(Location)	3	강한 준식별자	편의점 입구, 북카페 내부, 3층 복도
S5	취약성 (Vulnerability)	4	민감/취약성 단서	임산부, 장애인, 거동 불편, 보행 보조기
S6	건강·안전 상태 (Health/Safety)	4	관제 핵심 민감정보	실신, 쓰러짐, 출혈, 부상 (주체 인식 규칙)
S7	이상행동 (Abnormal Behavior)	3	운영상 민감 속성	폭행, 절도, 파손, 방화, fighting
S8	시간·반복성 (Time/Repetition)	1	결합 식별자	오후 1시경, 매일, 반복적으로, 2회

개인정보 유출 정도는 식 (2)와 같이 정의한다.

$$SLS(c) = \frac{1}{W_{sum}} \sum_{i=1}^{8} w_i I_i(c) \qquad (2)$$

여기서 $I_i(c)$는 캡션 c에 i번째 속성이 포함되면 1, 그렇지 않으면 0을 갖는 지시함수이며, w_i는 표 1의 각 속성의 가중치이다. 분모 W_{sum}는 가중치 총합 22이다. SLS는 [0, 1] 값을 가지며, 1에 가까울수록 유출 위험이 크다. UCA의 A0 평균 SLS 0.118은 평균적으로 약 2.6/22 분량의 의미 단서가 캡션당 노출됨을 의미한다.

관제 시스템 사용자는 업무 목적과 책임 범위

표 2. 접근 권한 등급, 추상화 레벨 및 속성 보존/제거 매핑
Table 2. Access tiers, abstraction levels, and attribute mapping

권한 등급	사용자 예시	접근 레벨	운영 시나리오	보존 / 제거 속성(신규)
Tier 0	수사기관, 법원 영장 기반 포렌식 분석자	A0	정식 수사·증거 분석	보존: S1~S8 전체 / 제거: 없음
Tier 1	관제 책임자, 보안 관리자	A1	사건 검토, 내부 보고	보존: S2~S8 / 제거: S1(성별)
Tier 2	일반 관제사, 실시간 운영 인력	A2	실시간 모니터링, 1차 대응	보존: S5·S6·S7 / 제거: S1·S2·S3·S4·S8
Tier 3	통계 담당자, 외부 보고·연구 사용자	A3	통계 보고, 공개 자료 작성	보존: 사건 카테고리·위험도만 / 제거: S1~S8 전체

에 따라 차등 정보 접근 권한을 가져야 한다. 표 2와 같이 4개 접근 권한 등급과 캡션 추상화 수준을 매핑한다. A1은 “명시적 인물 속성 완화” 단계로, UCA 캡션에서 빈번히 등장하는 성별 명사를 우선 완화한다. A2는 성별 외에도 연령·복장·세부 장소·시간을 일반화·제거하되, 관제 대응에 필요한 건강·안전 상태와 이상행동은 유지한다. A3는 인물 속성을 모두 제거하고 사건 카테고리 또는 위험도 수준만 남기는 고수준 요약을 생성한다.

속성 검출은 한국어·영어 정규식 사전 및 표제어 사전을 결합한 규칙 기반 탐지기로 수행한다. 특히 S6(건강·안전)은 “사람이 넘어졌다”와 “의자를 넘어뜨렸다”를 구분하는 주체 인식 규칙을 적용한다. 한편, 추상화 정책에 따른 유용성(Utility) 평가는 캡션 단위의 12개 클래스 이상행동 분류 정확도(Accuracy)로 측정한다. 평가의 목적은 특정 고성능 분류 모델의 절대 성능을 달성하는 것이 아니라, 추상화 레벨 변경에 따른 유용성의 '상대적 변화 폭'을 관찰하는 데 있다. 이에 따라 재현성이 높은 키워드 매칭 기반 약한 기준선(Weak Baseline)을 평가 지표로 채택하였다. 즉 본문의 “유용성(Utility)”은 평가 척도의 개념적 명칭이고 표의 “정확도(Accuracy)”는 그 측정값으로 동일 개념을 가리킨다. 약한 기준선의 분류 절차는 재현 가능하도록 다음과 같이 정의한다. 각 캡션 c에 대해 클래스 k의 사전 키워드 집합 K_k와의 매칭 점수를 수식 (3)으로 산출하고, 예측 라벨은 $\arg max_k\, score(c,k)$로 결정한다. 동점이거나 모든 점수가 0인 경우 “미분류”로 처리하여 오분류로 집계한다.

$$score(c,k) = \sum_{\omega \in K_K} \mathbb{1}\ [\omega \in c],$$

$$\mathbb{1}\ [\omega \in c] = \begin{cases} 1 & \text{if } \omega \text{ appears in } c \\ 0 & otherwise \end{cases} \qquad (3)$$

표 3. UCA 주 실험결과(300개)
Table 3. Primary experimental results on UCA

레벨	평균 SLS	A0 대비 SLS 감소율(%)	분류정확도(%)	A0 대비 변화(%p)
A0	0.118	기준	21.0	기준
A1	0.063	46.6	23.0	+2.0
A2	0.013	89.0	23.0	+2.0
A3	0.000	100.0	8.3	-12.7

프라이버시-유용성 균형 점수는 식 (4)와 같이 정의한다.

$$Balance(L) = U(L) - \lambda SLS(L) \qquad (4)$$

여기서 $U(L)$은 추상화 레벨 L에서의 분류 정확도이며, λ는 프라이버시 항의 상대적 중요도를 조절하는 정책 계수이다. 기본 분석에서 $\lambda = 1$을 사용한다.

4. 실험 결과 및 고찰

실험에서는 UCA를 주 실험 데이터셋으로 사용한다. 12개 이상행동 카테고리에서 카테고리당 25개씩 총 300개 영문 캡션을 층화 추출법으로 추출하였다. 각 캡션은 A0 원문에서 출발하여 규칙 기반 추상화기를 통해 A1, A2, A3로 변환되었으며, 이에 따라 총 1,200개 캡션이 UCA 주 실험에 사용된다. AI Hub 실내 이상행동 한국어 캡션 100개는 국내 CCTV 어노테이션에서의 고밀도 노출과 규칙 기반 추상화의 한계를 사례 중심으로 확인하기 위한 보조 분석 자료로 활용한다.

표 3에서 A1은 성별 단서 완화만으로 SLS를 0.118에서 0.063으로 46.6% 감소시켰고, A2는 연령·복장·세부 장소·시간의 일반화로 SLS를 0.013

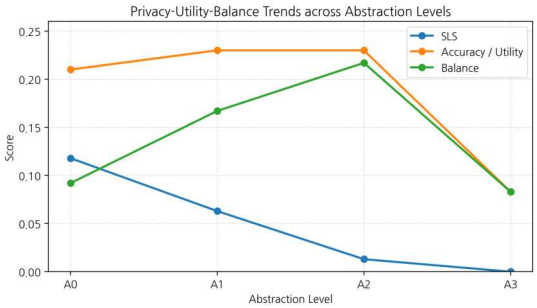
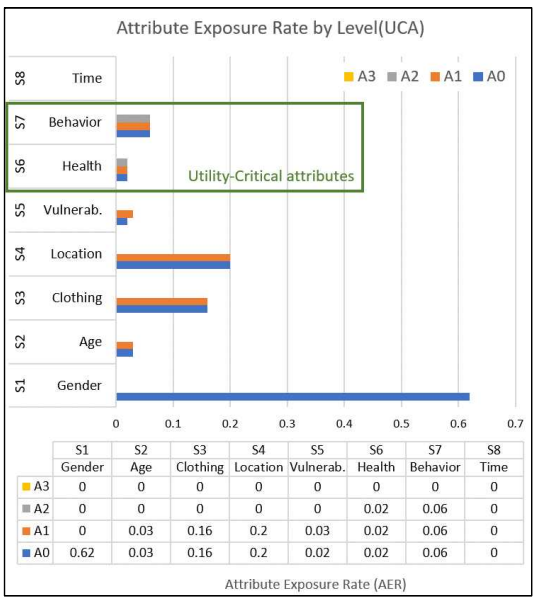


그림 1. UCA 데이터셋에서 추상화 레벨별 SLS, Accuracy, Balance 복합 추이
Fig. 1. Joint trends of SLS, Accuracy, and Balance across abstraction levels on UCA

까지 감소시켰다. A1과 A2의 정확도는 A0보다 0.02 높게 나타났으나, 본 연구에서는 이를 추상화에 따른 분류 성능 향상으로 해석하지 않는다. 그 근거는 다음과 같다. 첫째, 2.0%p의 차이는 300개 캡션 기준 약 6개 캡션의 분류 결과 변동



	S1 Gender	S2 Age	S3 Clothing	S4 Location	S5 Vulnerab.	S6 Health	S7 Behavior	S8 Time
A3	0	0	0	0	0	0	0	0
A2	0	0	0	0	0	0.02	0.06	0
A1	0	0.03	0.16	0.2	0.03	0.02	0.06	0
A0	0.62	0.03	0.16	0.2	0.02	0.02	0.06	0

그림 2. UCA 데이터셋에서 추상화 레벨별 속성 노출률
Fig. 2. Per-attribute exposure rate by abstraction level on UCA

표 4. UCA 기준 프라이버시-유용성 균형 (λ=1)
Table 4. Privacy-utility balance on UCA

레벨	SLS	Accuracy	Balance
A0	0.118	0.21	0.092
A1	0.063	0.23	0.167
A2	0.013	0.23	0.217
A3	0.000	0.083	0.083

에 해당하는 미세한 값으로, 표본 규모를 고려할 때 우연적 변동의 범위를 넘어선다고 보기 어렵다. 둘째, 본 연구의 정확도는 특정 모델의 절대 성능이 아니라 키워드 매칭 기반 약한 기준선에서 추상화 레벨 간 유용성의 '상대적 변화 폭'을 관찰하기 위한 비교 지표이므로, 이러한 소폭 증가를 성능 향상으로 해석하는 것은 평가 설계의 취지와 부합하지 않는다. 셋째, A2의 의의는 정확도 상승이 아니라 SLS를 89.0% 감소시키면서도 유용성을 A0 수준으로 유지하여 추가적 훼손이 없었다는 점에 있다. 반면 A3의 정확도 8.3%는 12개 클래스 무작위 추측 수준과 사실상 동일하다.

그림 1과 표 4는 A2가 가장 높은 균형 점수를 보임을 나타낸다. 복합 그래프에서 SLS는 A0→A3에서 단조 감소하고, Accuracy는 A2까지 유지된 뒤 A3에서 급락하며, Balance는 A2에서 정점에 도달하는 연동 추이가 드러난다. 한편 Balance 지표는 식 (4)을 통해 SLS, 즉 민감 의미 속성과 그 가중치에 종속되므로 균형 지표의 절대값은 가중치 설정에 영향을 받을 수 있다. 본 연구는 이 종속성을 그림 4의 가중치 민감도 분석으로 보완하여 결론의 강건성을 확인하였다.

그러나 균형 점수 단일 지표가 아니라 SLS 감소율, 키워드 매칭 기반 유용성 기준선 유지, A3의 무작위 수준 하락, 접근 권한 기반 관제 운영 적합성을 종합하여 A2가 실용적 균형점임을 확인하였다.

그림 2에서 A0의 주요 노출 속성은 S1 성별,

그림 3. UCA와 AI Hub의 어노테이션 방식 민감도 사례

Fig. 3. Annotation protocol sensitivity: UCA vs. AI Hub case examples

S3 복장, S4 장소이다. A1에서는 S1이 급감하고, A2에서는 S3와 S4가 추가로 감소한다. S6·S7은 UCA에서 원래 노출률이 높지 않으므로, 본 결과

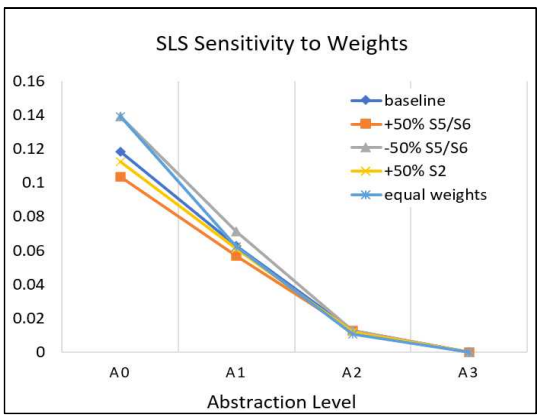


그림 4. 속성 가중치 변화에 따른 SLS 민감도 분석

Fig. 4. Sensitivity of SLS to attribute weight variations

는 “A2가 유용성 핵심 속성을 강하게 보존한다”는 해석보다는 “A2가 주로 노출되던 성별·복장·장소 단서를 제거하면서 잔여 안전·이상행동 단서를 추가로 훼손하지 않는다”는 해석에 더 부합한다.

그림 3은 UCA와 AI Hub의 절대 SLS를 직접 비교하기 위한 것이 아니라, 어노테이션 방식과 정보 밀도 차이를 설명하기 위한 사례 중심 그림이다. AI Hub 캡션은 “오후 1시경”, “북카페 내부”, “흰색 셔츠와 검정색 바지”, “30대 여성”, “파손 행위”처럼 시간·장소·복장·연령·행동 정보가 한 문장에 함께 포함될 수 있다. 이 경우 단순 PII 단어 제거만으로는 충분하지 않으며, 세부 장소·시간의 세분도 제거와 사람 상태/물체 행위 구분을 위한 주체 인식 규칙이 함께 요구된다.

그림 4는 기준선, S5/S6 가중치 ±50%, S2 가중치 +50%, 균등 가중치의 5개 시나리오에서 A0 > A1 > A2 > A3의 SLS 순서가 안정적으로 유지됨을 보여준다. 이는 제안 체계의 결론이 특정 가중치 조합에 과도하게 의존하지 않음을 의미한다.

5. 결론 및 향후 연구

본 연구는 VLM 기반 CCTV 관제 시스템에서 원본 영상이 비식별화되더라도 VLM이 생성하는 자연어 캡션이 별도의 텍스트 기반 개인정보 유출 채널이 될 수 있음을 문제로 제기하였다. 이를 정량화하기 위해 SLS, 8개 민감 의미 속성 분류체계, 사용자 접근 권한 기반 A0~A3 추상화 정책을 정의하고, UCA 주 실험과 AI Hub 보조 사례를 통해 추상화 수준에 따른 프라이버시-유용성 변화를 분석하였다.

실험 결과, VLM 캡션은 성별·연령·복장·장소·건강·이상행동·시간 정보를 텍스트로 명시함으로써 원본 영상 없이도 사적 속성이나 재식별 단서

를 제공할 수 있음을 확인하였다. A1은 성별 단서 완화만으로 SLS를 46.6% 감소시켰고, A2는 준식별자 일반화로 SLS를 89.0% 감소시키면서도 키워드 매칭 기반 약한 기준선에서 분류 정확도를 A0와 유사한 수준으로 유지하였다. 반면 A3는 SLS를 0으로 낮추지만 정확도가 12개 클래스 무작위 추측 수준으로 하락하여 세부 관제 임무에는 부적합하였다. 이에 따라 A2가 실시간 관제 환경에서 유용성과 프라이버시를 함께 고려할 수 있는 실용적 균형점임을 확인하였다.

다만 UCA와 AI Hub의 결과는 언어 간 우열이나 절대 SLS의 단순 비교로 해석되어서는 안 된다. 두 데이터셋은 어노테이션 목적, 라벨 공간, 문장 길이, 정보 밀도, 표현 세분화 정도가 다르다. AI Hub 보조 분석에서 잔여 노출이 나타난 것은 국내 CCTV 어노테이션이 시간·장소·복장·연령·행동 정보를 한 문장에 밀집해 포함할 수 있으며, 규칙 기반 추상화가 이러한 정보 밀도와 표현 세분화 정도에 영향 받음을 의미한다.

본 연구는 다음과 같은 한계를 가진다. 유용성 평가는 키워드 매칭 기반 캡션 단위 분류에 한정되며, 21~23%의 정확도는 강한 모델의 성능 상한이 아니라 추상화 레벨 간 상대 비교를 위한 약한 기준선의 측정값이다. 또한 속성 탐지는 규칙 기반 사전과 정규식에 의존하므로 문맥적 은유, 복합 속성, 사람 상태와 물체 행위의 경계, 세부 장소·시간 표현의 세분도 차이를 충분히 반영하지 못할 수 있다.

향후 연구에서는 매크로 F1, 이진 분류, CLIP/FAISS 기반 검색, LoRA 미세조정 CLIP, R@1·R@5·mIoU 등 실제 관제 시스템에 가까운 유용성 지표를 추가하고, 주체 인식 기반 속성 검출기와 세분도 인식 SLS를 도입할 필요가 있다. 또한 LLM 보조 추상화, 인간 검증, 정책 기반 접근통제, 실제 관제 시스템과 연계한 종단간 평가를 결합함으로써 VLM 기반 CCTV 관제에서 캡션 수준 프라이버시 보호 체계의 실효성을 검증할 계획이다.

본 과제(결과물)는 2026년도 교육부 및 전라남도의 재원으로 전라남도RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2026-RISE-14-001)

참 고 문 헌

[1] U. D. Silva, L. Fernando, B. L. P. Lik, Z. Koh, S. C. Joyce, B. Yuen and C. Yuen, “Large Language Models for Video Surveillance Applications,” arXiv Preprint arXiv:2501.02850, 2025, DOI: 10.48550/arXiv.2501.02850

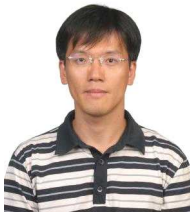
[2] W. Sultani, C. Chen, and M. Shah, “Real-world Anomaly Detection in Surveillance Videos,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479-6488, 2018, DOI: 10.1109/CVPR.2018.00678

[3] T. Yuan, X. Zhang, K. Liu, B. Liu, C. Chen, J. Jin, and Z. Jiao, “Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22052-22061, 2024, URL: https://openaccess.thecvf.com/content/CVPR2024/html/Yuan_Towards_Surveillance_Video-and-Language_Understanding_New_Dataset_Baselines_and_Challenges_CVPR_2024_paper.html

[4] 개인정보보호위원회, “고정형 영상정보처리기기 설치·운영 안내서,” 2024.12, URL: https://www.privacy.go.kr/front/bbs/bbsView.do?bbsNo=BBSMSTR_000000000049&bbscttNo=20779

[5] 국가법령정보센터, 「개인정보 보호법」 제23조 및 「개인정보 보호법 시행령」 제18조, URL: https://www.law.go.kr
[6] R. Staab, M. Vero, M. Balunovic, and M. Vechev, “Beyond Memorization: Violating Privacy via Inference with Large Language Models,” International Conference on Learning Representations (ICLR), 2024, https://openreview.net/forum?id=kmn0BhQk7p
[7] P. Hu, Z. Wang, R. Sun, H. Wang, and M. Xue, “M4I: Multi-modal Models Membership Inference,” Advances in Neural Information Processing Systems (NeurIPS), 2022, https://proceedings.neurips.cc/paper_files/paper/2022/file/0c79d6ed1788653643a1ac67b6ea32a7-Paper-Conference.pdf
[8] K. Xiu and S. Zhang, “CapRecover: A Cross-Modality Feature Inversion Attack Framework on Vision Language Models,“ arXiv Preprint, arXiv:2507.22828, 2025, URL: https://arxiv.org/abs/2507.22828
[9] S. Meisenbacher, A. Klymenko, and F. Matthes, "LLM-as-a-Judge for Privacy Evaluation? Exploring the Alignment of Human and LLM Perceptions of Privacy in Textual Data," Proc. 2025 Workshop on Human-Centered AI Privacy and Security (HAIPS), pp. 126-138, 2025, DOI: 10.1145/3733816.3760760
[10] C. Dwork, “Differential Privacy,” Proc. International Colloquium on Automata, Languages, and Programming (ICALP), pp. 1-12, 2006, DOI: 10.1007/11787006_1
[11] L. Sweeney, “k-anonymity: A Model for Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002, DOI: 10.1142/S0218488502001648
[12] E.-M. Schomakers, C. Lidynia, D. Müllmann, R. Matzutt, K. Wehrle, I. Spiecker genannt Döhmann, and M. Ziefle, “Putting Privacy into Perspective - Comparing Technical, Legal, and Users' View of Information Sensitivity,” INFORMATIK 2020, Gesellschaft für Informatik, pp. 857-870, 2021, DOI: 10.18420/inf2020_76
[13] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, “Estimating the Re-identification Risk of Clinical Data Sets,” BMC Medical Informatics and Decision Making, vol. 12, no. 66, 2012, DOI: 10.1186/1472-6947-12-66

저 자 소 개



한승완(Seungwan Han)

1994.2 전남대학교 전산학과 졸업
1996.2 전남대학교 전산통계학과 석사
2001.8 전남대학교 전산통계학과 석사
2001.12-2021.3 : ETRI 책임연구원
2021.3-현재 : 국립목포대학교 컴퓨터학부 부교수
<주관심분야> 알고리즘, 정보보호, 딥러닝, 데이터 사이언스, 컴퓨터비전 등

강승호(Seung-Ho Kang)

1994.8 전남대학교 전산학과 졸업
2003.2 전남대학교 전산학과 석사
2009.8 전남대학교 전산학과 박사
2010.9-2013.8 :국가수리과학연구소 연구원
2013.9-2025.2 : 동신대학교 컴퓨터학과 부교수
2025.3-현재: 국립목포대학교 컴퓨터학부 부교수
<주관심분야> 알고리즘, 강화학습 등