

Anomaly LVLM: LVLM을 활용한 사용자 맞춤형 비디오 이상 탐지 연구

Sunghyun Ahn*, Youngwan Jo, Kijung Lee, Sein Kwon, Sanghyun Park†

Department of Computer Science, Yonsei University

Korea Software Congress 2024

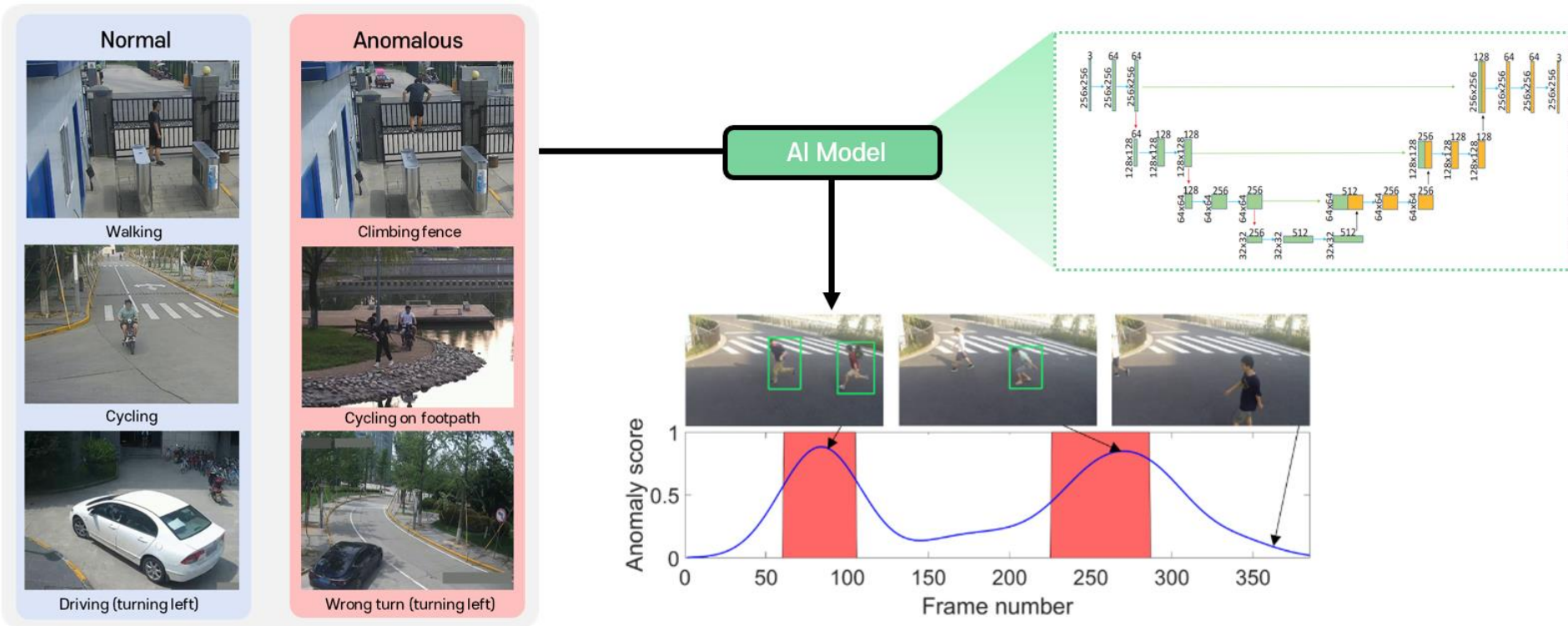
{skd, jyy1551, rlwjd4177, seinkwon97, sanghyun}@yonsei.ac.kr



01 Background Video Anomaly Detection



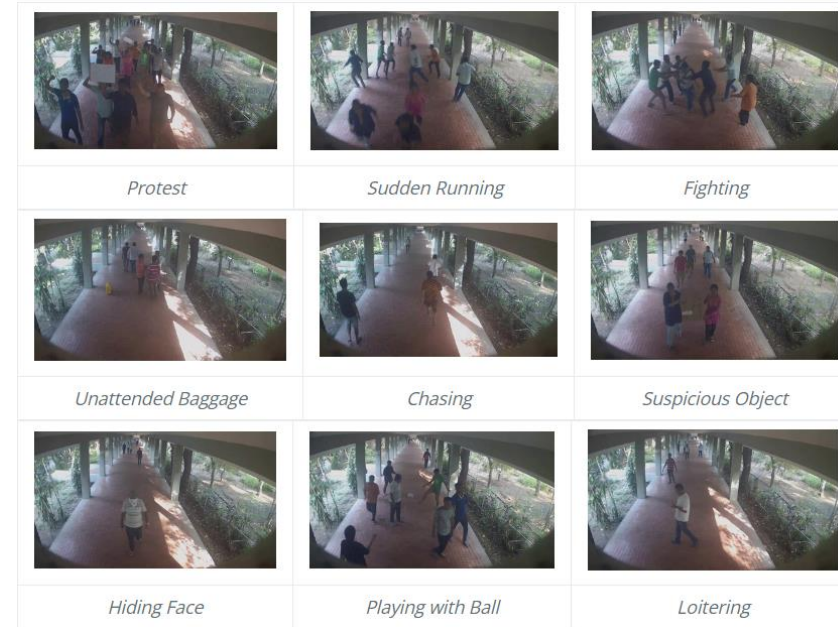
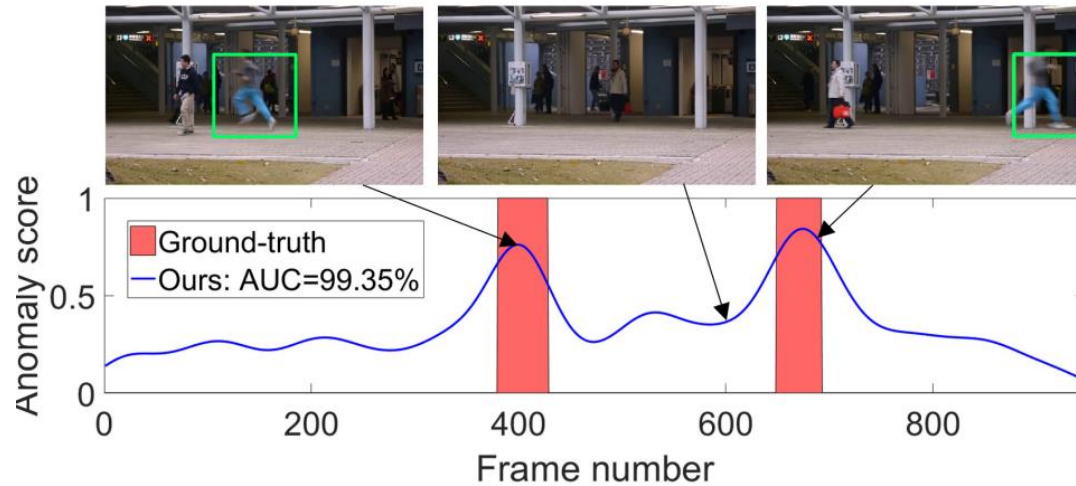
- 비디오 이상 탐지(VAD)는 비디오 스트림 내에서 비정상적인 사건이 발생했는지 여부를 판단하는 기술
- 비정상적인 사건에는 상황에 적합하지 않은 객체의 행동(i.e. climbing)이나 외형(i.e. bicycle) 정보가 포함됨
- 목표는 프레임별로 이진 분류를 수행하는 것



01 Background Main Problem



- 클래스 불균형 문제 $|\{x_i|y_i = 0\}| \gg |\{x_i|y_i = 1\}|$
- 다양한 비정상성 문제
- 정상 데이터만을 학습하고, 정상 데이터의 패턴과 유사하지 않은 모든 것을 비정상으로 분류하는 방법인 단항 분류(One Class Classification, OCC)가 주로 활용됨



Rodrigues, Royston, et al. "Multi-timescale trajectory prediction for abnormal human activity detection." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020.

01 Background One Class Classification



- 분류 기반 방법 (Classification-based): 정상적인 특징 벡터들을 특정 범위 안으로 군집화하는 클러스터 학습

$$\mathcal{A}_\theta(x) = \text{Diff}(\phi_\theta(x), c)$$

- Key idea
- 정상 데이터: 학습된 클러스터 안에 속함
 - 비정상 데이터: 학습된 클러스터로부터 멀리 위치함
- ϕ_θ : kernel function
 c : center of hypersphere

- 거리 기반 방법 (Distance-based): 정상 데이터의 특징을 메모리에 저장하고, 최근접 이웃 탐색(NN)을 사용하여 분류를 수행함

$$\mathcal{A}_\theta(x) = \min_{x_{tr} \in X_{tr}} \text{Diff}(\phi_\theta(x), \phi_\theta(x_{tr}))$$

- Key idea
- 정상 데이터: 특징 공간에서 정상 데이터와 가까이에 위치함
 - 비정상 데이터: 특징 공간에서 정상 데이터와 멀리 위치함

- 재구축 기반 방법 (Reconstruction-based): 생성 모델을 활용하여 정상 데이터를 재구축하는 학습

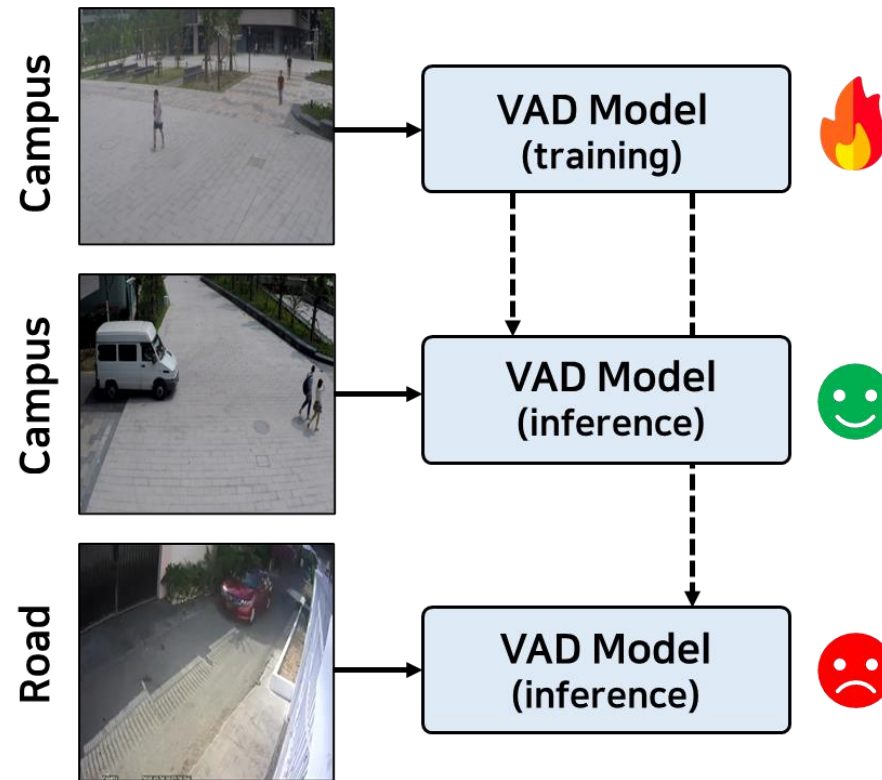
$$\mathcal{A}_\theta(x) = \text{Diff}(x, \text{Recon}_\theta(x))$$

- Key idea
- 정상 데이터: 재구축을 올바르게 수행함
 - 비정상 데이터: 재구축을 올바르게 수행하지 못 함

02 Introduction Problem Definition



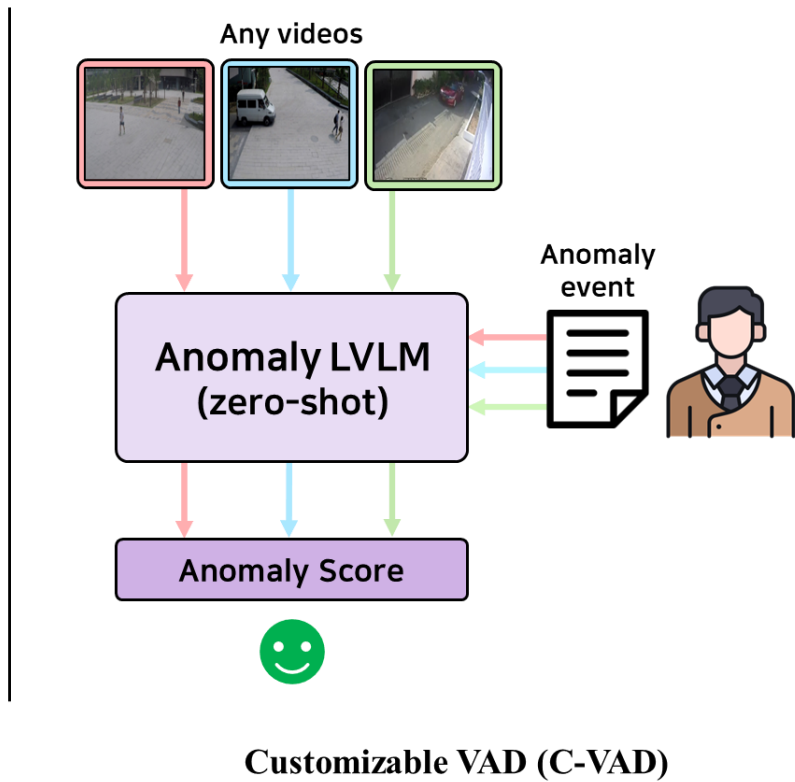
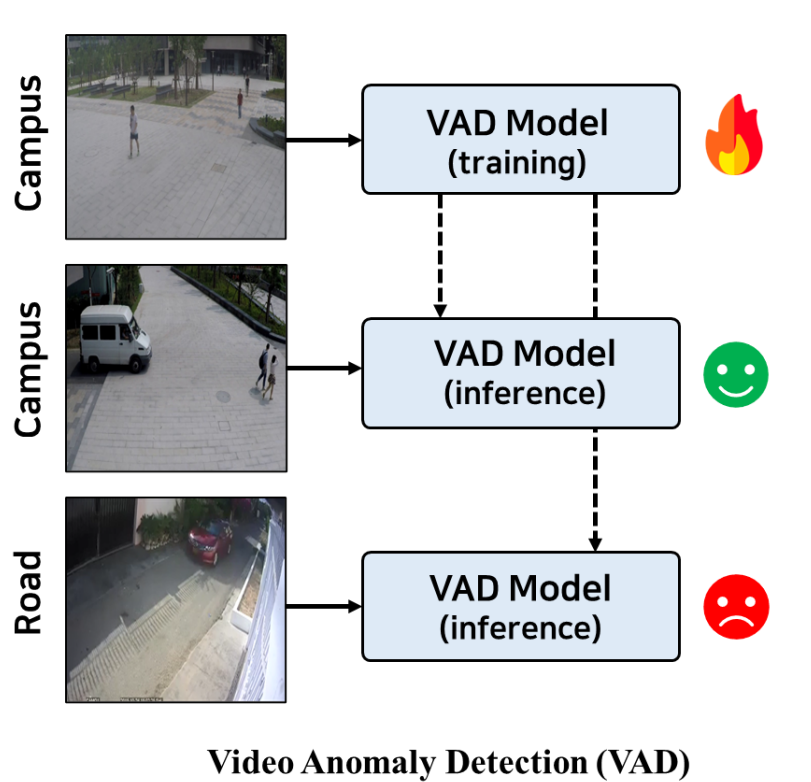
- VAD 모델은 특정 상황의 정상 패턴만 학습하므로 다양한 상황에 적용하기 어렵다는 한계가 존재함
- 보행자 구역에 대해 학습된 모델은 차량이 발견되면 이상 사건으로 판단하기 때문에, 차량이 통행하는 도로 상황에서 활용하기 힘들
- 사용자는 차도와 같은 다른 상황을 추가적으로 학습하거나, 별도의 AI 모델을 새롭게 만들어야 함 → 고성능 장비, 데이터 수집 등을 요구



02 Introduction Customizable VAD



- 사용자 맞춤형 비디오 이상 탐지 (Customizable VAD, C-VAD)
- 사용자가 정의한 텍스트를 비정상 이벤트로 간주하고, 비디오에서 해당 이벤트를 지닌 프레임을 탐지함
- Zero-shot 접근이 가능하면, 사용자는 각 상황에 맞는 모델을 따로 준비할 필요가 없고 학습을 할 필요도 없음



02 Introduction Large Vision Language Model



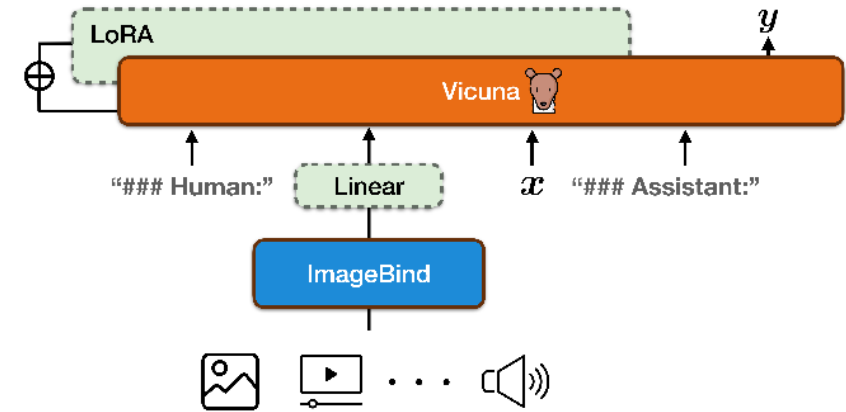
- LVLM: 이미지와 텍스트를 함께 이해하고 처리할 수 있는 인공지능 모델
- 이미지 처리를 위한 Image Encoder(i.e. ImageBind)와 텍스트를 처리를 위한 LLM (i.e. Vicuna)이 이용됨
- Image Captioning, Visual Question Answering, Text generation 등을 할 수 있음



Chatbot

Why this happened?

The image shows a washing machine overflowing with water, causing a flood in the kitchen. This happened because the washing machine's hose or drain pipe may have become clogged or broken, leading to water accumulation inside the machine. As a result, the water overflowed and flooded the kitchen, causing damage to the surrounding area. This situation highlights the importance of regular maintenance and inspection of appliances to prevent such incidents and ensure the safety of the home.

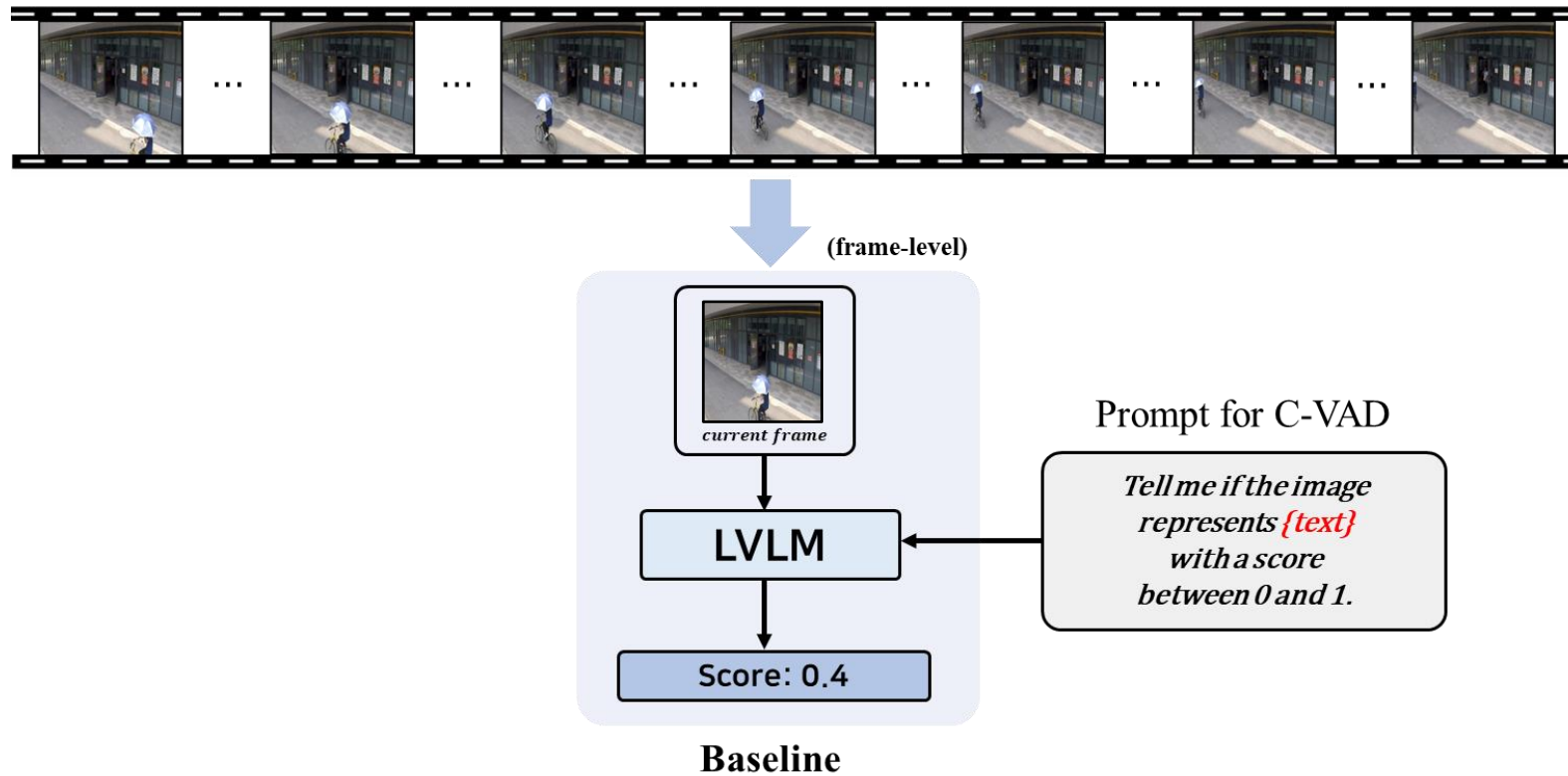


Su, Yixuan, et al. "Pandagpt: One model to instruction-follow them all." *arXiv preprint arXiv:2305.16355* (2023).

02 Introduction Basic Idea



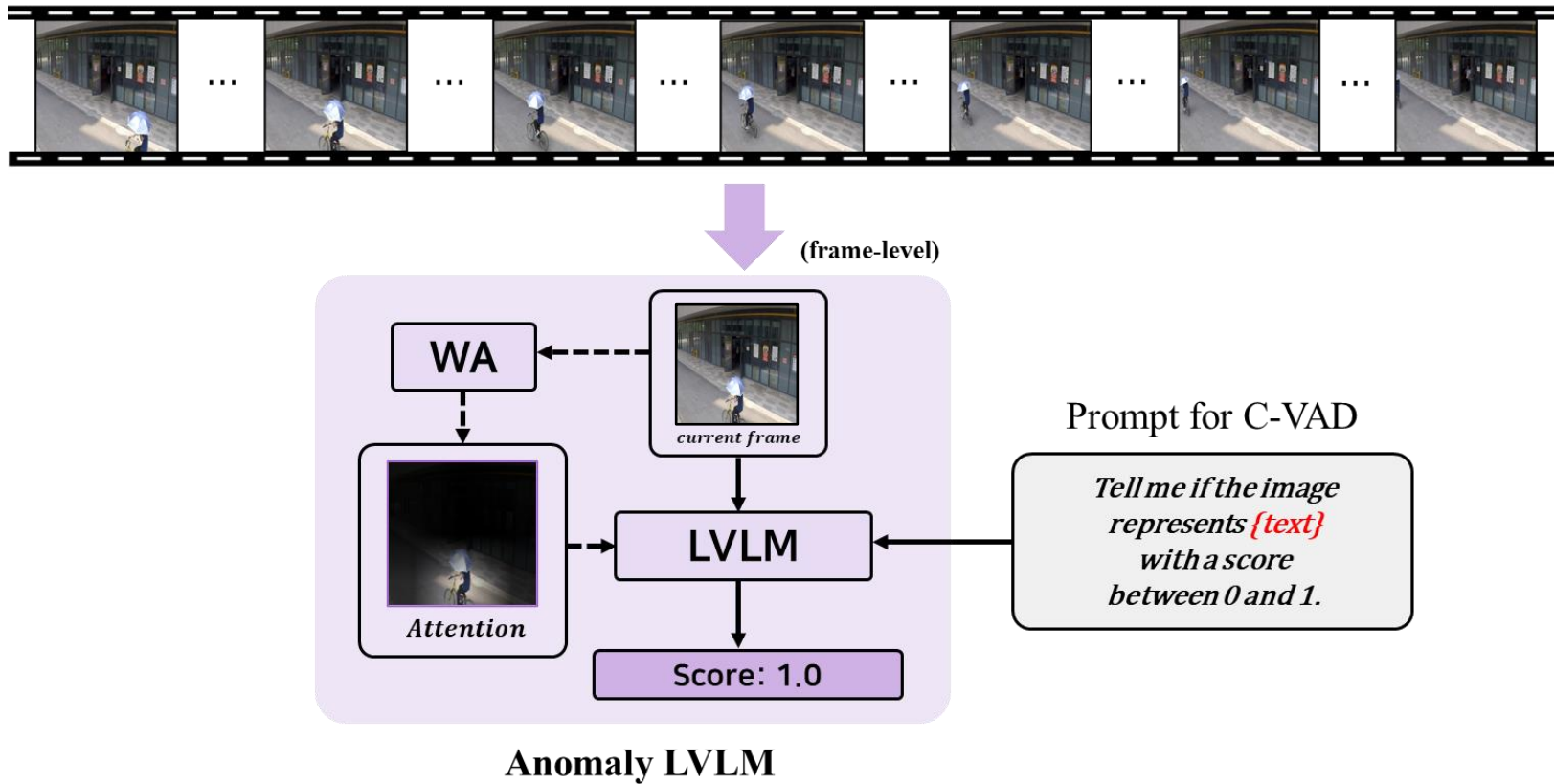
- LVLM을 통해 frame 단위로 VQA를 수행하여 Anomaly Score를 추정함
- 입력 이미지가 사용자가 입력한 텍스트(비정상 이벤트)를 표현하는 정도를 0(no)과 1(yes)사이로 반환하라고 함
- CCTV 영상은 전경보다 배경이 더 많이 포함되어 있어 작은 객체에 대한 올바른 분석이 어려울 수 있음



02 Introduction Anomaly LVLM



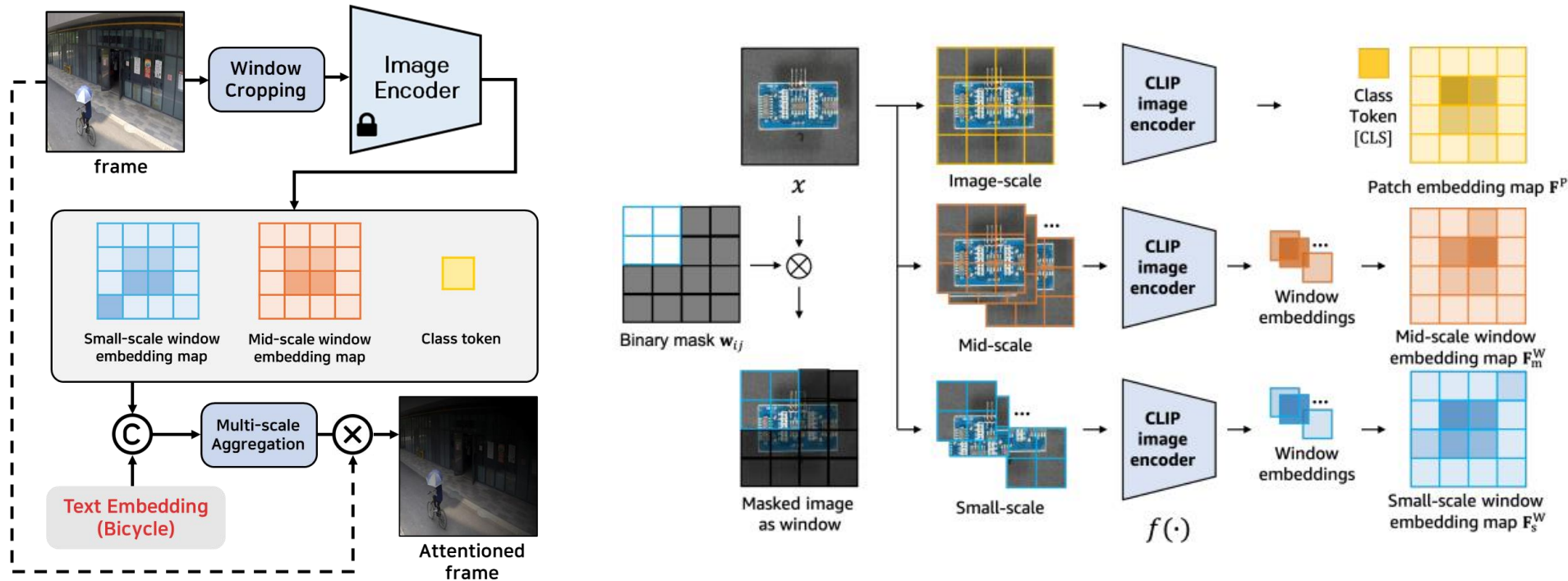
- Jeong이 제안한 WinCLIP 모델을 기반으로, 배경보다 객체가 강조된 영상을 제작하는 전처리 기법 (WA)을 적용함
- 제안 프레임워크(Anomaly LVLM)는 WA 모듈과 LVLM으로 구성되며, CCTV 비디오의 특성을 고려하여 C-VAD를 수행할 수 있음



03 Method WinCLIP based Attention



- WinCLIP에서 제안한 window embedding maps를 바탕으로 텍스트에 해당하는 영역을 파악함
- 다중 스케일 맵들은 각각 프레임의 작은 영역과 큰 영역에 대한 특징들을 나타내고, 클래스 토큰은 프레임의 대표 특징을 나타냄
- 입력 영상을 다시 곱하는 방식으로 텍스트 관련 부분에 집중하는 프레임을 생성할 수 있음



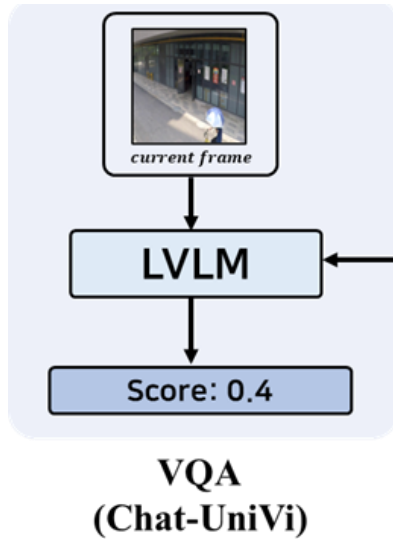
$$f_i^{text} = \Phi_{WA}(f_i, e^{text})$$

Jeong, J et al., "WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation", Proceedings of the IEEE conference on computer vision and pattern recognition, pp.19606-19616, 2023.

03 Method Prompt Tuning



- Task는 LVLM이 수행해야 하는 작업으로, 프레임 내 텍스트 포함 여부를 평가하는 명령어
- Consideration은 평가 시 고려 사항으로, 텍스트에 해당하는 요소가 프레임의 중심 내용이 아니더라도 높은 점수를 부여하라는 지침 (이미지가 {text}를 포함하더라도 이미지의 중심 내용이 아니면 낮은 점수를 반환하는 경향이 있으므로 이용함)
- Output은 평가 결과의 출력 형식을 지정하며, 이상 점수를 소수점 첫째 자리까지 표현하도록 명시함



Prompt for C-VAD

Tell me if the image represents {text} with a score between 0 and 1.

Detailed prompt

Task: Evaluate whether the given image includes {text} on a scale from 0 to 1. A score of 1 means {text} is clearly present in the image, while a score of 0 means {text} is not present at all. For intermediate cases, assign a value between 0 and 1 based on the degree to which {text} is visible.

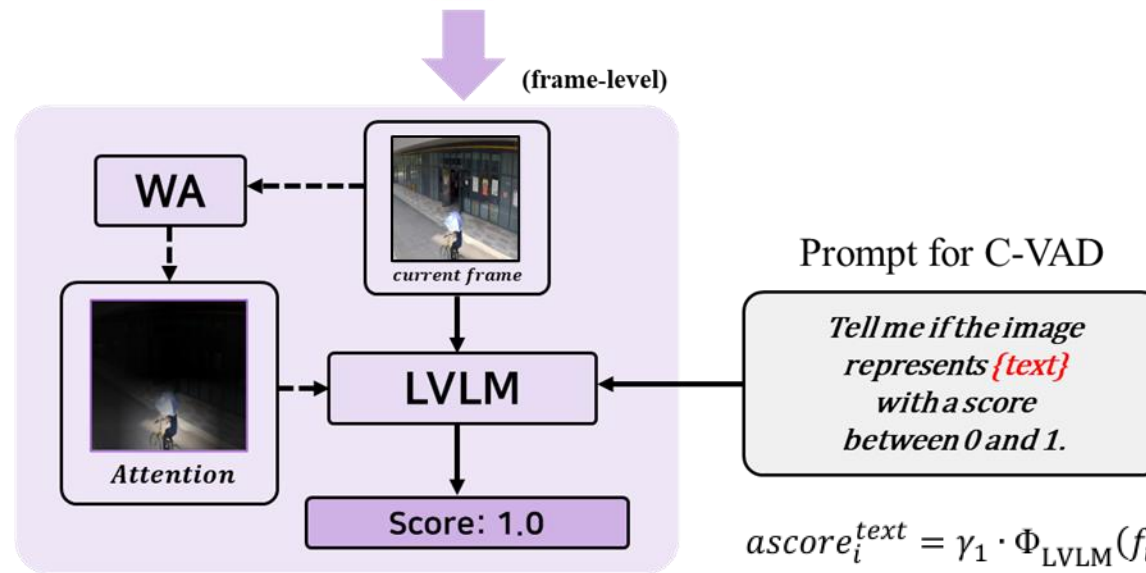
Consideration: The key is whether {text} is present in the image, not its focus. Thus, if {text} is present, even if it is not the main focus, assign a higher score like 1.0.

Output: Provide the score as a float, rounded to one decimal place.

03 Method Video Anomaly Detection



- 입력 프레임으로는 원본 프레임과 WA를 통해 전처리된 프레임이 각각 사용됨
- LVLM이 응답한 결과들은 조합되어 최종 이상 점수가 산출됨



$$ascore_i^{text} = \gamma_1 \cdot \Phi_{LVLM}(f_i, p^{text}) + \gamma_2 \cdot \Phi_{LVLM}(f_i^{text}, p^{text})$$

Anomaly LVLM

04 Experiments Dataset



- ShanghaiTech Campus (STC)
- 학교 캠퍼스 환경으로, 보행자 전용 도로를 나타내는 데이터셋
- 정상 이벤트: walking, standing 등
- 비정상 이벤트: riding a bike, jumping, running, car, skateboarding 등

→ 정답(frame-level label)이 범주형이 아니라 이진 형태로만 존재하기 때문에 해당 데이터셋을 가공하여 C-VAD 데이터셋을 제작함



04 Experiments Results





- VAD 성능을 평가하기 위해 Area Under the ROC Curve(AUC)를 이용함
- 제안된 프레임워크는 단일 LVLM(Baseline) 대비 평균 AUC가 4.74% 향상되었으며, 최종적으로 0.81의 준수한 성능을 달성함
- 이는 WA가 작은 객체의 중요 정보를 강조하여, LVLM이 보다 정확한 이상 점수를 반환하는 데 기여했기 때문



	Class	Baseline	Proposed	Improvement
Action	Skateboarding	0.6163	0.6489	+5.29%
	Throwing	0.9218	0.9302	+0.92%
	Running	0.5316	0.5880	+10.61%
	Loitering	0.6150	0.7169	+16.57%
	Jumping	0.8259	0.8822	+6.82%
	Falling	0.7727	0.8325	+7.74%
	Fighting	0.8449	0.8789	+4.03%
Appearance	Car	0.8846	0.8846	+0.00%
	Hand truck	0.9528	0.9548	+0.21%
	Bicycle	0.7219	0.7219	+0.00%
	Motorcycle	0.8759	0.8759	+0.00%
	Average	0.7784	0.8104	+4.74%

04 Experiments Qualitative Analysis



- WA가 작은 객체의 중요 정보를 강조하여 탐지 성능이 향상된 예시
- Running (0.5 -> 0.9), skateboarding (0.0 -> 0.8)

P	<p>Task: Evaluate whether the given image includes running on a scale from 0 to 1. A score of 1 means running is clearly present in the image, while a score of 0 means running is not present at all. For intermediate cases, assign a value between 0 and 1.</p> <p>Consideration: The key is whether running is present in the image, not its focus. Thus, if running is present, even if it is not the main focus, assign a higher score like 1.0.</p> <p>Output: Provide the score as a float, rounded to one decimal place.</p>	
R1		<p>The score for the image is 0.5. This indicates that there is a low level of activity related to running, but it is not clearly happening in the image.</p>
R2		<p>The image shows a group of people running down a path, with one person clearly in motion. Based on the visual content, the score for running would be 0.9.</p>

P	<p>Task: Evaluate whether the given image includes skateboarding on a scale from 0 to 1. A score of 1 means skateboarding is clearly present in the image, while a score of 0 means skateboarding is not present at all. For intermediate cases, assign a value between 0 and 1.</p> <p>Consideration: The key is whether skateboarding is present in the image, not its focus. Thus, if skateboarding is present, even if it is not the main focus, assign a higher score like 1.0.</p> <p>Output: Provide the score as a float, rounded to one decimal place.</p>	
R1		<p>The image does not show any skateboarding activity. The score is 0.0.</p>
R2		<p>The image includes a person riding a skateboard down a brick road, which suggests that skateboarding is present in the scene. I would assign a score of 0.8.</p>

04 Experiments Qualitative Analysis



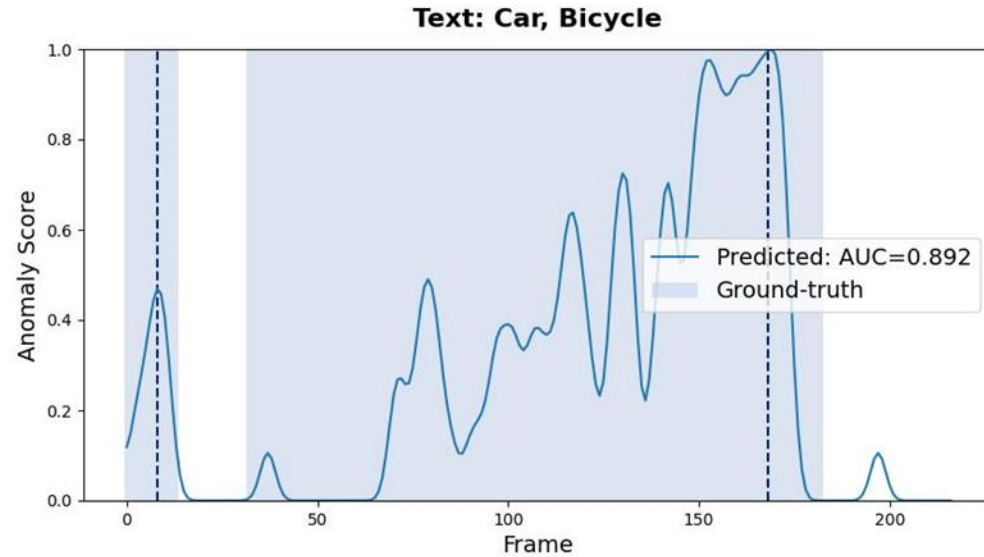
- (a)는 점프를 하는 한 명의 대상에 집중하여 더 높은 점수를 반환함
- (b)는 작은 손수레(Hand truck) 영역을 강조하여 더 정확한 탐지가 가능함
- (c)와 같이 객체가 충분히 큰 경우, WA를 수행하지 않고도 정확한 이상 탐지가 가능함

(a) Jumping		(b) Hand truck		(c) Car	
The image shows two people playing with a frisbee, and one of them is jumping, so the score is 0.7 .	The image shows a person doing a jump, so the score would be 1.0 .	The hand truck is partially visible, so the score would be 0.8 .	The image features a hand truck, and it is clearly visible. Assign a score of 1.0 .	The silver car is prominently parked on the brick-paved road, taking up much of the image. The score is 1.0 .	The image shows a car driving down a street, so the score would be 1.0 , as the car is clearly present in the image.

04 Experiments Qualitative Analysis



- 다중 클래스에 대해 비디오의 이상 점수를 시각화함
- Anomaly LVLM이 다양한 비정상 유형에 대해서도 안정적인 이상 탐지 성능을 발휘함



05 Conclusion Contributions



- 다양한 비정상 상황에서 효과적으로 이상 탐지를 수행하기 위해, 사용자 맞춤형 비디오 이상 탐지(C-VAD) 기술과 Anomaly LVLM 프레임워크를 새롭게 제안함
- CCTV 비디오에서 작은 객체를 효과적으로 감지할 수 있는 WinCLIP 기반 어텐션 방법을 제안함
- C-VAD 작업의 정량적 평가를 위해 C-STC 데이터셋을 구축하고, 이를 통해 Anomaly LVLM의 우수성을 입증함
- 향후 연구에서는 행동 이상 유형을 보다 정교하게 판단하기 위해 시간 정보를 활용한 추가적인 연구를 진행할 계획

Thank you