WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

# AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM

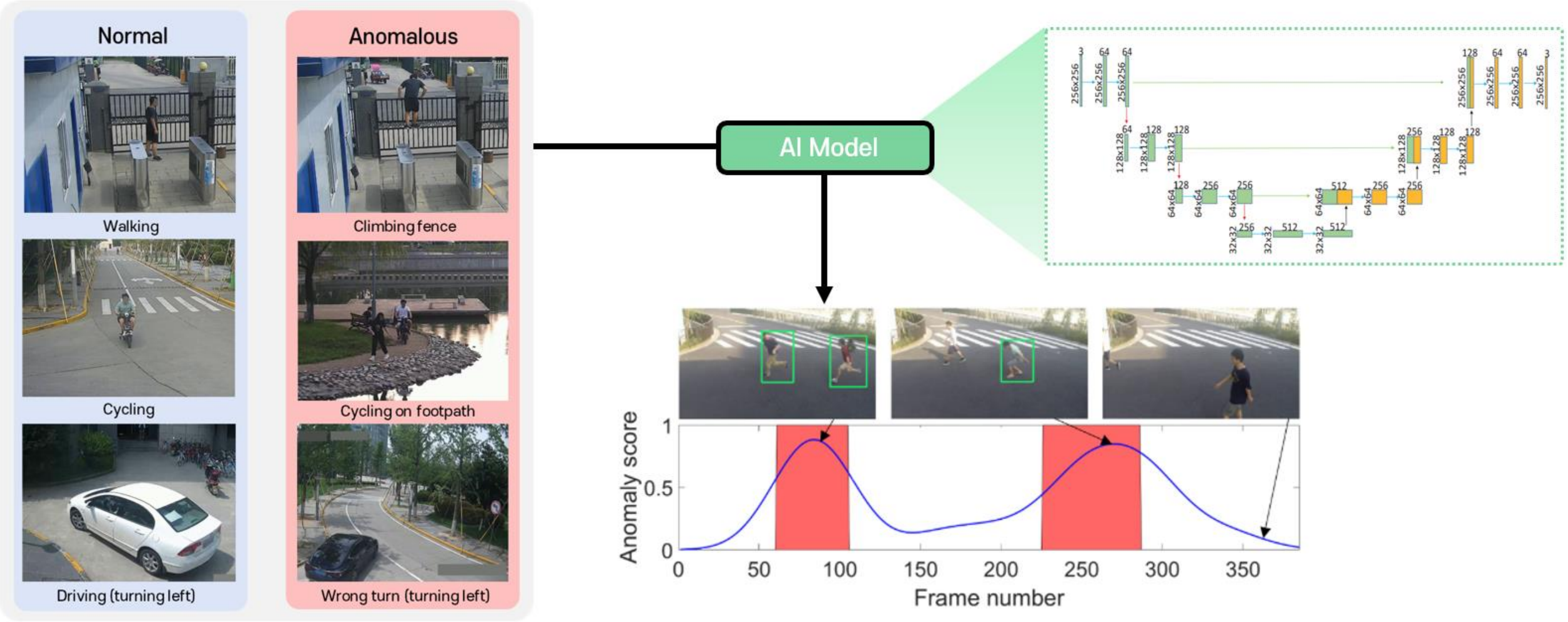
Sunghyun Ahn⋆, Youngwan Jo⋆, Kijung Lee, Sein Kwon, Inpyo Hong, Sanghyun Park†

Data Engineering Lab, Yonsei University
Seoul, Republic of Korea
{skd, jyy1551, rlwjd4177, seinkwon97, hip9863, sanghyun}@yonsei.ac.kr

⋆ Equal contribution
† Corresponding author

# Video Anomaly Detection

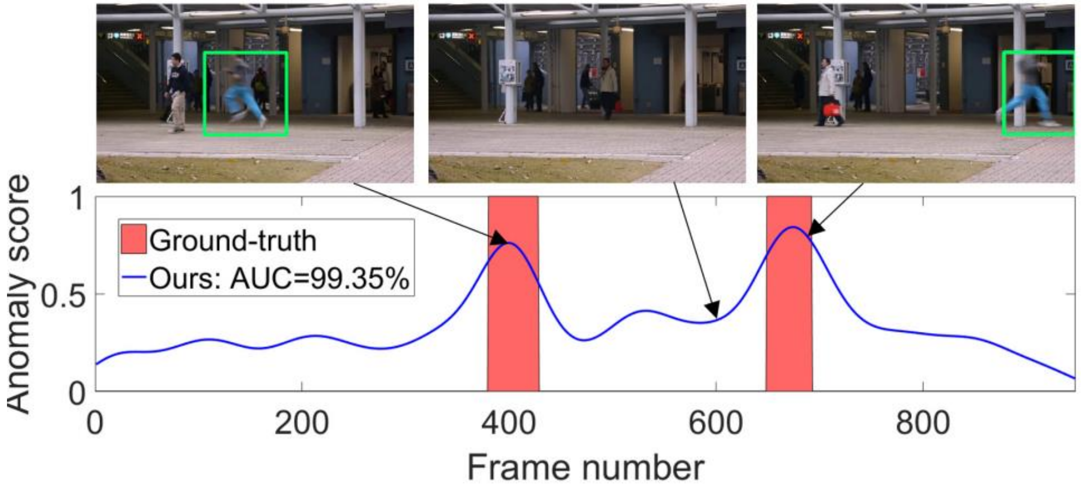
- Video Anomaly Detection (VAD) aims to determine whether abnormal events occur within video streams
- Abnormal events include the appearance or action of objects that are not suitable for the situation
- The goal is to do **Binary Classification** on each frame

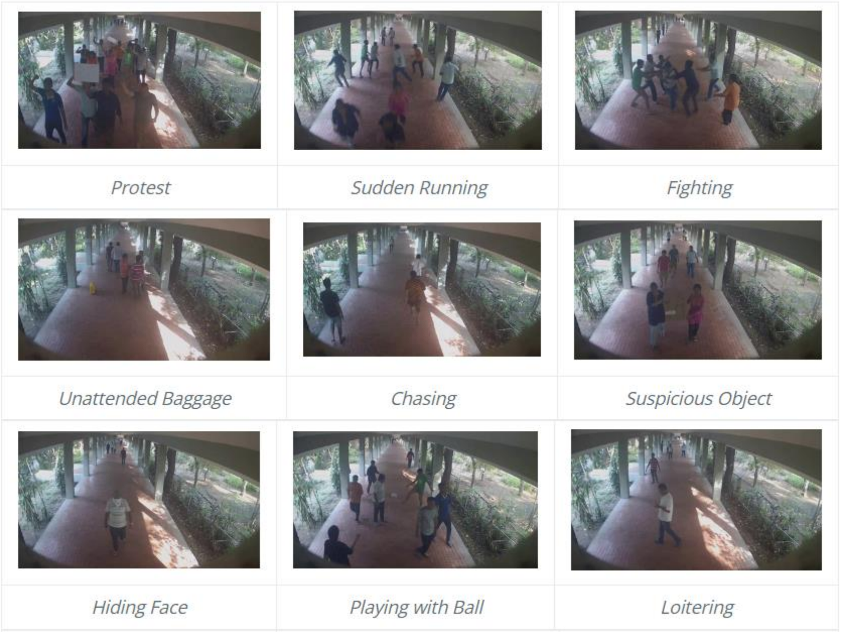


## Background

# One Class Classification

- Class imbalance problem $|\{x_i \mid y_i=0\}| \gg |\{x_i \mid y_i=1\}|$
- Diverse anomaly
- **One-Class Classification (OCC)** is utilized that learns exclusively from normal data and classifies anything not resembling the patterns of normal data as abnormal



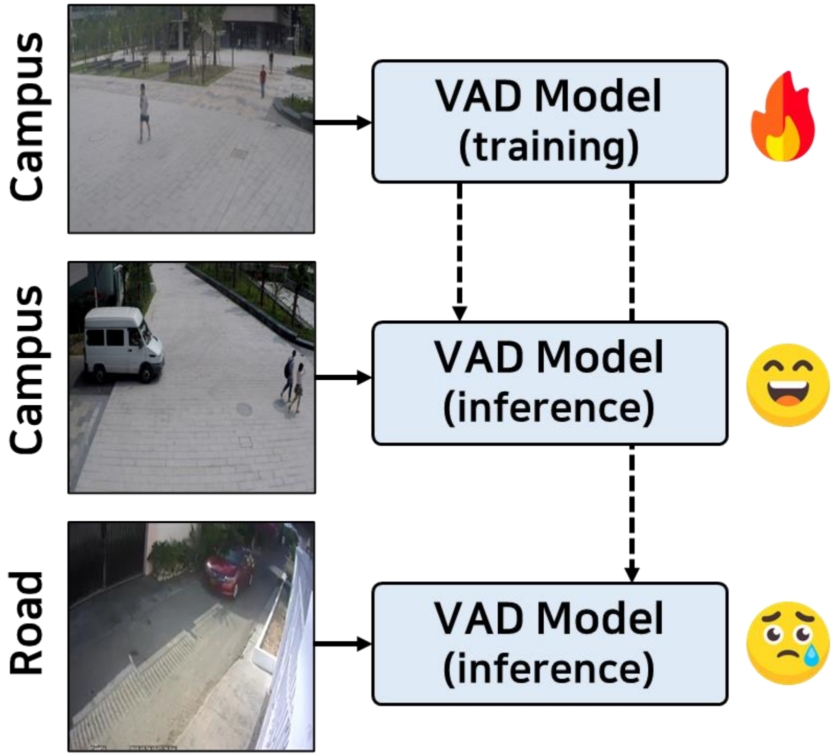

Rodrigues, Royston, et al. "Multi-timescale trajectory prediction for abnormal human activity detection." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020.

## Introduction

# Problem Definition

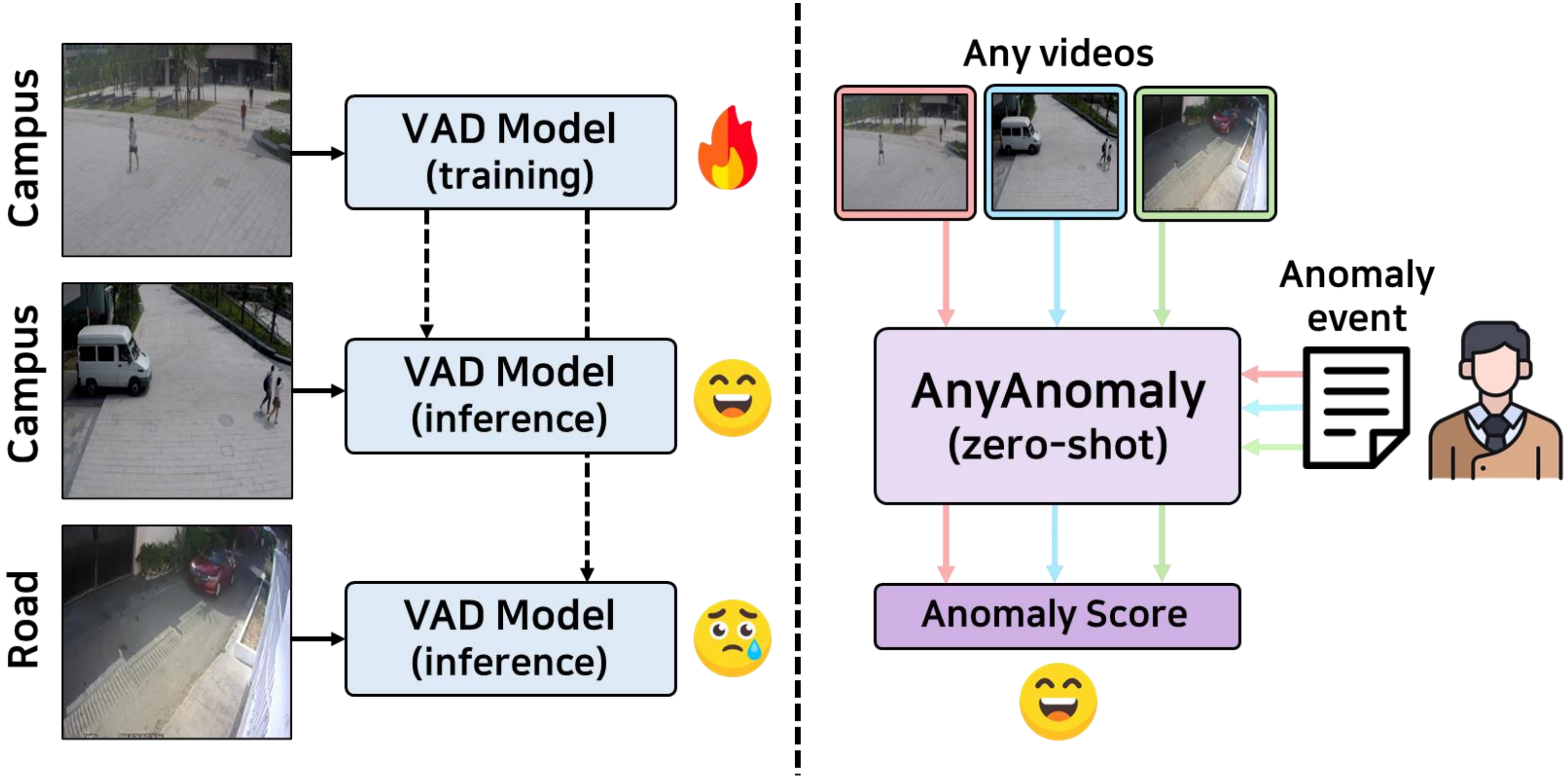
- VAD models learn normal patterns under specific scenarios, limiting generalization to diverse environments
- Models trained on pedestrian zones classify vehicles as anomalies, making deployment in road scenes difficult
- Adapting to new scenarios requires additional training or building separate models

  → increased computational cost and data collection



# Key Idea

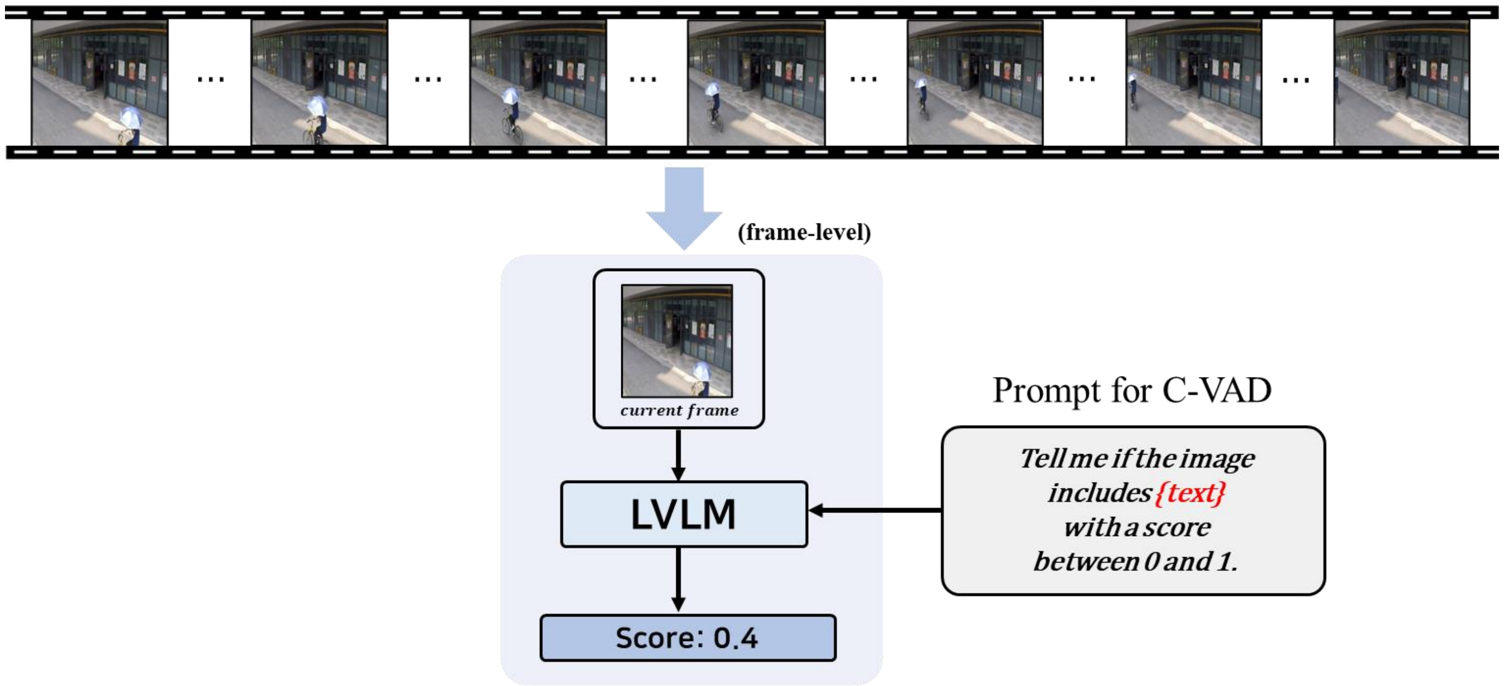
- **Customizable Video Anomaly Detection (C-VAD)**: a zero-shot, text-driven VAD technique
- User-defined text descriptions specify abnormal events; frames containing the described events are detected
- No scenario-specific retraining or separate models are required



## Introduction

# Baseline

- Perform **frame-level VQA using an LVLM** to estimate an anomaly score
- Prompt the model to return a value between 0 (no) and 1 (yes),
  indicating the degree to which the input image contains the user-defined abnormal event
- However, this approach may suffer from several limitations

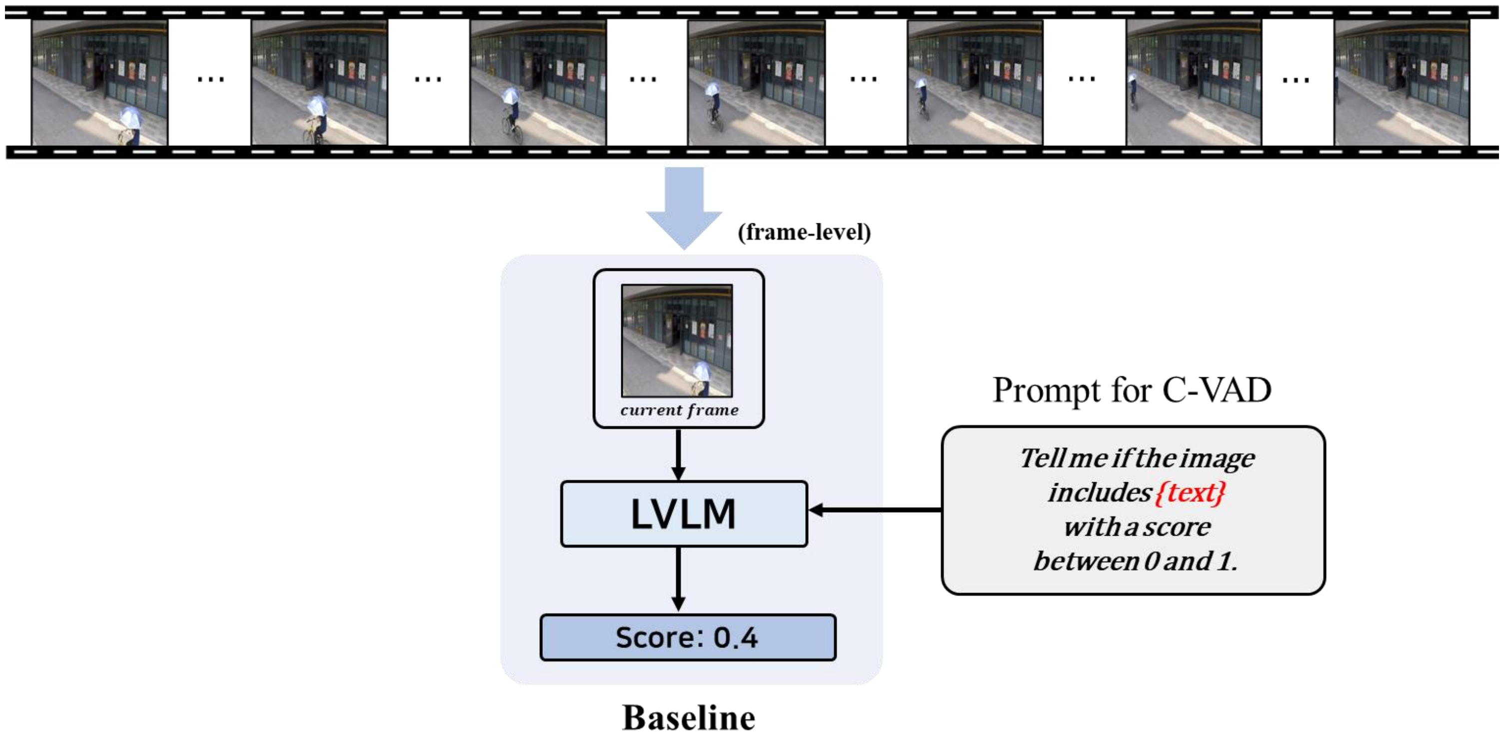



**Baseline**

- VQA: Visual Question Answering
- LVLM: Large Vision Language Model

# Limitations of Baseline

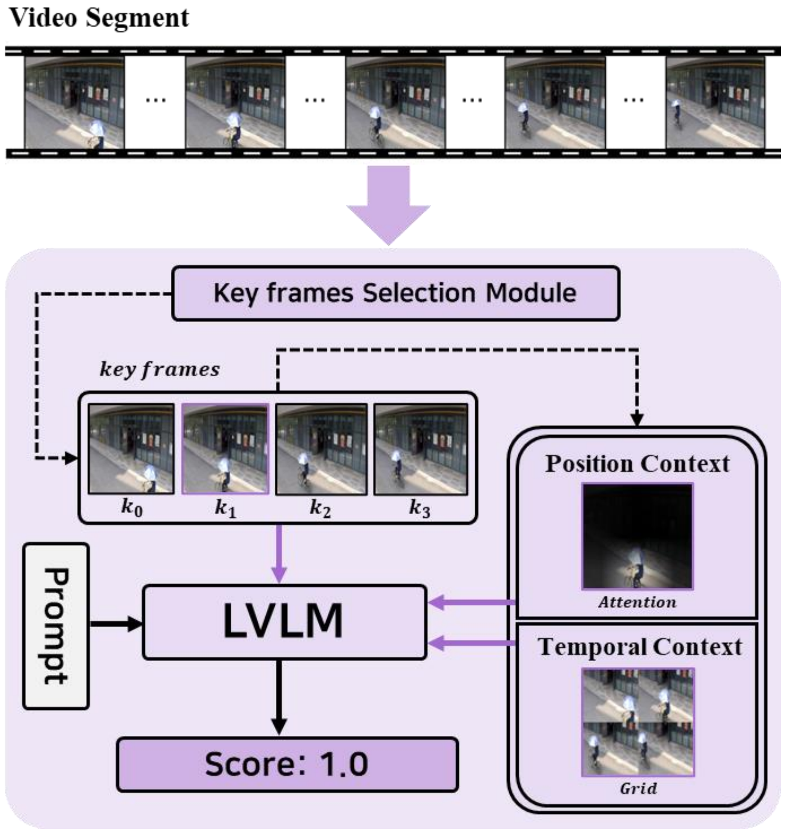
- As the number of frames increases, inference latency grows significantly
- Due to the characteristics of CCTV footage (foreground-background imbalance, object congestion), accurate object-level analysis can be challenging
- Processing images independently limits temporal reasoning and makes behavior analysis difficult



## Introduction

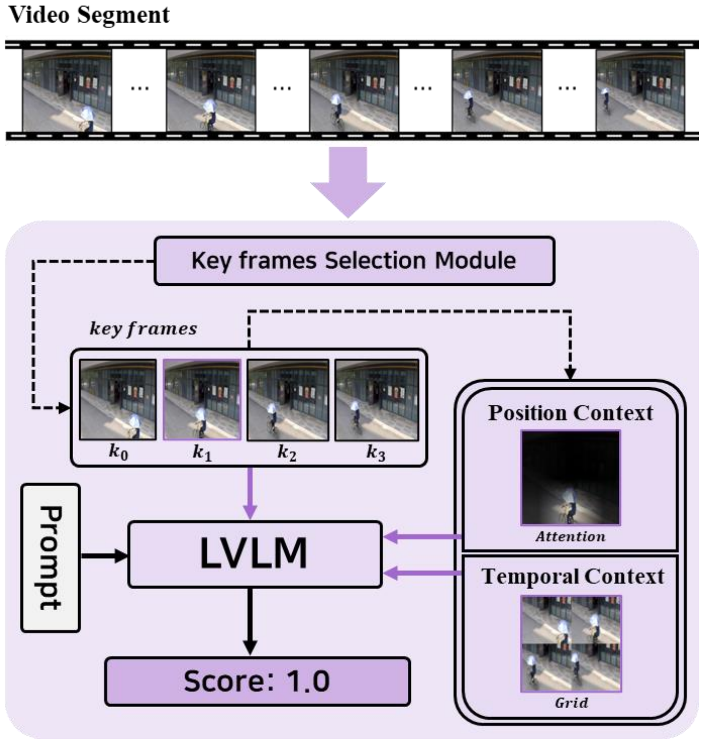
# AnyAnomaly

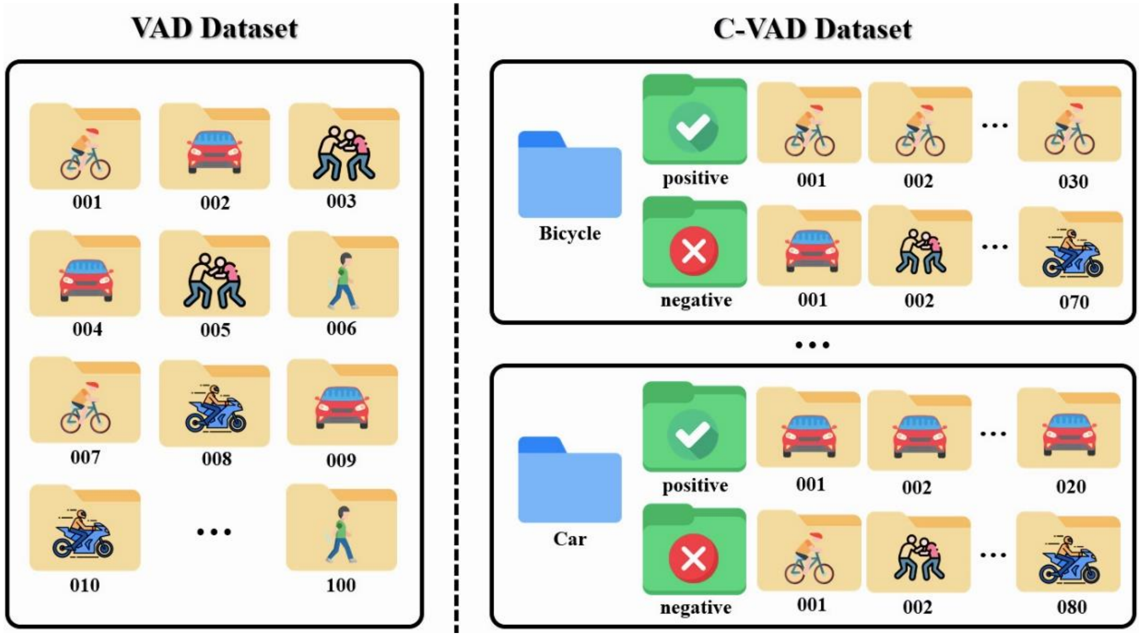
- **Segment-level approach:** select key frames and perform VQA per segment
- **Context-aware VQA:** perform VQA using additional contexts that represent the image
- **Position Context** emphasizes important spatial regions, while **Temporal Context** represents scene changes over time



# Contributions

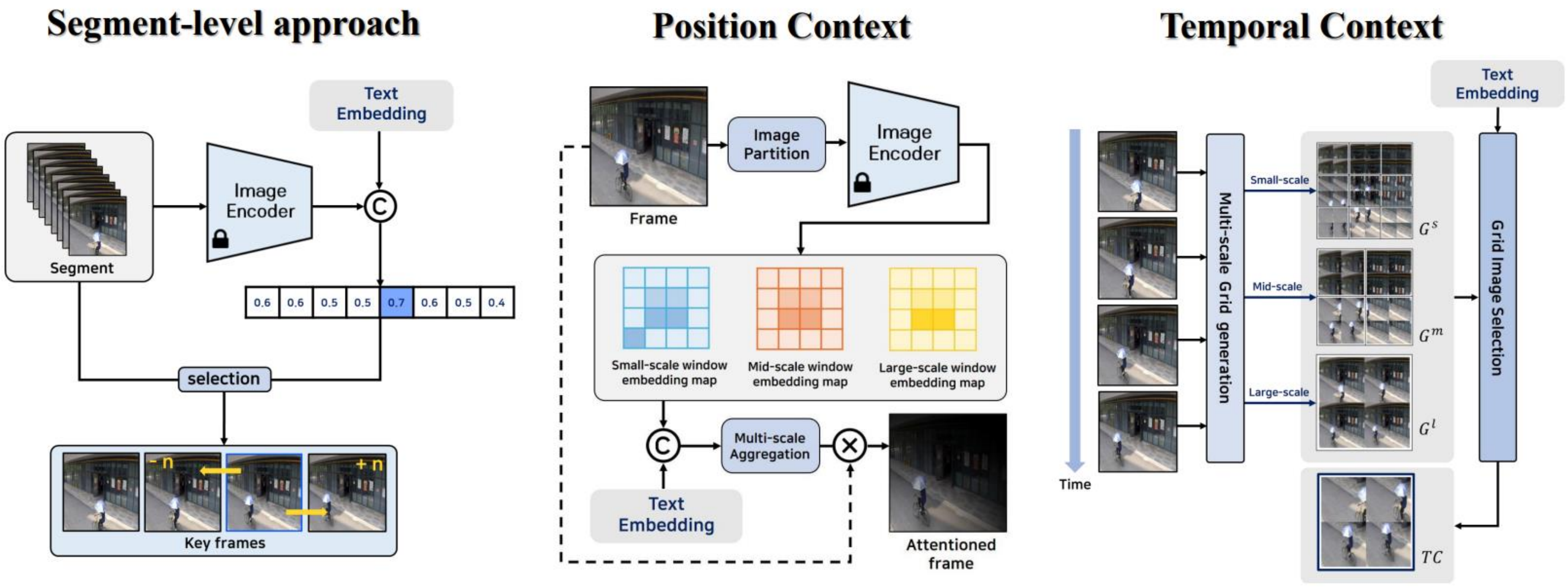
- Propose **C-VAD** for anomaly detection across diverse environments
- Develop **AnyAnomaly**, a context-aware VQA model for C-VAD
- Construct **C-VAD datasets** for rigorous evaluation and validate the effectiveness of the proposed method



# Method
## Architecture

- **KSM:** method for segment-level approach *(reduces inference time)*
- **WA:** method for generating position context *(improves object-level analysis)*
- **GIG:** method for generating temporal context *(enhances behavior analysis)*

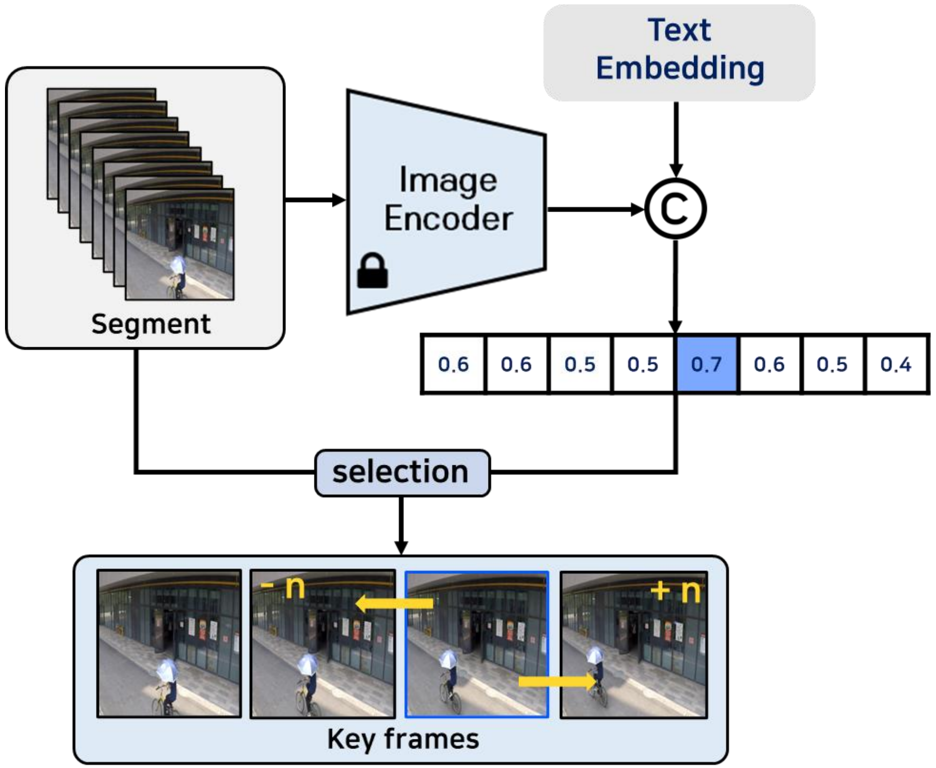



(a) Key frames Selection Module (KSM)  (b) WinCLIP-based Attention (WA)  (c) Grid Image Generation (GIG)

## Method

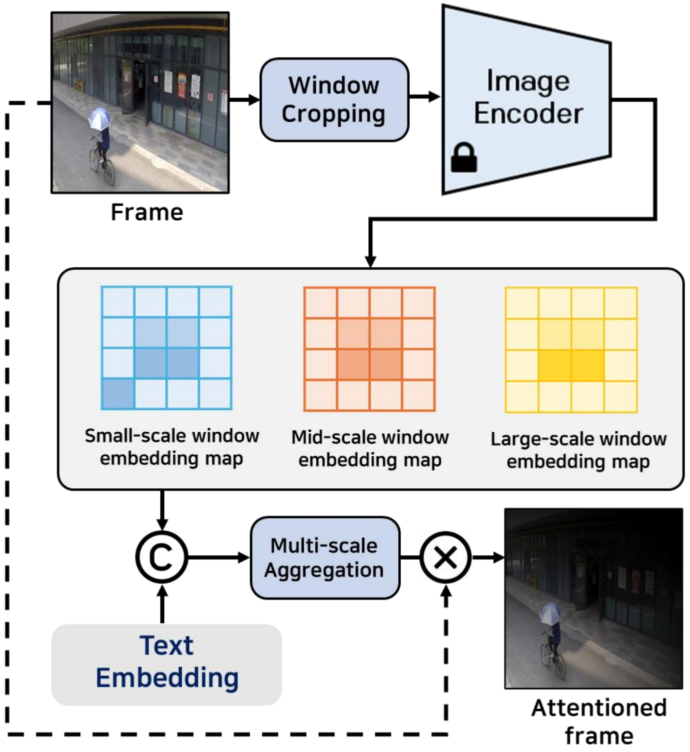
# Key frames Selection Module

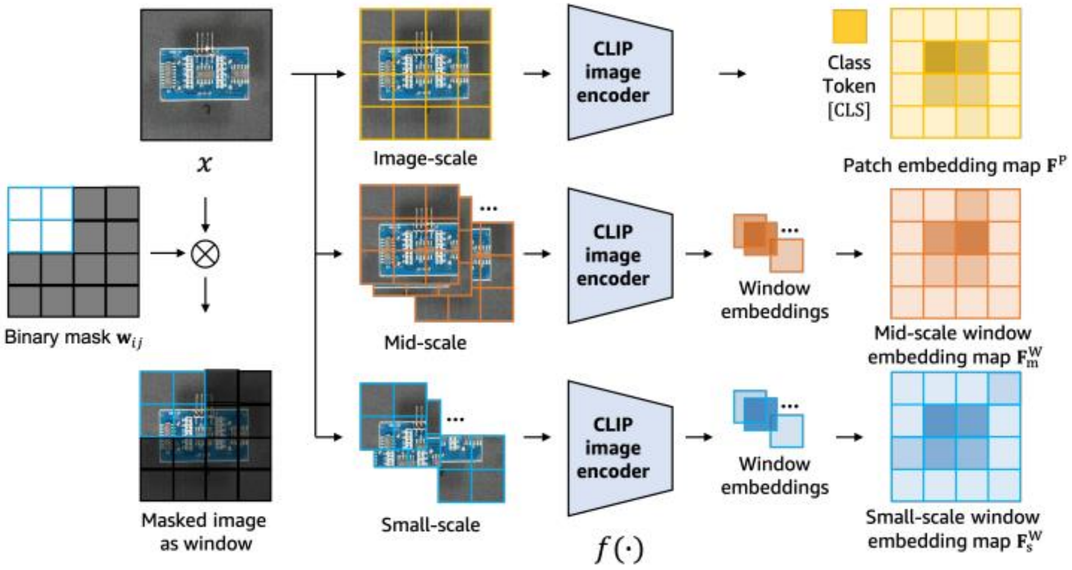
- **Temporal uniformity** and **text alignment** are critical in key frames extraction
- Use an image encoder (e.g., CLIP) to select frame most similar to the text embedding
- Divide the segment into multiple groups and select remaining frames based on the positions of the selected frame



# WinCLIP-based Attention

- Based on the **window embedding maps** proposed in WinCLIP, **identify regions corresponding to the text**
- Multi-scale embedding maps provide representations at both local and global scales
- The similarity map is obtained by averaging similarities across multiple scales and then used to reweight the input image



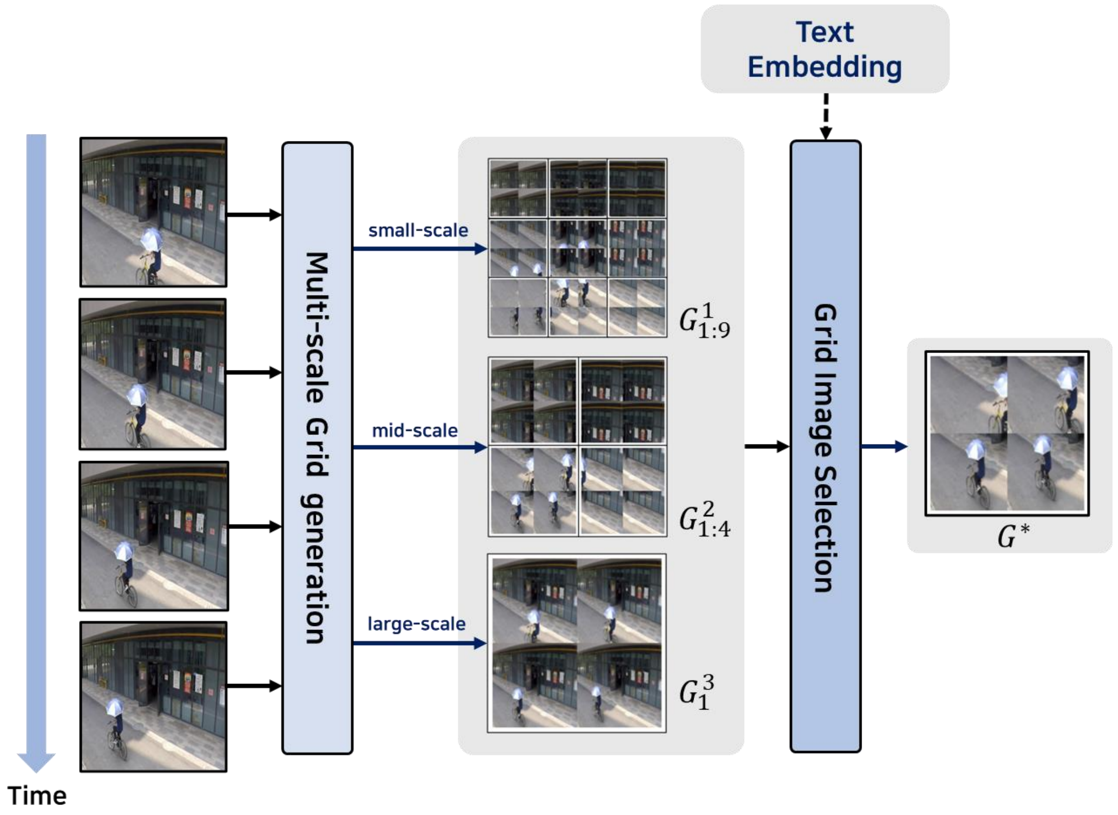




Jeong, J et al., "WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation", Proceedings of the IEEE conference on computer vision and pattern recognition, pp.19606-19616, 2023.

## Method

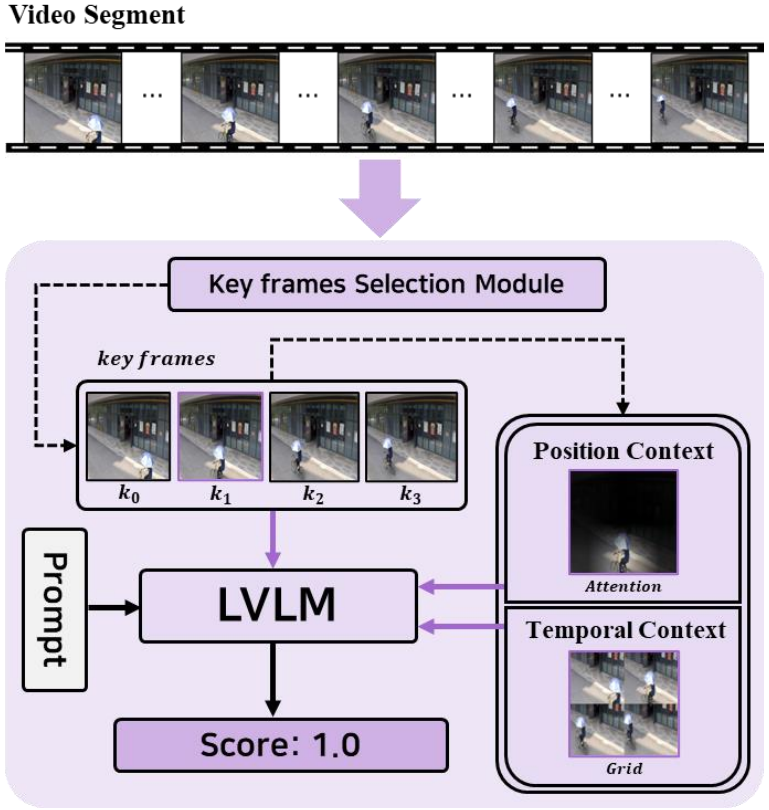
# Grid Image Generation

- Following the same procedure as WA, **windows at the same spatial position across key frames are grouped** at each scale
- The grouped windows are arranged into a 2×2 grid to form a grid image
- The grid image most relevant to the text embedding is selected



# Anomaly Scoring

- LVLM evaluates the representative key frame $\hat{k}$ and position context $PC$ using the shared prompt $P$
- Temporal context $TC$ is evaluated with a modified prompt $P^*$
- The outputs are aggregated to compute the final anomaly score via late fusion




**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, including a brief reason for the score in one short sentence.

**Prompt for temporal context**

**Context:** The given image represents a sequence (row 1 column 1 → row 1 column 2 → row 2 column 1 -> row 2 column 2) illustrating temporal progression.

$$ascore = \gamma_1 \cdot \Phi_{\text{LVLM}}(\hat{k}, P) + \gamma_2 \cdot \Phi_{\text{LVLM}}(PC, P) + \gamma_3 \cdot \Phi_{\text{LVLM}}(TC, P^*)$$

# Results

# Main Results

- In existing VAD datasets, videos are not categorized by anomaly classes
- The **proposed C-VAD datasets** categorize videos by anomaly classes
- Each class contains positive and negative samples for evaluating user-specified anomaly detection

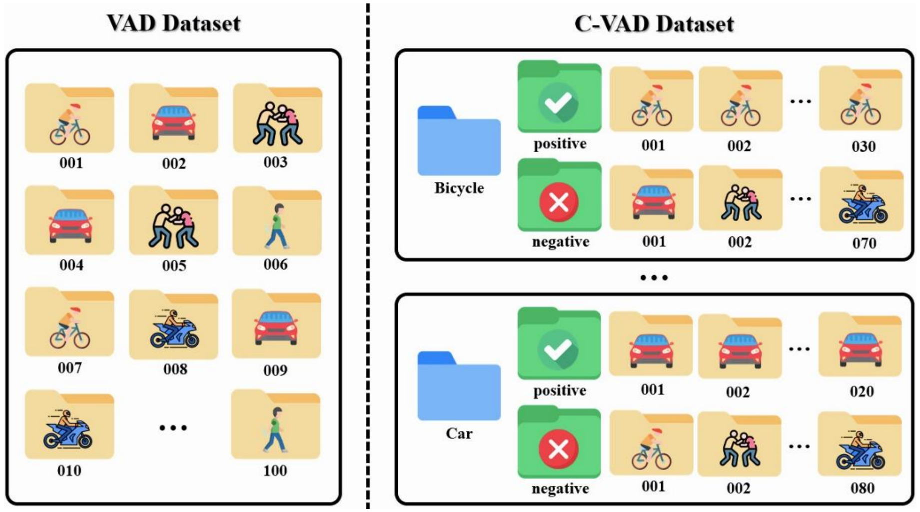



Table 1. Performance comparison on C-ShT dataset. The best results are **bolded**. The second-best results are underlined.

| Category | Class | Baseline | +KSM | +KSM/PC | +KSM/TC | Proposed | Improvement (%) |
|---|---|---|---|---|---|---|---|
| Action | Skateboarding | 61.30 | 57.06 | 57.79 | **73.66** | **73.66** | +20.16 |
| | Throwing | **91.41** | 72.82 | 88.74 | 82.53 | 90.67 | -0.81 |
| | Running | 53.13 | 51.93 | 53.68 | 59.77 | **60.11** | +13.14 |
| | Loitering | 61.98 | 51.96 | **81.27** | 76.94 | **81.27** | +31.12 |
| | Jumping | 82.84 | 92.89 | 92.91 | **95.31** | **95.31** | +15.05 |
| | Falling | 78.31 | 78.95 | 79.24 | **88.01** | **88.01** | +12.39 |
| | Fighting | 84.48 | 91.18 | 91.18 | **98.06** | **98.06** | +16.07 |
| | **Average** | 73.35 | 72.00 | 77.83 | 82.04 | **83.87** | +14.34 |
| Appearance | Car | 88.72 | 90.96 | **91.46** | 90.96 | **91.46** | +3.09 |
| | Hand truck | 95.50 | 98.20 | 98.91 | **99.20** | **99.20** | +3.87 |
| | Bicycle | 72.36 | 72.46 | **78.47** | 72.46 | **78.47** | +8.44 |
| | Motorcycle | **88.04** | 86.72 | 86.72 | 86.72 | 86.72 | -1.50 |
| | **Average** | 86.16 | 87.09 | 88.89 | 87.34 | **88.95** | +3.25 |
| **Overall Average** | | 78.01 | 77.48 | 81.85 | 83.97 | **85.72** | +9.88 |

Table 2. Performance comparison on C-Ave dataset

| Category | Class | Baseline | +KSM | +KSM/PC | +KSM/TC | Proposed | Improvement (%) |
|---|---|---|---|---|---|---|---|
| Action | Throwing | 78.44 | 80.13 | **89.77** | 82.40 | **89.77** | +14.44 |
| | Running | 75.82 | 77.67 | 77.67 | **77.90** | **77.90** | +2.74 |
| | Dancing | 85.65 | 72.28 | 76.64 | **91.92** | **91.92** | +7.32 |
| | **Average** | 79.97 | 76.69 | 81.36 | 84.07 | **86.53** | +8.2 |
| Appearance | Too close | 57.23 | 61.48 | 61.48 | **91.78** | **91.78** | +60.37 |
| | Bicycle | 99.99 | 99.84 | 99.99 | 99.93 | **100.00** | +0.01 |
| | **Average** | 78.61 | 80.66 | 80.74 | 95.86 | **95.89** | +21.98 |
| **Overall Average** | | 79.43 | 78.28 | 81.11 | 88.79 | **90.27** | +13.65 |

# Ablation Study

Table S2. Comparison on segment length

| Segment length | C-ShT | C-Ave | FPS |
|---|---|---|---|
| Baseline | 78.01 | 79.43 | 0.96 |
| 8 | 83.83 | 83.96 | 2.67 |
| 16 | 83.45 | 87.45 | 4.49 |
| 24 | **85.72** | **90.27** | 6.67 |
| 32 | 82.50 | 85.94 | **8.45** |

Table 4. Comparison on window size.

| Window Size | C-ShT | | | C-Ave | | |
|---|---|---|---|---|---|---|
| | Act. | App. | Total | Act. | App. | Total |
| small | 78.8 | **90.6** | 83.1 | 84.7 | 87.1 | 85.7 |
| middle | 81.2 | 89.0 | 84.1 | **87.5** | 92.0 | 89.3 |
| large | 82.1 | 89.7 | 84.9 | 86.8 | 86.4 | 86.6 |
| all | **83.9** | 89.0 | **85.7** | 86.5 | **95.9** | **90.3** |

Table S4. Comparison of diverse LVLMs. The model highlighted in blue represents the most efficient model for the C-VAD task, while the one highlighted in purple indicates the most effective model. For further comparison, additional experiments were conducted using Qwen-based models. *: Experiment conducted using vLLM.

| LVLM | Pre-trained | C-ShT | | C-Ave | | FPS |
|---|---|---|---|---|---|---|
| | | w/o context | Proposed | w/o context | Proposed | |
| Chat-UniVi[14] | Chat-UniVi-7B | 77.5 | 85.7 | 78.3 | 90.3 | 6.67 |
| MiniGPT-4[44] | LLaMA-2 Chat 7B | 54.0 | 67.4 | 53.9 | 55.3 | 1.26 |
| MiniCPM-V[41] | MiniCPM-Llama3-V-2_5 (8B) | 87.7 | **90.1** | 86.3 | **91.0** | 1.36 |
| LLAVA++[28] | LLaVA-Meta-Llama-3-8B-Instruct-FT | 73.3 | 82.8 | 59.0 | 69.4 | 7.25 |
| Qwen2.5-VL[6] | Qwen2.5-VL-3B-Instruct | 89.0 | 90.2 | 78.0 | 87.0 | 11.18 |
| Qwen2.5-VL*[6] | Qwen2.5-VL-3B-Instruct | 88.6 | 90.2 | 78.3 | 88.1 | 34.78 |
| Qwen2.5-VL*[6] | Qwen2.5-VL-7B-Instruct | 93.0 | **95.5** | 86.9 | **92.4** | 24.08 |

## Results

# Performance Comparison

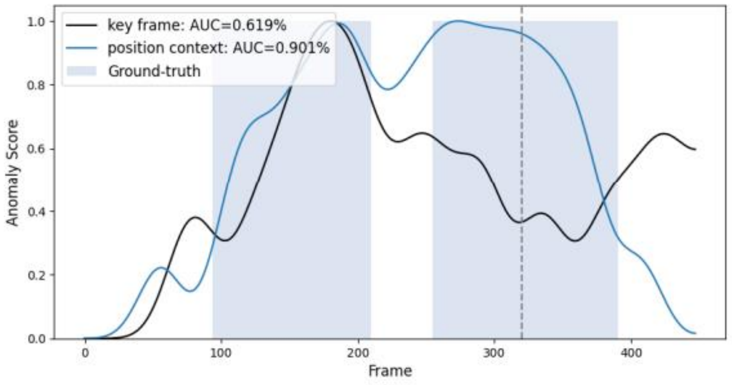
Table 5. Comparison with state-of-the-art VAD methods. * indicates testing without context.

| Method | Zero-shot | Ave | ShT | UB | UCF |
|---|---|---|---|---|---|
| AMMC-Net[7] | ✗ | 86.6 | 73.7 | - | - |
| STEAL-Net[5] | ✗ | 87.1 | 73.7 | - | - |
| MPN[23] | ✗ | 89.5 | 73.8 | - | - |
| DLAN-AC[39] | ✗ | 89.9 | 74.7 | - | - |
| UBnormal[1] | ✗ | - | - | 68.5 | - |
| FPDM[37] | ✗ | 90.1 | 78.6 | 62.7 | 74.7 |
| SLM[33] | ✗ | 90.9 | 78.8 | - | - |
| USTN-DSC[40] | ✗ | 89.9 | 73.8 | - | - |
| AnomalyRuler[38] | ✗ | 89.7 | **85.2** | 71.9 | - |
| MULDE[26] | ✗ | - | 81.3 | 72.8 | 78.5 |
| AED-MAE[30] | ✗ | **91.3** | 79.1 | 58.5 | - |
| MA-PDM[43] | ✗ | **91.3** | 79.2 | 63.4 | - |
| AccI-VAD[29] | ✗ | - | 76.2 | 66.8 | 60.3 |
| AnyAnomaly* | ✓ | 81.4 | 77.2 | 73.1 | 77.8 |
| AnyAnomaly | ✓ | 87.3 | 79.7 | **74.5** | **80.7** |

Table 6. Generalization performance comparison. Tr.: cross-domain training where models trained on one VAD dataset are evaluated on another. Few.: methods that adapt to the target domain using only a few training samples, Aux.: methods that utilize auxiliary datasets, *: since competitors did not perform cross-domain evaluations on ShT, we present their same-domain results instead.

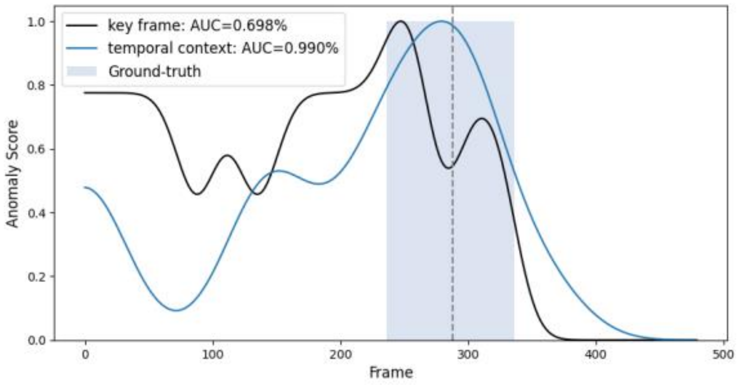
| Method | Tr. | Few. | Aux. | Ave | ShT |
|---|---|---|---|---|---|
| STEAL-Net[5] | ✓ | ✗ | ✗ | 54.3 | 51.7 |
| Jigsaw[35] | ✓ | ✗ | ✗ | 62.9 | 59.3 |
| rGAN[21] | ✓ | ✓ | ✗ | 76.6 | 77.9* |
| MPN[23] | ✓ | ✓ | ✗ | 78.9 | 73.8* |
| zxVAD[3] | ✓ | ✗ | ✓ | 82.2 | 71.6* |
| Shibao et al.[9] | ✓ | ✗ | ✓ | 86.2 | 78.7 |
| ZS CLIP[27] | ✗ | ✗ | ✗ | 62.3 | 60.9 |
| ZS ImageBind[10] | ✗ | ✗ | ✗ | 64.5 | 61.3 |
| LLaVA-1.5[18] | ✗ | ✗ | ✗ | 67.4 | 59.6 |
| Video-ChatGPT[24] | ✗ | ✗ | ✗ | 76.9 | 69.1 |
| AnyAnomaly | ✗ | ✗ | ✗ | **87.3** | **79.7** |

# Qualitative Results




The image does not show a clear **bicycle**, so the score is 0.4.

The image includes a group of people riding **bicycles**, with at least one clearly visible bicycle. The score would be 0.9.

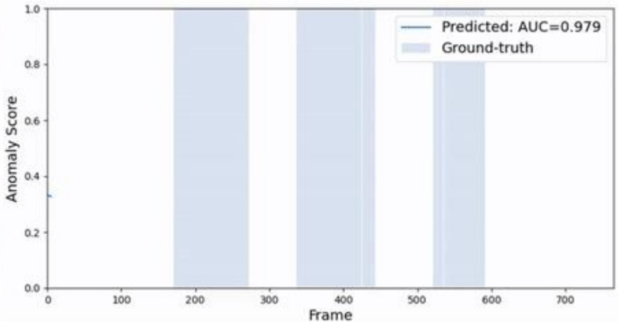



The image shows a group of young men standing on a paved area, but there is no evidence of **fighting**. The score is 0.5.

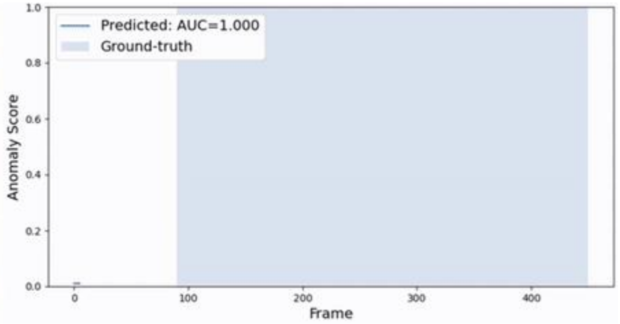
The image shows a group of people **fighting**, with at least two individuals engaged in physical altercations. The score is 0.9.

## Results

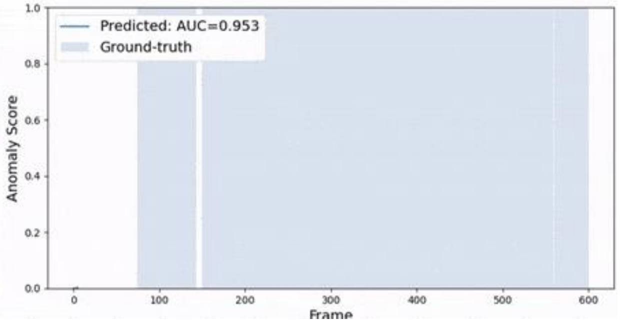
# Demos

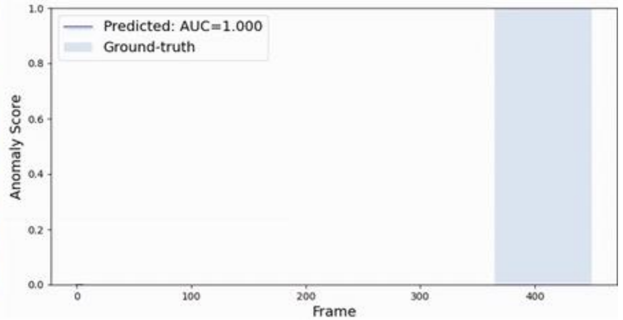




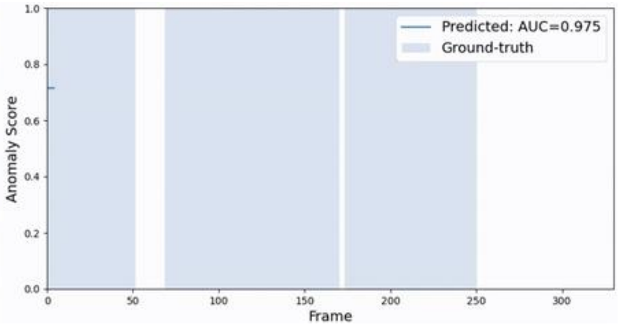




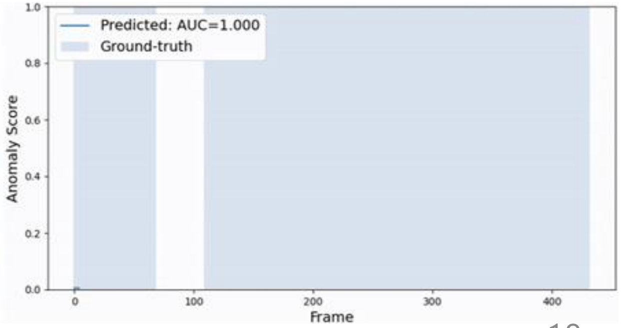












# Thank you