Harnessing Large Language Models for Training-free Video Anomaly Detection

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, Elisa Ricci

Presenter: Sunghyun Ahn

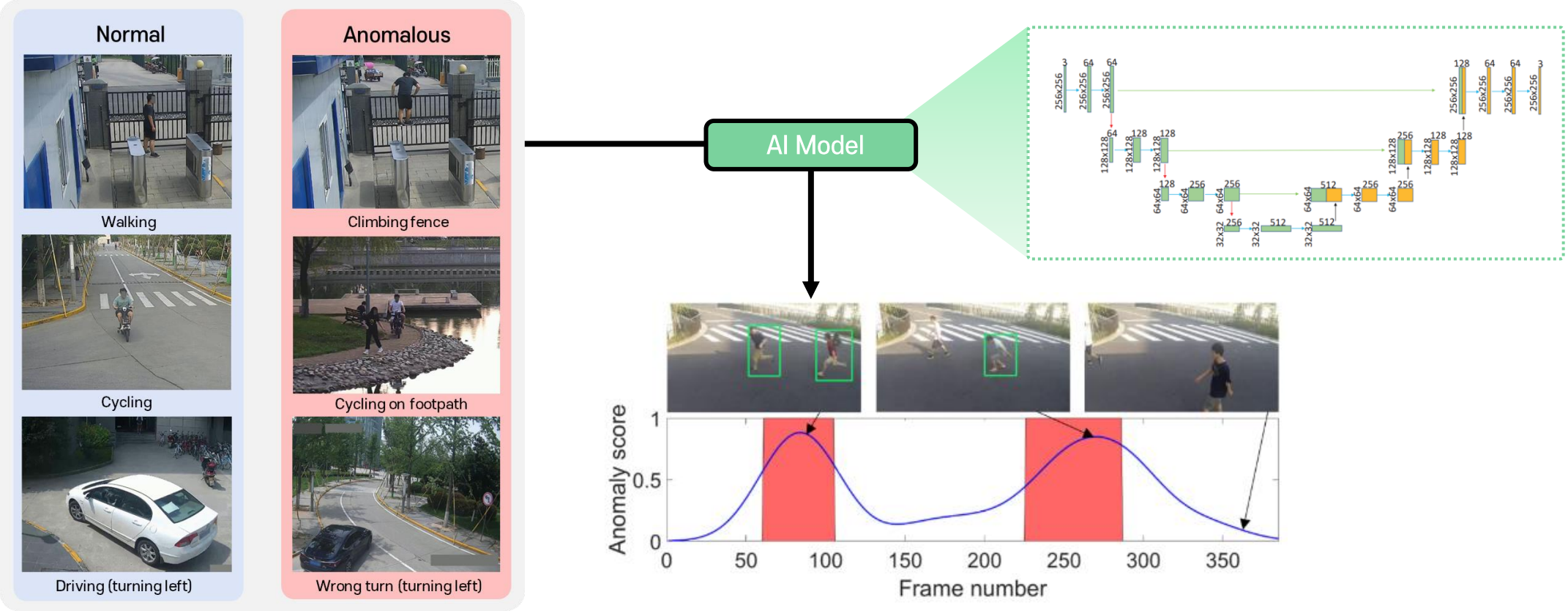
Department of Computer Science, Yonsei University

skd@yonsei.ac.kr

Introduction

Video Anomaly Detection

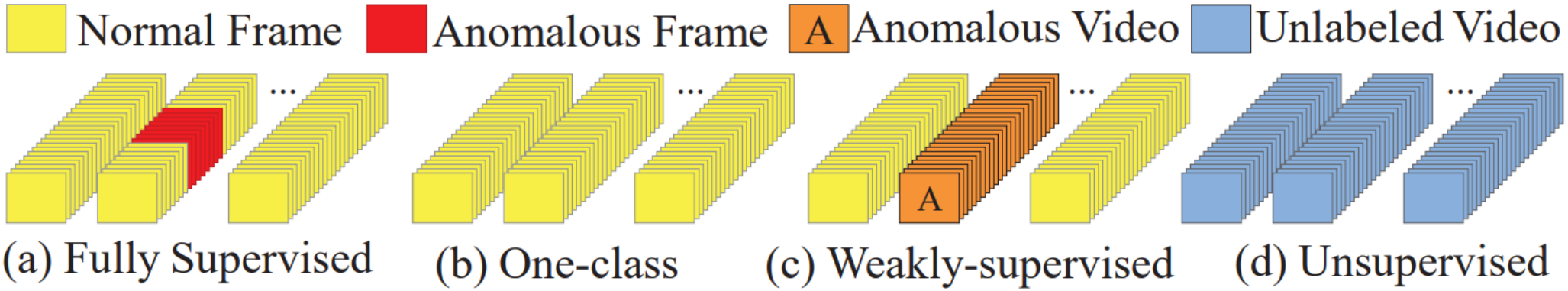
- Video Anomaly Detection (VAD) aims to identify abnormal events within video streams
- Abnormal events include the appearance or behavior of objects that are not suitable for the situation
- VAD has high annotation costs, various effective training methods are used, not just supervised learning



Introduction

Four Training Methods

- **Fully Supervised:** frame-level normal/abnormal annotations in the training data
- **One-class:** only normal training data
- **Weakly-supervised:** video-level normal/abnormal annotations
- **Unsupervised:** no training data annotations

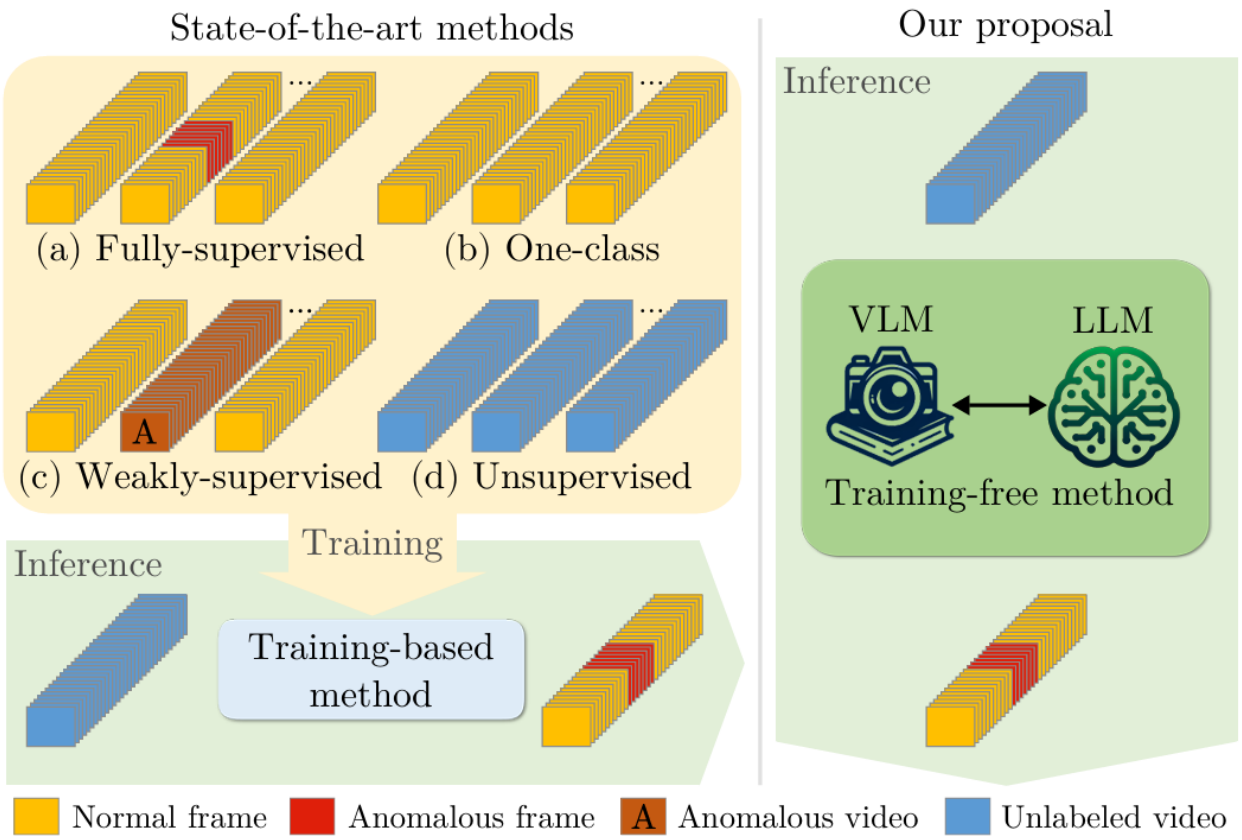


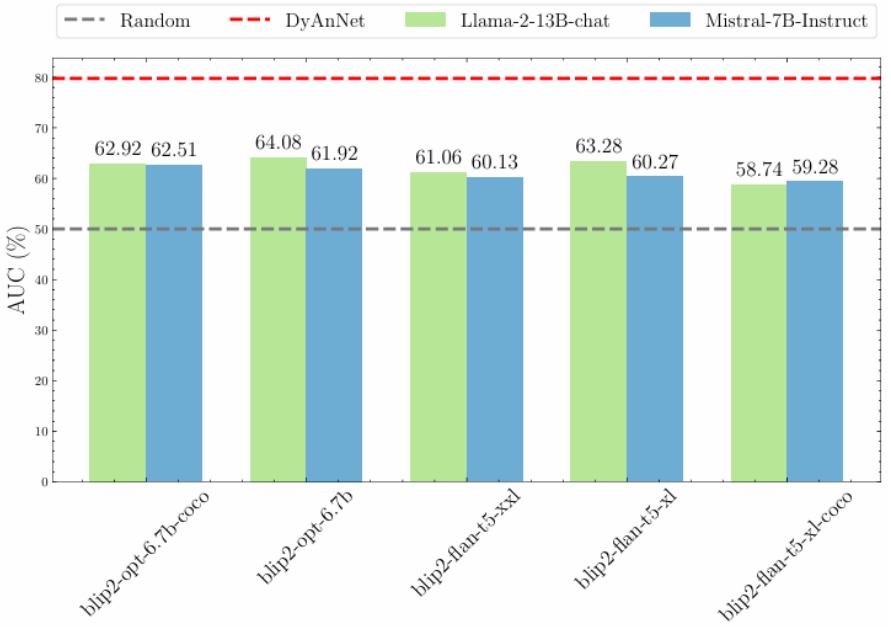
Zaheer, M. Zaigham, et al. "Generative cooperative learning for unsupervised video anomaly detection." Proceedings of the IEEE/CVF conference on c omputer vision and pattern recognition. 2022.

Introduction

Language-based VAD

- Propose **Language-based VAD (LAVAD)** using only LLM and VLM without training
- Using VLM (BLIP) and LLM (LLaMA) directly results in lower performance compared to unsupervised methods
- Address the issue of insufficient performance when using VLM and LLM directly





Limitations of Frame-level caption

- Incorrect captions may occur *(captions unrelated to the frame, normal captions extracted via abnormal frames)*
- Difficult to secure global context *(i.e. running vs chased)*, lack of dynamic information in the scene *(i.e. standing vs loitering)*

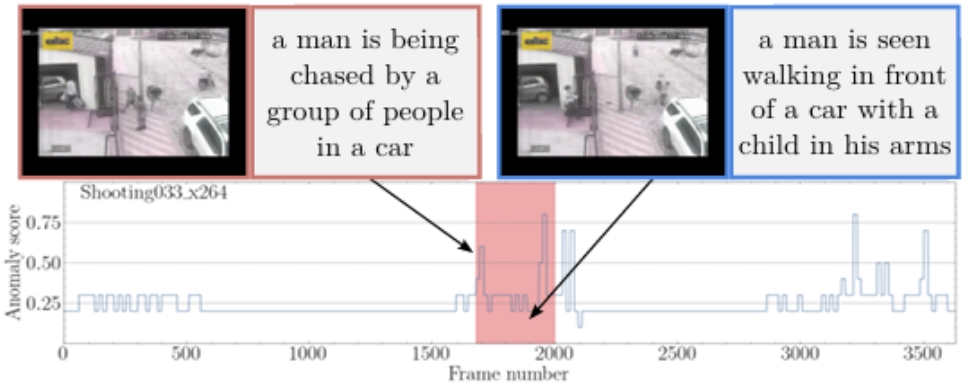
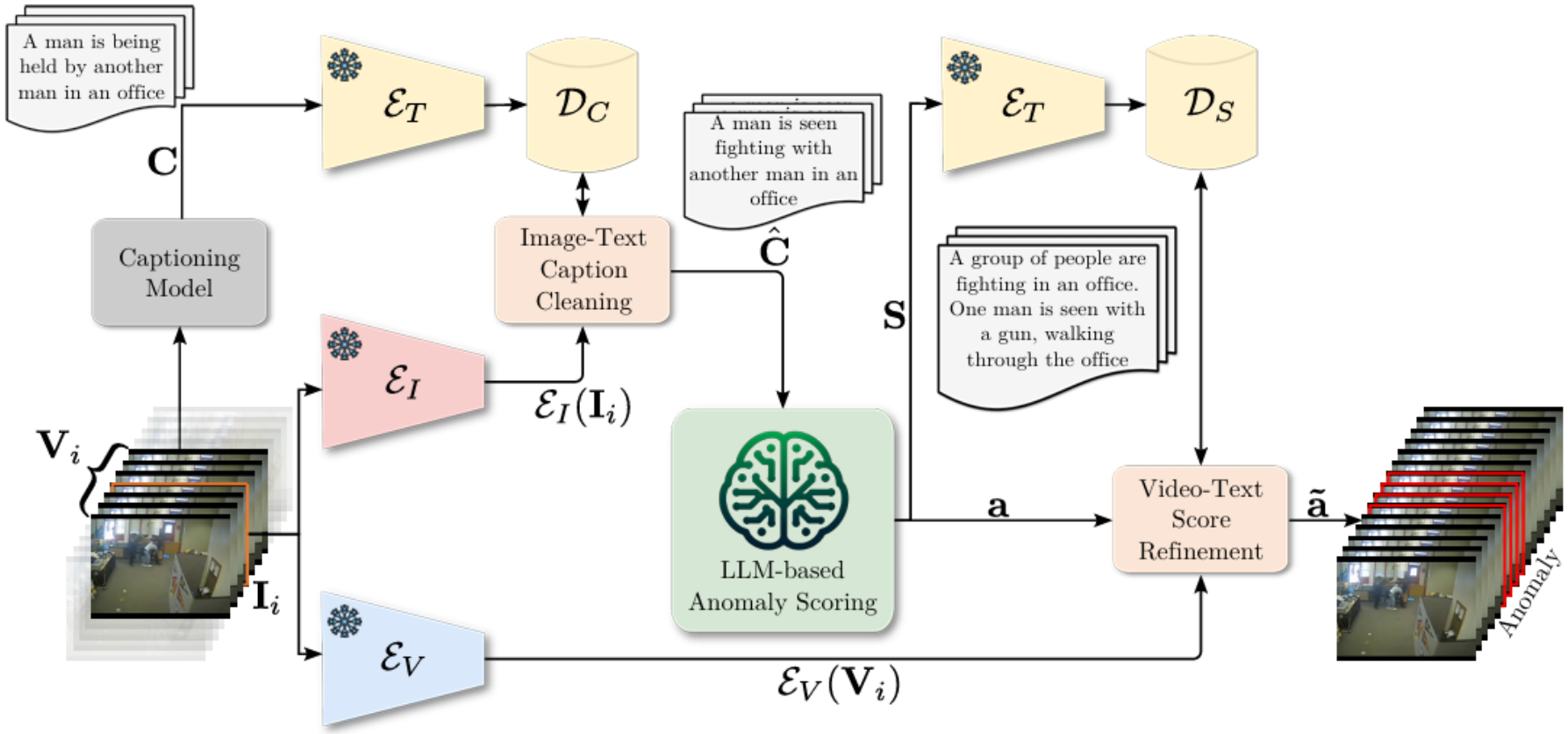


Figure 3. The anomaly score predicted by Llama [29] over time for video *Shooting033* from UCF-Crime. We highlight some sample frames with their associated BLIP-2 captions to demonstrate that the caption can be semantically noisy or incorrect (red bounding boxes are for abnormal predictions while blue bounding boxes are for normal predictions). Ground-truth anomalies are highlighted. In particular, the caption of the frame enclosed by a blue bounding box within the ground truth anomaly fails to accurately represent the visual content, leading to a wrong classification due to the low anomaly score given by the LLM.

Method

Language-based VAD

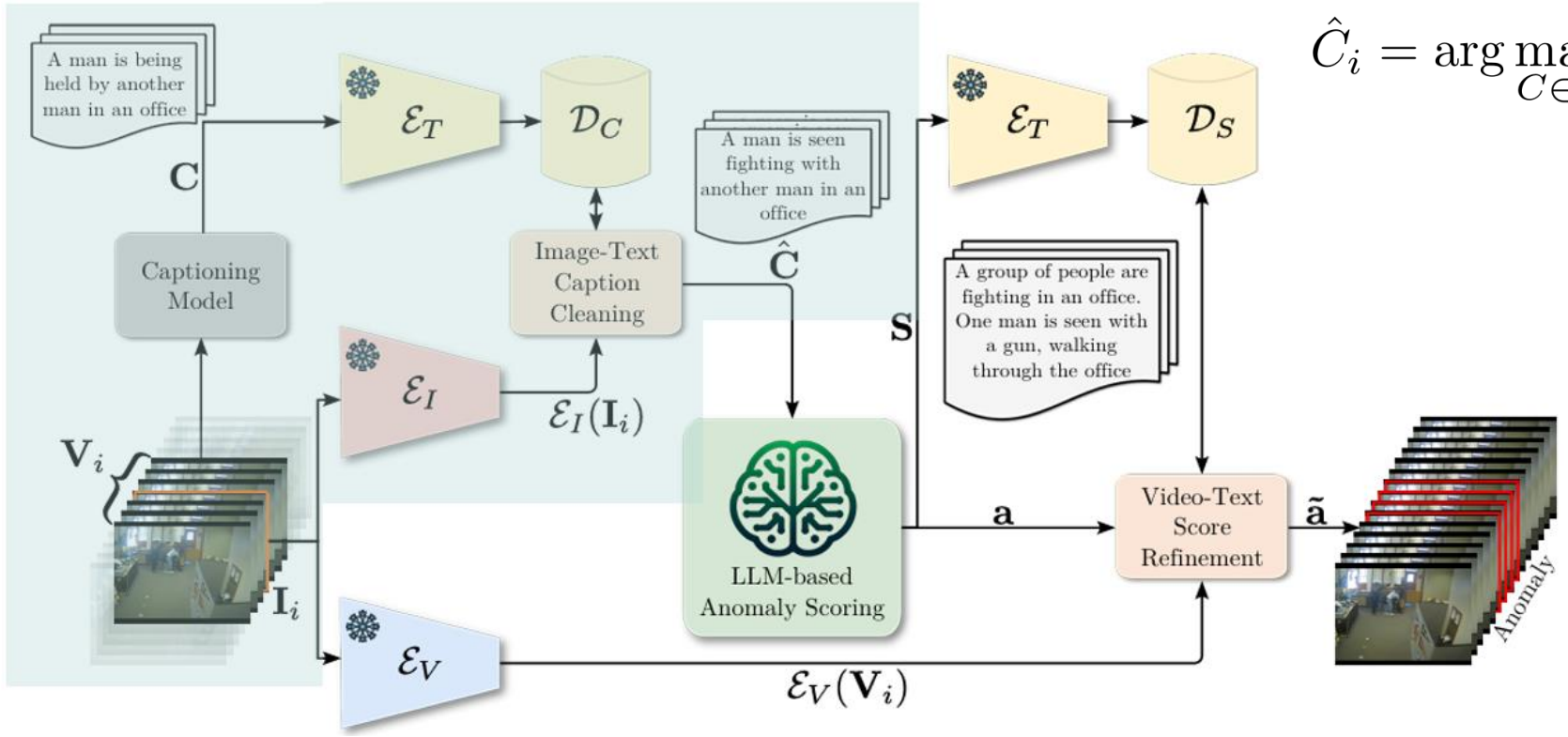
- **Image-Text Caption Cleaning** resolves noisy caption issues
- **LLM-based Anomaly Scoring** addresses the limitations related to lack of scene dynamics
- **Video-Text Score Refinement** improves performance by refining anomaly scores



Method

Image-Text Caption Cleaning

- Assume the existence of captions that better capture visual content among surrounding frames
- Replace the caption with text features that best correspond to the visual features of the frame
- Obtains more accurate frame captions (e.g., being held vs fighting)



$$\hat{C}_i = \arg\max_{C \in \mathbf{C}} \langle \mathcal{E}_I(\mathbf{I}_i) \cdot \mathcal{E}_T(C) \rangle,$$

Method

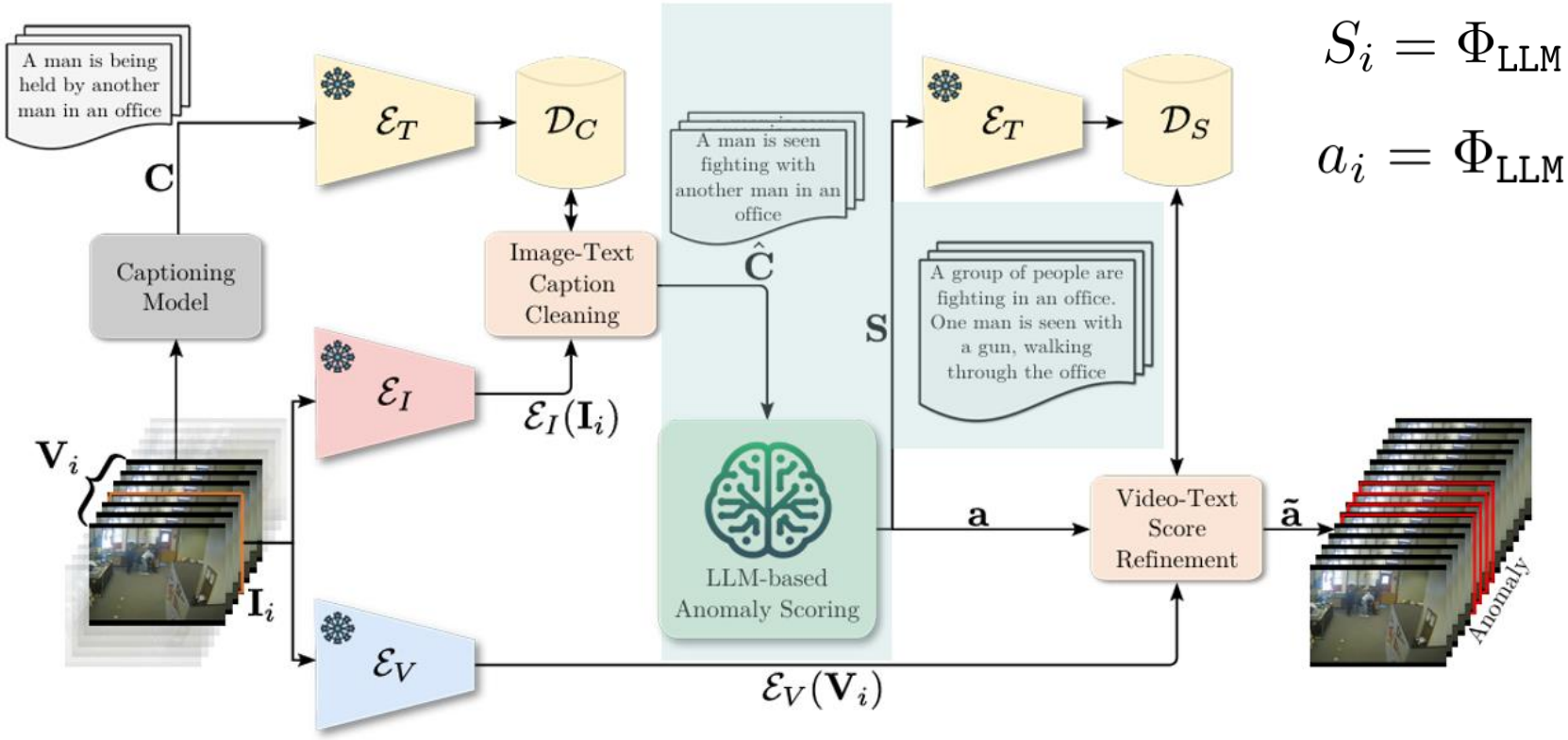
LLM-based Anomaly Scoring

- Perform **summarization** and **scoring** processes using LLM
- Summarize captions of surrounding frames (V_i) to obtain global and dynamic context
- Provide prompts for VAD to output an anomaly score (0-1)

P_S "Please summarize what happened in few sentences, based on the following temporal description of a scene."

P_C "If you were a law enforcement agency, how would you rate the scene described on a scale from 0 to 1,with 0 representing a standard scene and 1 denoting a scene with suspicious activities?"

P_f "Please provide only one number in the provided list below [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. "



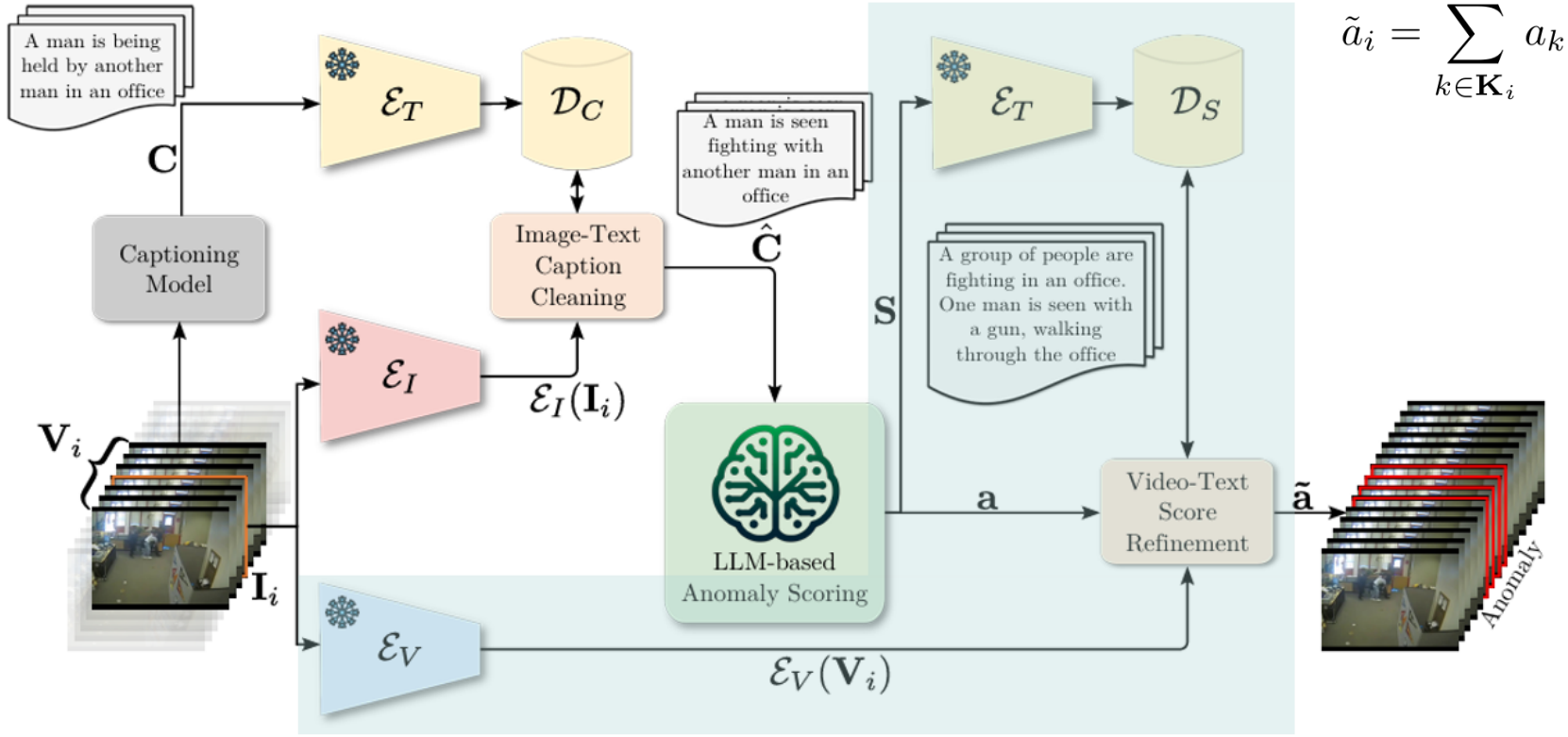
$$S_i = \Phi_{\text{LLM}}(\mathrm{P}_S \circ \hat{\mathbf{C}}_i)$$

$$a_i = \Phi_{\text{LLM}}(\mathrm{P}_C \circ \mathrm{P}_F \circ S_i)$$

Method

Video-Text Score Refinement

- Based on the summary of V_i (S_i), the initial Anomaly Score (a_i) may be inaccurate
- Because summaries are not always accurate, so the entire score set must be considered
- Find k summaries semantically closest to V_i and compute a weighted sum of anomaly scores corresponding to those summaries

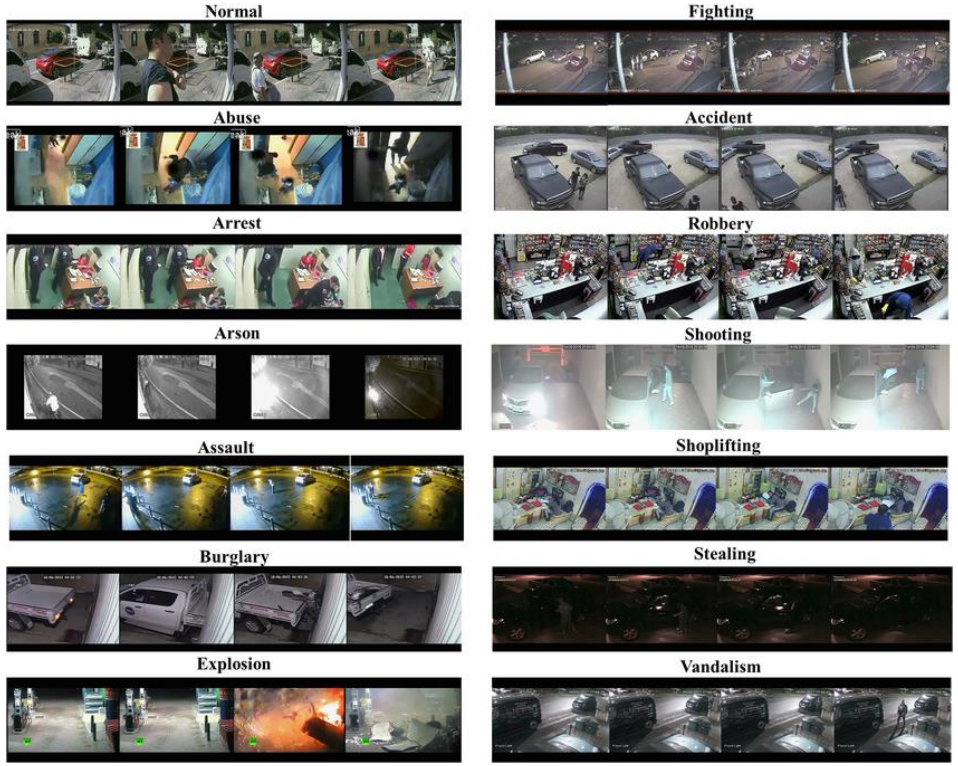


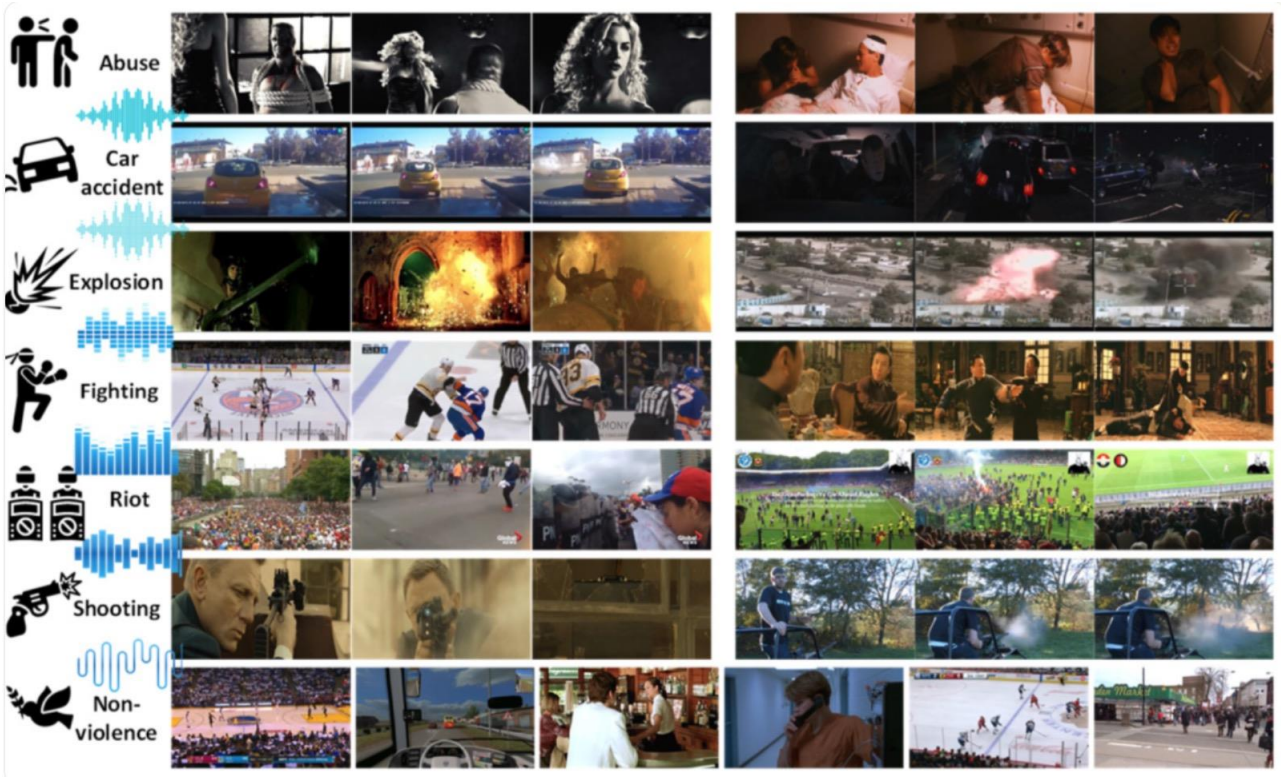
$$\tilde{a}_i = \sum_{k \in \mathbf{K}_i} a_k \cdot \frac{e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k) \rangle}}{\sum_{k \in \mathbf{K}_i} e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k) \rangle}}$$

Experiments

Datasets

- **UCF-Crime** is a large-scale dataset that is composed of 1900 long untrimmed real-world surveillance videos
- **XD-Violence** is another large-scale dataset for violence detection, comprising 4754 untrimmed videos that are collected from both movies and YouTube





Experiments

Ablation study

- When the Image-Text Caption Cleaning component is omitted, the performance degrades by –3.8%
- Skipping temporal summary and relying only on cleaned captions with refinement leads to a –7.58% drop
- Using only temporal summary anomaly scores on cleaned captions, without aggregating similar frames, leads to a –7.49% drop

Image-Text Caption Cleaning	LLM-based Anomaly Scoring	Video-Text Score Refinement	AUC (%)
✗	✓	✓	76.48
✓	✗	✓	72.70
✓	✓	✗	72.79
✓	✓	✓	**80.28**

Table 3. Results of LAVAD variants w/o each proposed component on the UCF-Crime Dataset.

Experiments

Ablation study

- Investigate the impact of different priors in the context prompt P_c
- Incorporating both priors does not further boost the AUC
- More stringent context might limit the detection of a wider range of anomalies

ANOMALY PRIOR	IMPERSONATION	AUC (%)
✗	✗	79.32
✓	✗	79.38
✗	✓	**80.28**
✓	✓	79.77

Table 4. Results of LAVAD on UCF-Crime with different priors in the context prompt when querying the LLM for anomaly scores.

BASE PROMPT: "How would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities?"

ANOMALY PRIOR: "or potentially criminal activities"

IMPERSONATION: "If you were a law enforcement agency,"

Experiments

Ablation study

- Investigate how the VAD performance changes in relation to the number of semantically similar temporal summaries
- AUC metric consistently increases as K increases
- Considering similar frames improves the accuracy of VAD

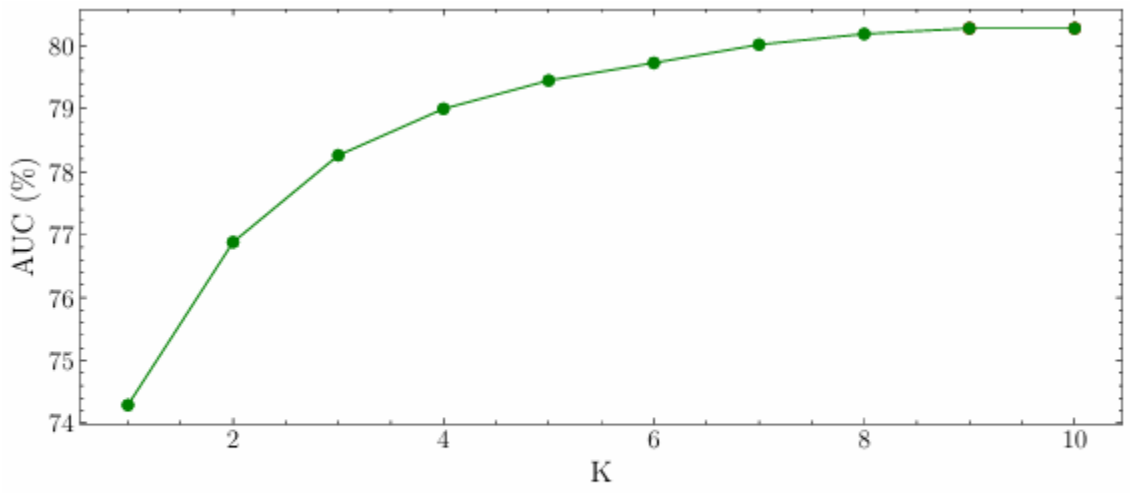


Figure 6. Results of LAVAD on UCF-Crime over the number of K semantically similar frames used for anomaly score refinement.

Experiments

Comparison with state of the art

- Notably, LAVAD without any training demonstrates superior performance compared to both the one-class and unsupervised baselines

METHOD	BACKBONE	AUC(%)
SULTANI *et al.* [24]	C3D-RGB	75.41
SULTANI *et al.* [24]	I3D-RGB	77.92
IBL [41]	C3D-RGB	78.66
GCL [40]	ResNext	79.84
GCN [42]	TSN-RGB	82.12
MIST [5]	I3D-RGB	82.30
WU *et al.*[36]	I3D-RGB	82.44
CLAWS [39]	C3D-RGB	83.03
RTFM [28]	VideoSwin-RGB	83.31
RTFM [28]	I3D-RGB	84.03
WU & LIU [35]	I3D-RGB	84.89
MSL [15]	I3D-RGB	85.30
MSL [15]	VideoSwin-RGB	85.62
S3R [34]	I3D-RGB	85.99
MGFN [2]	VideoSwin-RGB	86.67
MGFN [2]	I3D-RGB	86.98
SSRL [13]	I3D-RGB	87.43
CLIP-TSA [11]	ViT	87.58
SVM [24]	-	50.00
SSV [23]	-	58.50
BODS [33]	I3D-RGB	68.26
GODS [33]	I3D-RGB	70.46
GCL [40]	ResNext	74.20
TUR *et al.* [30]	ResNet	65.22
TUR *et al.* [31]	ResNet	66.85
DYANNET [27]	I3D	79.76
ZS CLIP [22]	ViT	53.16
ZS IMAGEBIND (IMAGE) [6]	ViT	53.65
ZS IMAGEBIND (VIDEO) [6]	ViT	55.78
LLAVA-1.5 [17]	ViT	72.84
LAVAD	ViT	**80.28**

Table 1. Comparison with state-of-the-art weakly-supervised, one-class, unsupervised and training-free methods on the UCF-Crime dataset. The best results among training-free methods are highlighted in bold.

METHOD	BACKBONE	AP(%)	AUC(%)
WU *et al.* [36]	C3D-RGB	67.19	-
WU *et al.* [36]	I3D-RGB	73.20	-
MSL [15]	C3D-RGB	75.53	-
WU AND LIU[35]	I3D-RGB	75.90	-
RTFM [28]	I3D-RGB	77.81	-
MSL [15]	I3D-RGB	78.28	-
MSL [15]	VideoSwin-RGB	78.58	-
S3R[34]	I3D-RGB	80.26	-
MGFN [2]	I3D-RGB	79.19	-
MGFN [2]	VideoSwin-RGB	80.11	-
HASAN *et al.* [8]	AE^{RGB}	-	50.32*
LU *et al.* [19]	Dictionary	-	53.56*
BODS [33]	I3D-RGB	-	57.32*
GODS[33]	I3D-RGB	-	61.56*
RAREANOM [26]	I3D-RGB	-	68.33*
ZS CLIP [22]	ViT	17.83	38.21
ZS IMAGEBIND (IMAGE) [6]	ViT	27.25	58.81
ZS IMAGEBIND (VIDEO) [6]	ViT	25.36	55.06
LLAVA-1.5 [17]	ViT	50.26	79.62
LAVAD	ViT	**62.01**	**85.36**

Table 2. Comparison with state-of-the-art weakly-supervised, one-class, unsupervised and training-free methods on the XD-Violence dataset. $*$ denotes results reported in [26]. The best results among training-free methods are highlighted in bold.

Experiments

Qualitative Results

- LAVAD correctly detect the anomaly scene in the three abnormal videos
- In the case of *Normal_Videos_722*, LAVAD consistently predicts a low anomaly score throughout the video

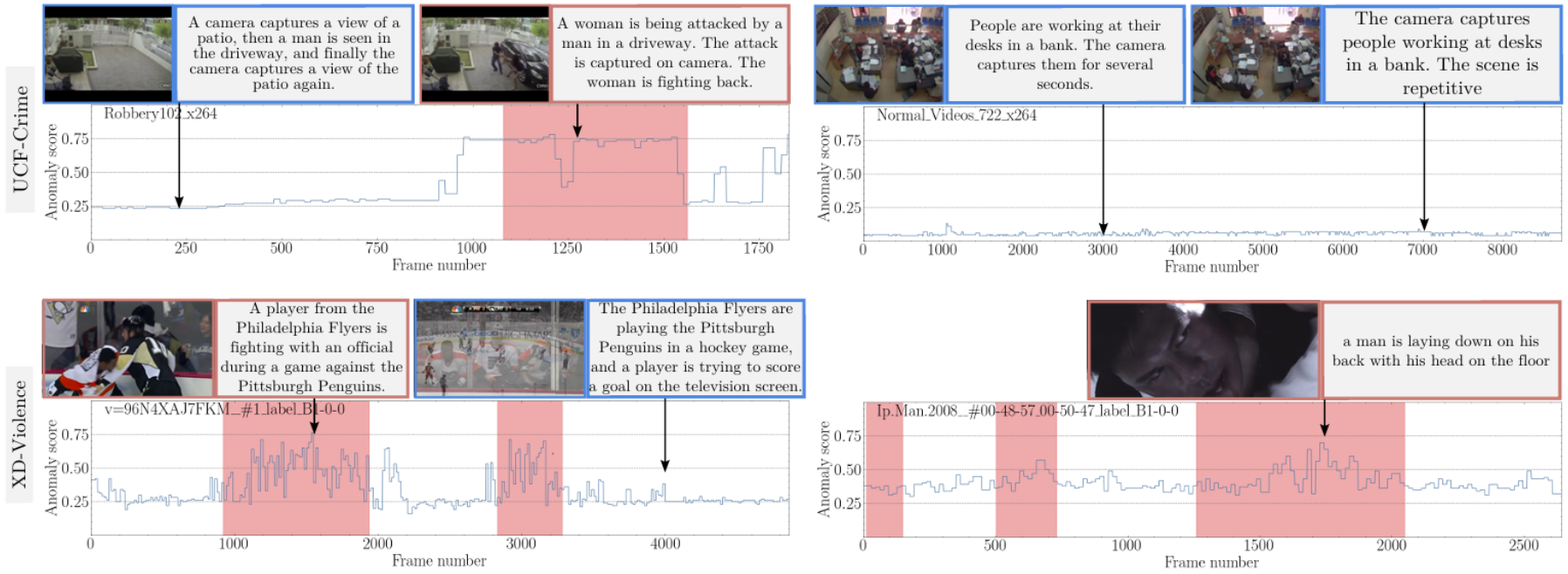
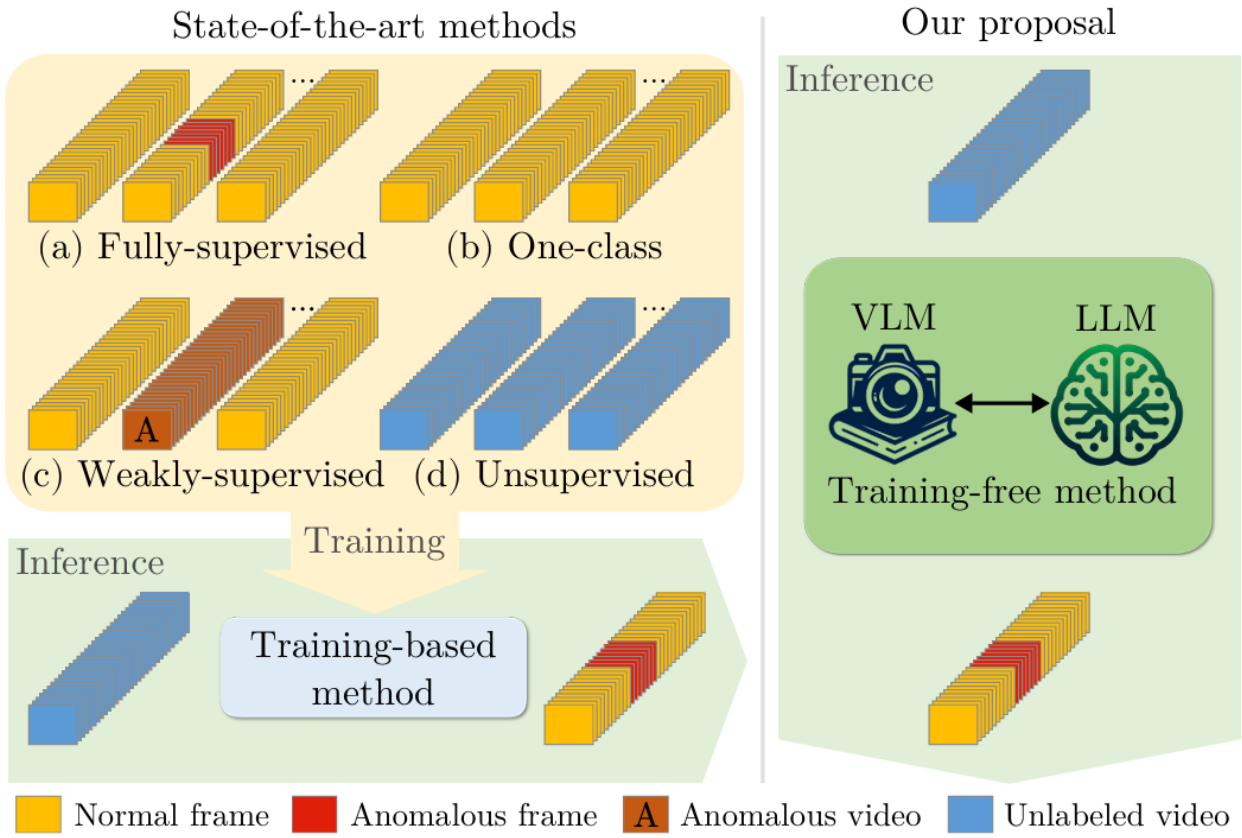


Figure 5. We showcase qualitative results obtained by LAVAD on four test videos, including two videos (top row) from UCF-Crime and two videos from XD-Violence (bottom row). For each video, we plot the anomaly score over frames computed by our method. We display some keyframes alongside their most aligned temporal summary (blue bounding boxes for normal frame predictions and red bounding boxes for abnormal frame predictions), illustrating the relevance among the predicted anomaly score, visual content, and description. Ground-truth anomalies are highlighted.

Conclusions

Language-based VAD

- Introduce LAVAD, a pioneering method to address training-free VAD
- Demonstrate superior performance compared to training methods in unsupervised and one-class setting
- Expected to significantly contribute to VAD, where data collection is difficult



Thank you