AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM

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

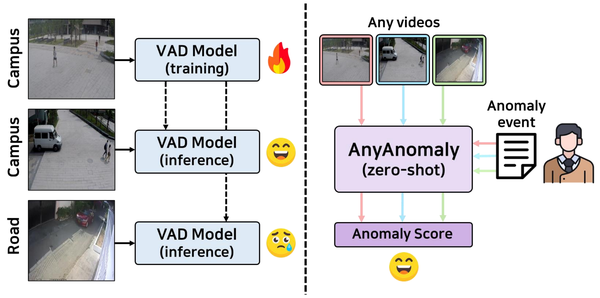
Data Engineering Lab, Yonsei University, Seoul, Republic of Korea

WACV 2026 TUCSON, AZ 3/6 - 3/10

MOTIVATION

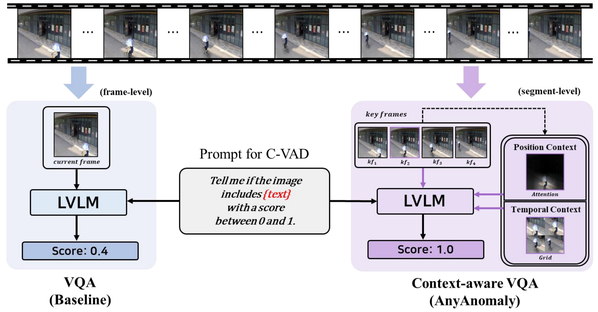
- **Video Anomaly Detection (VAD)** models typically learn normal patterns under specific scenarios, **limiting generalization to diverse environments**
- Adapting to new scenarios requires retraining or separate models
  → **increased computational cost and data collection**

What if we could detect anomalies dynamically using user-defined events, instead of relying on static normal patterns?



CONTRIBUTIONS

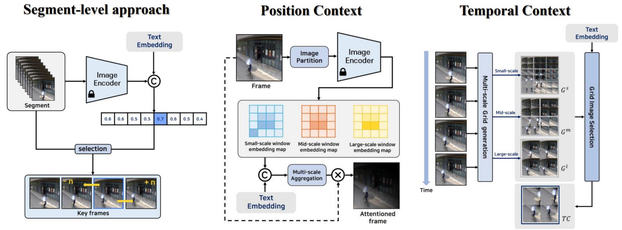
- **Customizable Video Anomaly Detection (C-VAD) : a zero-shot, text-driven VAD technique**
  - ✓ **User-defined text descriptions specify abnormal events; frames containing the described events are detected**
1. Propose **C-VAD** for anomaly detection across diverse environments
2. Develop **AnyAnomaly**, a context-aware VQA model for C-VAD
3. Construct **C-VAD datasets** for rigorous evaluation and validate the effectiveness of the proposed method



METHODS

- **Baseline**: perform frame-level VQA using an LVLM to estimate an anomaly score (return a value between 0 (no) and 1 (yes))
  (latency issue, limit object-level analysis and temporal reasoning)
- **AnyAnomaly**
  1. **Segment-level approach**: select key frames and perform VQA per segment
  2. **Context-aware VQA**: perform VQA using additional contexts that represent the frames
     Position Context emphasizes important spatial regions, while **Temporal Context** represents scene changes over time

Method Overview



- **Key frames Selection Module**: selects representative frames from a video segment by considering **temporal uniformity** and **alignment with the text query**
- **WinCLIP-based Attention**: captures important spatial regions using **multi-scale window embedding**
- **Grid Image Generation**: combines windows from the same spatial locations to form **grid image**

Anomaly Scoring

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.

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

EXPERIMENTS

Performance Comparison

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

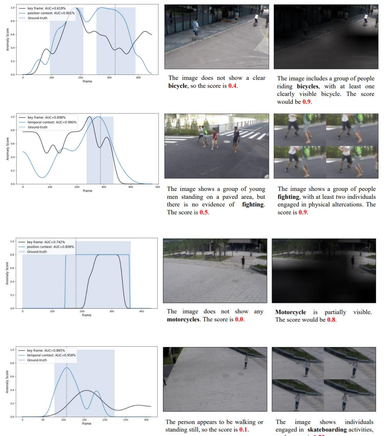
Category	Class	Baseline	+KSM	+KSM/PC	+KSM/TC	Proposed	Improvement (%)
Action	Throwing	78.44	80.13	**89.77**	82.40	**89.77**	+14.44
	Running	75.82	77.67	77.67	**77.90**	**77.90**	+2.74
	Dancing	85.65	72.28	76.64	**91.92**	**91.92**	+7.32
	Average	79.97	76.69	81.36	84.07	**86.53**	+8.2
Appearance	Too close	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

Method	Zero-shot	Ave	SbT	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[73]	✗	90.1	78.6	62.7	74.7
SLM[33]	✗	90.9	78.8	-	-
USTN-DSC[40]	✗	89.9	73.8	-	-
AnomalyRater[58]	✗	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	-
Accl-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**

Segment length	C-SbT	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**

Dataset	Method	Value	AUC
Ave	w/o context		81.4
	w/o tuning	1.0, 1.0, 1.0	84.4
	w/ tuning	0.6, 0.3, 0.1	**87.3**
SbT	w/o context		77.2
	w/o tuning	1.0, 1.0, 1.0	79.4
	w/ tuning	0.5, 0.3, 0.2	**79.7**
UB	w/o context		73.1
	w/o tuning	1.0, 1.0, 1.0	73.8
	w/ tuning	0.6, 0.1, 0.3	**74.5**
UCF	w/o context		77.8
	w/o tuning	1.0, 1.0, 1.0	80.5
	w/ tuning	0.6, 0.1, 0.3	**80.7**

Qualitative Evaluation



VAD in complex scenarios