AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM

Supplementary Material

A. Experiment Details

A.1. Dataset Details

VAD Dataset. We used the CUHK Avenue (Ave) [5], ShanghaiTech Campus (ShT) [6], UBnormal (UB) [1], and UCF-Crime (UCF) [8] datasets. Ave comprises of videos captured by a single camera on a university campus, containing five types of abnormal events; throwing paper, running, dancing, approaching the camera (Too close) and bicycle. ShT is a campus CCTV dataset that includes 13 different background scenes and 11 types of abnormal events; such as bicycles, cars, fighting, and jumping. UB is a synthetic dataset generated using the Cinema4D software, encompassing 29 diverse background scenes, including indoor environments, sidewalks, and *etc*. It provides a total of 22 abnormal events, including not only challenging-to-detect events such as smoking and stealing but also complex scenarios such as driving outside the lane and people-car accidents. UCF is a large-scale dataset containing 1,900 untrimmed real-world surveillance videos, covering 13 anomaly categories such as assault, burglary, explosion, and stealing.

C-VAD Dataset. We constructed the Customizable-ShT (C-ShT) and Customizable-Ave (C-Ave) datasets. C-ShT reorganizes the test data of ShT into 11 abnormal event types and assigns new labels to each type. For example, in the bicycle category, videos containing bicycles were assigned to positive, whereas all other videos were assigned to negative. The frame-level labels were set to 1 only for frames in which a bicycle appeared in the positive videos. C-Ave was constructed by reorganizing the test data of Ave into 5 abnormal event types, following the same labeling methodology as C-ShT.

A.2. Implementation Details

In a key experiment using the C-VAD datasets, we employed an efficient Chat-UniVi [3] 7B model, considering the balance between performance and speed. For the VAD dataset experiment, we utilized the effective MiniCPM-V [10] 8B model to achieve optimal performance and compared it with state-of-the-art (SOTA) models. At the time, both models represented a well-validated choice for context-aware VQA. Additional results with Qwen2.5-VL [2] are reported in Table S4 for further comparison. The CLIP model used for key frames selection and context generation was ViT-B/32. For context generation, we adopted large, middle, and small window sizes of (120,120), (80,80), and (48,48), respectively. For C-Ave and Ave, the large window size was set to (240,240). All the experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU.

Table S1. Comparison on prompt tuning

Prompt Tuning	C-ShT	C-Ave
Baseline (simple)	70.38	67.58
Baseline (+reasoning)	71.58	72.79
Baseline (+reasoning, consideration)	78.01	79.43
Proposed (simple)	79.29	74.01
Proposed (+reasoning)	79.79	82.09
Proposed (+reasoning, consideration)	**85.72**	**90.27**

A.3. Prompt Details

Figure S1 shows the detailed prompts used in the experiments. First, a reasoning prompt is designed to obtain the chain-of-thought [9] effect by requiring a simple reason along with the anomaly score. This helps to break down the problem step-by-step, guiding the model to resolve complex issues more systematically. For example, the question *"Does the image include jumping?* can be divided into two steps: *1. "Is there an object related to jumping (e.g., a person)?"* and *2. "Is the object performing a jumping action?"* This allows object-level image analysis, leading to more refined predictions. The consideration prompt encourages the assignment of a high score even when X is not central within the image. This prompt was introduced to address the issue where low scores are assigned simply because X exists but is not the central element. The effectiveness of this prompt tuning is compared and analyzed in Table S1.

The simple prompt instructs the LVLM to output only the anomaly score, while adding reasoning prompt the model to perform reasoning during the score calculation process, and applying consideration prompt encourages the model to focus on the given text. Experimental results showed that using both reasoning and consideration prompt achieved the best performance, suggesting that when the LVLM includes reasoning in the process, it produces more accurate results and can respond more precisely to user instructions through consideration prompt.

B. Additional Quantitative Evaluation

B.1. Segment length and FPS

Table S2 presents the performance comparison and FPS based on different segment lengths. The baseline segment length was set to 1. It was observed that deriving anomaly

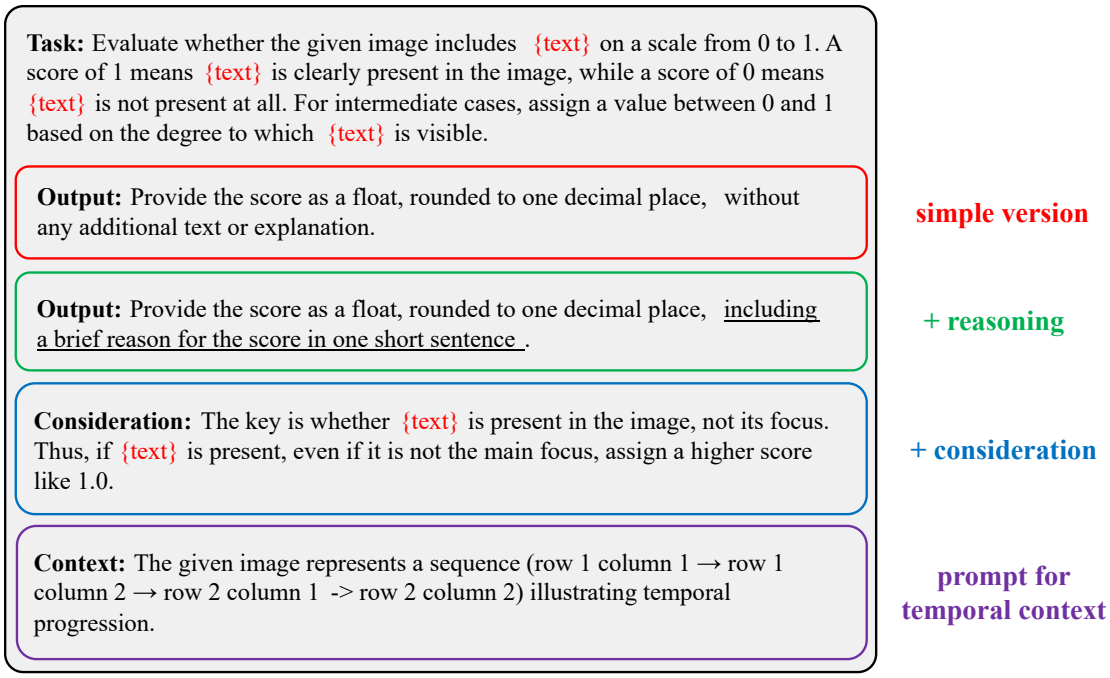


Figure S1. Prompt details. The content written in the simple version is not utilized when applying reasoning.

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 S3. Comparison of different methods on various datasets

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**
ShT	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**

scores at the segment level yields superior performance compared to the baseline, which relies on a single frame. The highest AUC performance was achieved when the segment length is set to 24, reaching 85.72% and 90.27% for C-ShT and C-Ave, respectively. However, excessively long segment length introduces irrelevant information into the temporal context, leading to a decrease in accuracy. Furthermore, performing VAD at the segment level resulted in a 594% improvement in the FPS compared with the baseline.

B.2. Hyperparameter Tuning

We tuned the three hyperparameters γ_1, γ_2, and γ_3 used for the final anomaly score calculation for each VAD dataset. Each hyperparameter controls the influence of the anomaly score derived from the frame, position, and temporal contexts. As shown in Table S3, the optimal hyperparameter values vary across datasets owing to differences in object sizes and abnormal events. Additionally, comparing w/o context, which does not utilize context information, and w/o tuning, where all hyperparameters were set to the same value, we observed performance improvements of 3.0%, 2.2%, 0.7%, and 2.7%, even without hyperparameter tuning. In contrast, the performance differences owing to hyperparameter tuning were 2.9%, 0.3%, 0.7% and 0.2%, respectively. This demonstrates the effectiveness of our proposed approach in utilizing context information in VAD and proves that it achieves a strong generalization performance even without hyperparameter tuning.

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[3]	Chat-UniVi-7B	77.5	85.7	78.3	90.3	6.67
MiniGPT-4[11]	LLaMA-2 Chat 7B	54.0	67.4	53.9	55.3	1.26
MiniCPM-V[10]	MiniCPM-Llama3-V-2_5 (8B)	87.7	**90.1**	86.3	**91.0**	1.36
LLAVA++[7]	LLaVA-Meta-Llama-3-8B-Instruct-FT	73.3	82.8	59.0	69.4	7.25
Qwen2.5-VL[2]	Qwen2.5-VL-3B-Instruct	89.0	90.2	78.0	87.0	11.18
Qwen2.5-VL*[2]	Qwen2.5-VL-3B-Instruct	88.6	90.2	78.3	88.1	34.78
Qwen2.5-VL*[2]	Qwen2.5-VL-7B-Instruct	93.0	**95.5**	86.9	**92.4**	24.08

B.3. Diverse LVLM Comparison

Table S4 presents the results for C-ShT and C-Ave when using various LVLMs. We evaluated the performances of four SOTA LVLMs: Chat-UniVi [3], MiniGPT-4 [11], MiniCPM-V [10], and LLAVA++ [7]. All experiments were conducted using the default settings, and 'Pre-trained' refers to the names of the pre-trained model weights. The experimental results demonstrate that incorporating the proposed context-aware VQA improves the performance of all LVLMs. Specifically, the use context-aware VQA leads to improvements ranging from 2.6% to 24.8%. Notably, even MiniCPM, which achieved the best performance without context-aware VQA, and showed additional improvements of 2.7% and 5.4% for C-ShT and C-Ave, respectively, when context-aware VQA was applied. This confirms that leveraging the proposed context-aware VQA is effective for C-VAD. Additionally, we observed that Chat-UniVi, with an FPS of 6.67, was the most efficient model, whereas MiniCPM-V achieved the highest performance on both datasets, scoring 90.1% and 91.0%, respectively. Therefore, as mentioned in Appendix A.1, Chat-UniVi was used for the C-VAD experiments and MiniCPM-V was used for the VAD dataset experiments.

B.4. Additional Experiments with vLLM

With recent advances in LLM inference libraries such as vLLM [4], latency issues can be largely mitigated. vLLM supports continuous batching and KV-caching, which substantially improve inference speed. Our method is not restricted to a specific LVLM and can be applied to various models supported by vLLM.

To evaluate efficiency, we conducted experiments by batching the key frame, PC, and TC during C-VAD. The experiments were conducted using Qwen2.5-VL [2], one of the representative models supported by vLLM. The results, presented in the lower part of Table S4, show that using vLLM maintained the AUC while achieving a 211% improvement in FPS (from 11.18 to 34.78). Furthermore, even with a 7B model, higher FPS was observed compared to other models. These findings suggest that AnyAnomaly has strong potential for real-time applications.

C. Additional Qualitative Evaluation

C.1. Further Analysis on Contexts

To provide a deeper understanding of the role of PC and TC, we present both qualitative analyses of their effectiveness and representative failure cases.

Figure S2 illustrates the improvements achieved by incorporating PC and TC. Without PC, the motorcycle appeared too small to be detected, resulting in an anomaly score of 0. With PC applied, however, the relevant region was emphasized, and the score increased to 0.8 despite the object's small size (top row). Likewise, without TC, the model misinterpreted skateboarding as walking and assigned a low anomaly score of 0.1. By integrating temporal information, the model captured motion cues such as positional changes and the enlarged appearance of the skateboard, correctly identifying skateboarding with a score of 0.75 (bottom row). These results indicate that our context-aware VQA is more effective than conventional VQA.

Nevertheless, Figure S3 demonstrates that PC and TC can also introduce errors under certain conditions. In the top row, before applying PC, the LVLM detected the abnormal action of throwing a bag with an anomaly score of 0.7. After applying PC, however, the model attended more to the person than the bag, making the action ambiguous and reducing the score to 0.3. In the bottom row, without TC, a normal walking scene yielded a score of 0, but after applying TC, walking was misinterpreted as a hard negative, raising the score to 0.4. While these failure cases exist, they provide valuable insights into the challenges of context-aware VQA and guide future directions for improvement.

Figure S2. Anomaly score comparison with context visualization (success cases)

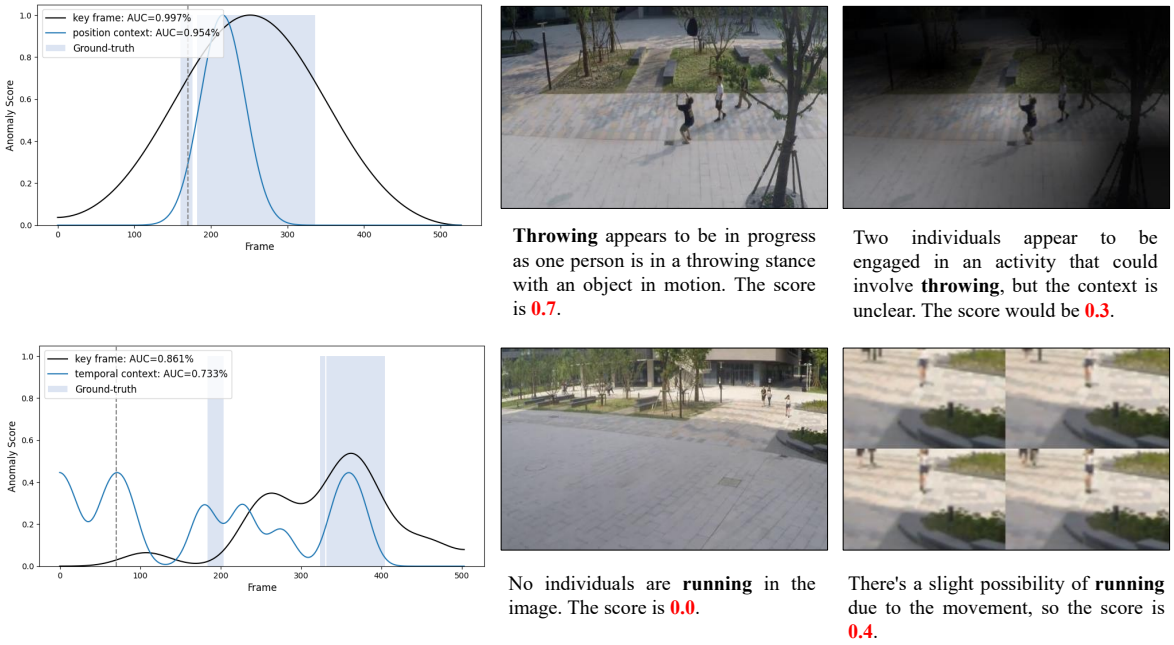
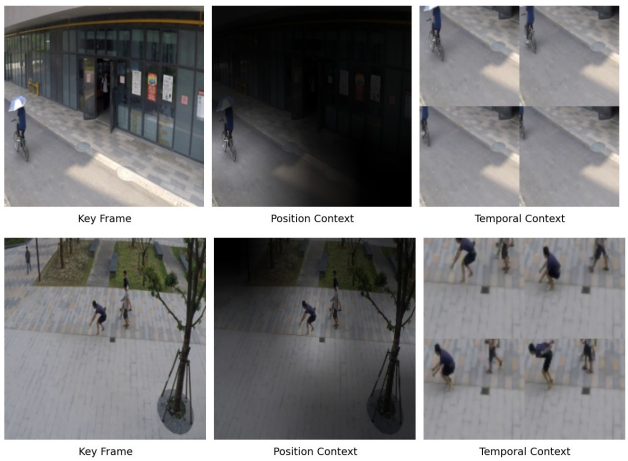


Figure S3. Anomaly score comparison with context visualization (failure cases)

C.2. Context Complementarity

In this section, we explain the complementarity between PC and TC in context-aware VQA. Figure S4 visualizes the key frame of a specific segment along with the images generated using WA and GIG for of PC and TC. We also present the results of a context-aware VQA that utilizes these contexts.

In the first row, when the text input was 'bicycle', PC



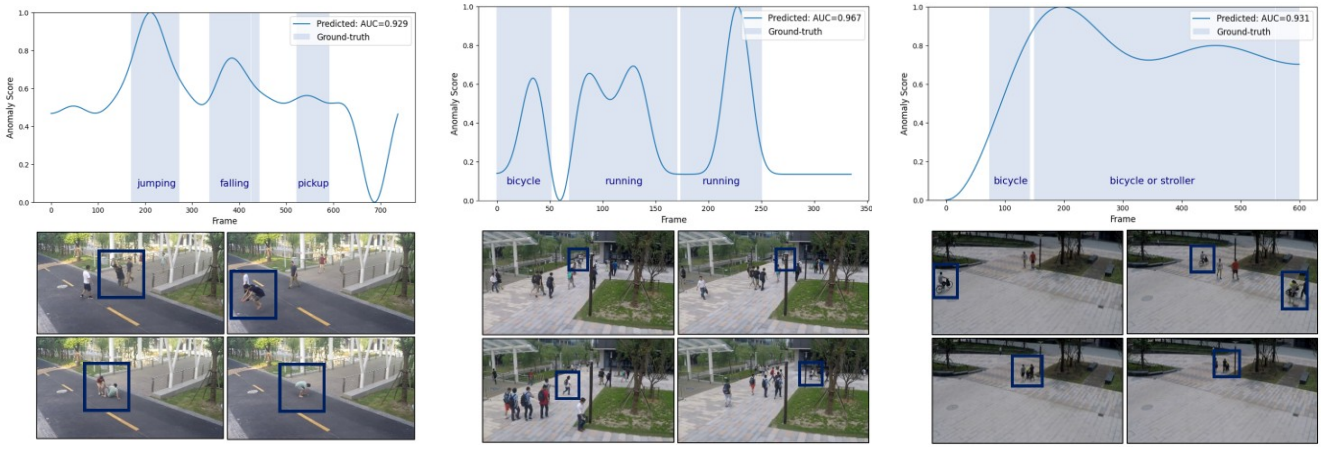
Position context: The image includes a bicycle. The score is **1.0**.

Temporal context: The score is **0.5**. The bicycle is only partially visible in different instances.

Position context: The image does not show any jumping. Assigning a score of **0.0**.

Temporal context: The image shows a person jumping in different instances, so the score would be **0.9**. The person is captured in various positions, including a jump.

Figure S4. Example of complementarity between position and temporal context. The first example highlights the importance of position context and the second example emphasizes the importance of temporal context.



(a) jumping-falling-pickup (b) bicycle-running (c) bicycle-stroller

Figure S5. Anomaly detection in diverse scenarios. Various abnormal events can emerge over time.

successfully identified the bicycle via WA, yielding a score of 1.0. However, the temporal context suffers from a cropping effect due to motion over time, resulting in a lower score of 0.5. In the second row, when the text input is 'jumping,' the attention result from WA fails to accurately locate the 'jumping' person. Additionally, because of the lack of temporal information, PC was unable to recognize the jumping action, resulting in a score of 0.0. In contrast, TC captured the entire jumping action over time, achieving a score of 0.9.

These results demonstrate that the proposed PC, which focuses on the object appearance, and TC, which leverages temporal information, are complementary. By integrating both approaches, we enable an effective generalization of the VAD.

C.3. Anomaly Detection in Diverse scenarios

Figure S5 visualizes the results of VAD performed on videos containing multiple abnormal classes. The captions in each figure indicate the abnormal classes used in the corresponding video. We input the user-defined abnormal keywords as text individually to obtain the scores, and assigned the highest score as the anomaly score for the corresponding segment. As shown in the visualization results, the proposed AnyAnomaly enables VAD across various types of abnormal events. This demonstrates that AnyAnomaly can be effectively utilized even when the user aims to simultaneously detect multiple abnormal types.

C.4. Anomaly Detection in Complex scenarios

Figure S6 presents the visualization results of AnyAnomaly on complex scenarios. 'Key Frame', 'Position Context',

and 'Temporal Context' visualize $\hat{k}$, PC, and TC, respectively. The text below each figure represents the LVLM output. These visualization results demonstrate that the proposed context-aware VQA, which utilizes PC and TC, is effective and contributes to improving VAD performance.

Additionally, in Figure S6d, we observe that the model can detect certain frames of "walking drunk" even without utilizing context information. This suggests that the strong visual reasoning capabilities of the LVLM enable VAD in complex scenarios. However, as shown in Figure S6a–S6c, relying solely on individual frames is insufficient for fully leveraging these reasoning abilities. Therefore, the proposed context-aware VQA approach is essential for effective VAD.

D. Discussion

D.1. Comparison with traditional VAD

Traditional VAD methods and our zero-shot C-VAD each have distinct strengths and limitations. Traditional VAD detects anomalies as deviations from learned normal patterns, requiring no prior knowledge of specific anomaly types and delivering strong performance within the trained environment. However, it often exhibits poor generalization to unseen environments and typically necessitates retraining. In contrast, C-VAD requires prior knowledge of anomaly types but removes the need for retraining or additional data collection even when the definition of "normal" varies across users or environments. This makes it a practical and cost-effective solution for real-world applications. We anticipate that, with continued advances in LVLM technology, the proposed C-VAD will become even more effective in the future.

D.2. Limitation

Efficiency is crucial in VAD; therefore, we adopted the most lightweight model among the SOTA LVLMs and employed a segment-level approach to reduce latency. However, when multiple abnormal events occur simultaneously, each event must be processed independently, which increases latency. In future work, we plan to improve the efficiency of C-VAD in handling multiple abnormal events concurrently.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3

[3] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1, 3

[4] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 3

[5] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 1

[6] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017. 1

[7] Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S. Khan. Llava++: Extending visual capabilities with llama-3 and phi-3, 2024. 3

[8] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1

[10] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1, 3

[11] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

(a) Anomaly event: jaywalking

Proposed (w/o context): The score would be **0.0**, as there is no evidence of individuals crossing the street.

Proposed: The image shows a man jaywalking across the street, so the score is **1.0**. This assessment is based on the clear depiction of a person crossing the street at a point other than a designated crosswalk.

(b) Anomaly event: driving outside lane

Proposed (w/o context): The image does not show any cars driving outside their lanes, so the score is **0.0**. There are no instances of driving outside the lanes in this particular image.

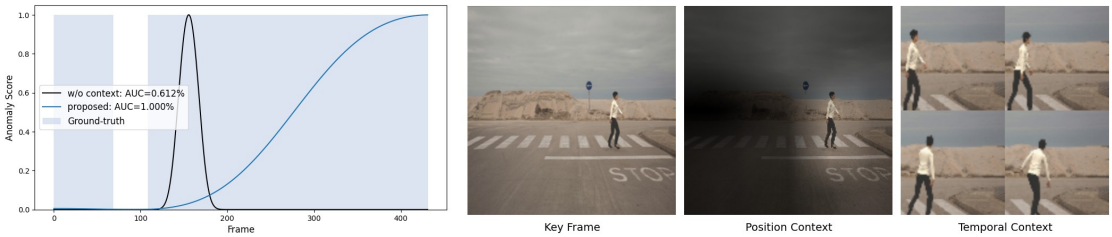
Proposed: The image shows a car driving outside the lane in multiple instances, with the highest score being **0.9**. The car is seen drifting into the next lane in some frames.

(c) Anomaly event: people and car accident

Proposed (w/o context): The score would be **0.0**. This is because there is no evidence of an accident involving people and a car in the visible parts of the image.

Proposed: The score would be **1.0**. This is because each frame in the sequence shows a person lying on the ground near a vehicle, which is indicative of an accident involving a person and a car.

(d) Anomaly event: walking drunk



Proposed (w/o context): The image does not include the "walking_drunk" condition, so the score is **0.0**.

Proposed: The image includes a man who appears to be walking while slightly drunk, so the score would be **0.8**.

Figure S6. Anomaly detection in complex scenarios. Results with and without the inclusion of context are presented.