Explainable AI for Real-Time Video Anomaly Anticipation

David C. Anastasiu*

Abstract

With computational power expanding on the edge and deep learning models now capable of real-time inference, it is time to rethink our approach to anomalies. Instead of waiting for anomalies to happen and then detecting them, why not predict them before they occur—and prevent them altogether? Imagine a system that can look at the data at time t and warn us that something might go wrong at $t + h$. If h is long enough, we can act—automatically or semi-automatically—to stop the anomaly in its tracks. Of course, that might mean changing some people's plans, and when the anomaly does not happen (because it was prevented), they might wonder why those changes were necessary. That is why these systems need to explain themselves—showing us, visually or descriptively, what they thought was about to go wrong. In this paper, we explore the challenges and opportunities of building real-time video anomaly anticipation systems and share a vision for how these tools could make a real-world impact.

1 What is the Blue Sky Idea?

Real-time event detection is essential for preventing potential catastrophes. Video surveillance systems, now ubiquitous in cities, workplaces, and retail environments, present a valuable opportunity to predict and prevent anomalies before they occur. This capability spans multiple domains, including intelligent transportation systems (ITS), online education, smart homes, and smart cities [17]. Video anomaly detection (VAD) involves identifying abnormal events from video sequences, such as detecting a traffic collision using freeway surveillance cameras. Thus far, efforts in VAD have focused primarily on accurately identifying anomalies and, to a lesser extent, reducing the delay between an anomaly and its detection. However, these approaches identify anomalies only after they occur, relying on features in the video that signal an abnormal event. With advancements in AI hardware and modeling, we now have the means to develop tools that can accurately predict future anomalies in real-time, anticipate these events, and offer clear explanations for their predictions. It is time to push the boundaries of VAD into the transformative domain of *explainable anomaly anticipation*.

2 Why is it a Blue Sky Idea? Why should the community ponder over it? Why now?

Ideal video anomaly detection is defined as the process of detecting and tracking abnormal events online and in real time [12]. Online video analysis methods are algorithms that process video streams incrementally, producing output as new frames arrive. In contrast, offline (or batch) methods require access to the entire input data before generating any output. While many state-of-the-art VAD methods achieve high detection accuracy, they are offline methods that rely on extracting complex visual features, making them unsuitable for real-time applications [24].

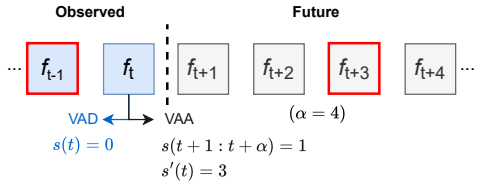

Figure 1: Video anomaly detection (VAD) vs. anticipation (VAA). f_t is the frame at time t, α is the anticipation time horizon, $s()$, $s'()$ are anomaly scores. Red squares denote ground truth anomalies.

A fundamental limitation of VAD is that it requires anomalies to occur before they can be detected. We propose a shift in focus toward **video anomaly anticipation** (VAA), which aims to predict whether an anomaly will occur within a future time horizon, such as 10 seconds [1]. As illustrated in Figure 1, VAD determines whether the current frame contains an anomaly ($s(t) =$ 1 or 0), while VAA predicts whether any of the next α frames will exhibit an anomaly ($s(t+1:t+\alpha) = 1$ or 0). Additionally, we propose to detect *when* the anomaly will occur, for instance predicting $s'(t) = 3$, meaning the anomaly is expected to happen three frames into the future, and its *confidence level*. In an analysis we performed on the Woven Traffic Safety dataset [10], we found that independent human annotators could predict an accident that was about to occur 2.76 seconds ahead

*Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA.

of time on average. With sufficiently accurate predictions and a long enough anticipation horizon, proactive measures could be deployed to prevent the anomaly.

A crucial research question here is, *What is the maximum anticipation horizon, α, that allows accurate anomaly prediction with minimal false positives and no false negatives for various anomaly types?*.

Existing methods for describing anomalies also have significant limitations. These approaches typically analyze the anomaly during or after it has occurred, relying on frames from the anomalous period to generate descriptions. Many state-of-the-art techniques leverage Vision Language Models (VLMs) [5, 21, 20] to construct a narrative based on identified scene facts. However, these methods focus on what has happened. Instead, we propose describing what might happen if the predicted anomaly were to occur.

Another key research question here is, *Are there sufficient clues in the scene at time t to accurately describe the scenario at future time $t + \beta$ when the predicted anomaly might occur ($\beta = 3$ in our example)?*

Now is the ideal time to address VAA due to advancements in AI modeling, edge computing, and video surveillance technology, which enable real-time analysis and prediction. The widespread deployment of video systems across industries, coupled with the growing need for proactive safety solutions, makes it feasible to develop methods that can prevent accidents and reduce risks by anticipating anomalies before they occur.

3 Does the Blue Sky Idea challenge our current set of assumptions or does it take a bold approach to solve a wicked problem?

The Blue Sky Idea of VAA takes a bold approach to solve a wicked problem. It challenges the current assumption that anomaly detection should occur only after the event has happened, by shifting the focus to predicting and preventing anomalies in real time. This requires rethinking traditional methods of video analysis, expanding the role of AI and edge computing, and addressing complex challenges like explainability, inference efficiency, and real-world deployment. The idea pushes the boundaries of existing technologies to not only detect but anticipate and intervene before catastrophic events occur.

The need for VAA solutions is most pressing in scenarios where predicting the future with sufficient lead time could allow for interventions to prevent anomalies. However, not all anomalies lend themselves to this approach. For instance, a catastrophic boiler failure may not be predictable unless clear warning signs, such as visible stress fractures, are present beforehand. Given the vast range of possible anomalies, efforts should prioritize high-reward scenarios where real-time anticipation could feasibly lead to prevention. One prominent example involves **motion-based anomalies**, such as a vehicle veering onto a sidewalk, a traffic accident, or a forklift colliding with shelving in a warehouse.

4 What are the challenges?

Several challenges, described below, must be overcome to unlock the full potential of VAA methods.

4.1 VAA Datasets Current general VAD datasets, such as UCF-Crime and XD-Violence, primarily support weakly- or semi-supervised VAD models, while datasets like Iowa DOT [14], CADP, and SUTD-TrafficQA focus on vehicle accident detection or anticipation, often using dashcam videos instead of surveillance inputs. These datasets are limited by their narrow representation of anomaly types and lack of diverse recording conditions (e.g., variations in time, weather, and season).

A key research question is *whether VAD datasets can be used to train and evaluate models that solve the VAA problem.* It may be possible to transform existing VAD datasets into VAA datasets. For instance, datasets like NWPU Campus [1] include anomaly timestamps that could support VAA research but lack explicit anticipation labels, while WTS [10] is limited to pedestrian-vehicle accidents and lacks precise anomaly start annotations.

Another avenue of research would be leveraging recent advances in VLMs and CGI simulations to directly generate VAA datasets. Synthetic data generation offers a complementary approach, addressing the limitations of real-world datasets and enabling the creation of rare or hazardous scenarios. Technologies such as NVIDIA Omniverse Replicator [16] allow for generating diverse, physically accurate data for training DNN models. These virtual environments can simulate rare anomalies, such as traffic or warehouse accidents, that are impractical to collect in real-world settings. A key research question is *whether models trained on synthetic data can generalize effectively to real-world scenarios.*

4.2 VAA Models Early VAD methods relied on extracting features from previous frames to model normality or employed reconstruction- and prediction-based approaches, using autoencoders or frame prediction to identify anomalies via reconstruction or prediction errors [15]. Some methods incorporated optical flow to enhance frame prediction [11, 22], while Cao et al. [1] extended these to VAA by estimating future-frame prediction errors.

The key research question here is *whether there is enough signal in the data before the anomaly has occurred to anticipate it.* In the case of motion-based anticipation models, this may be possible through accurate prediction of future object motion tracks, similar to some works in traffic accident anticipation that use LSTMs, GRUs, and GNNs to determine future collisions [9]. *Is it possible to extend these models to solve the general anomaly anticipation problem?*

4.3 Real-Time VAA To enable real-time prevention of anomalies, VAA models must operate at the edge, close to the cameras capturing the scenarios. It is essential to explore various methods that enhance inference efficiency while maintaining model effectiveness. One such approach is the quantization of weights and activations [18], which involves reducing model weights to lower numerical precision. This technique significantly decreases model latency and size with minimal impact on accuracy. Achieving extreme low-precision quantization (e.g., fewer than 3 bits per parameter) often requires data-dependent approaches [7], which leverage a small calibration dataset to determine optimal quantization parameters. Recent advances in LLM quantization [13] have pushed these limits further, restricting weights to +1, 0, and -1. This setup is particularly well-suited for specialized hardware, such as Field Programmable Gate Arrays (FPGAs), where inference execution can be highly optimized.

Numerous techniques have been proposed to enhance DNN inference performance on FPGAs, including approximate computing, stochastic computing, pipelining, quantization, and efficient mapping of binarized neural networks to FPGA hardware [8]. Recent research has also emphasized FPGA-based inference for resource-constrained scenarios, focusing on applications such as LLMs [3], multi-modal foundation models, and Mixture of Experts (MoEs), including vision-specific MoEs [6]. Once accurate anticipation models proposed in Section 4.2 have been developed, the key research question here is, *What are the best ways to improve their efficiency for optimum real-time inference?*

4.4 VAA Prevention Once an anomaly is anticipated with sufficient confidence, immediate action must be taken to prevent it. Alerts can be delivered to relevant actors through visual or audio signals. For instance, visual alerts such as flashing lights or highlighted warnings can be displayed on monitors, while audio alerts like alarms or instructions can guide operators to take corrective actions. In environments like warehouses or transportation systems, these alerts can be crucial for preventing accidents.

In automated environments, such as intelligent transportation systems (ITS) or autonomous vehicles, vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I) communication can enable coordinated responses. For example, if an accident is predicted, nearby vehicles can receive alerts to adjust speed or change trajectory to avoid collisions. In extreme cases, automated systems could intervene directly by triggering braking or rerouting traffic to prevent the anomaly from occurring.

The challenge lies in ensuring these alerts and interventions are both timely and accurate, minimizing risks and maximizing safety. Real-time, context-aware prevention is key to avoiding incidents while maintaining the safety of all involved. As the best prevention techniques are likely to depend on the type of anticipated anomaly, a key research question here is *how to choose the most effective means of interacting with the actors in the scene to effectively prevent the anomaly.*

4.5 Explainable VAA While some methods can detect anomalies in video data, they often operate as black-box systems without offering explanations. Explainability is defined as "the capacity to clarify or provide straightforward meaning to humans in easily understandable terms" [23]. Current approaches explain detected anomalies at a high level, such as anomaly localization using attention-based models [2], action recognition through reasoning-based models [4], or relying on intrinsically interpretable methods [19]. Prediction-based methods that forecast frames or flow maps in video clips may also be applicable to the VAA problem.

Recent anomaly description methods [5, 21, 20] use LLMs and VLMs like LLaVA, Qwen-VL, and Video-LLaVA. These models process several video frames from the anomaly, encoded by a VLM encoder such as CLIP ViT-L/14, and generate textual descriptions of the anomaly scenario. Some methods enhance these descriptions through an LLM. A key research question here is *whether the signals used to effectively anticipate an anomaly are sufficient to also explain it.*

5 What will success look like?

Although our understanding of anomaly anticipation is still in its early stages, the potential benefits of this technology are profound. Success in video anomaly anticipation will be characterized by the development of real-time, explainable systems capable of accurately predicting anomalies before they occur, allowing for proactive interventions. These systems will operate with minimal latency, ensuring that predictions are made with enough lead time to prevent incidents. Additionally, success will involve creating datasets and AI models that can generalize across different environments and scenarios, as

well as implementing robust alerting mechanisms that effectively communicate predictions to relevant actors for timely action. Ultimately, the ability to prevent accidents and reduce risks in various domains, such as transportation and smart cities, will demonstrate the transformative impact of this technology.

References

[1] Cao, C., Lu, Y., Wang, P., and Zhang, Y. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 20392–20401.

[2] Cao, Y., and Wu, J. Tobias: A random cnn sees objects. *IEEE Transactions on Pattern Analysis & Machine Intelligence 46*, 02 (feb 2024), 1290–1304.

[3] Chee, J., Cai, Y., Kuleshov, V., and Sa, C. D. Quip: 2-bit quantization of large language models with guarantees, 2024.

[4] Doshi, K., and Yilmaz, Y. Towards interpretable video anomaly detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2023), pp. 2654–2663.

[5] Duan, Z., Cheng, H., Duo, X., Wu, X., Zhang, X., Xi, Y., and Xie, Z. Cityllava: Efficient fine-tuning for vlms in city scenario. In *CVPR Workshop* (Seattle, WA, USA, 2024).

[6] Frantar, E., and Alistarh, D. Qmoe: Practical sub-1-bit compression of trillion-parameter models, 2023.

[7] Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., and Soudry, D. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning* (18–24 Jul 2021), M. Meila and T. Zhang, Eds., vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 4466–4475.

[8] Ji, M., Al-Ars, Z., Hofstee, P., Chang, Y., and Zhang, B. Fpqnet: Fully pipelined and quantized cnn for ultra-low latency image classification on fpgas using opencapi. *Electronics 12*, 19 (2023).

[9] Karim, M. M., Yin, Z., and Qin, R. An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos. *IEEE Transactions on Intelligent Vehicles 9*, 1 (2024), 1792–1803.

[10] Kong, Q., Kawana, Y., Saini, R., Kumar, A., Pan, J., Gu, T., Ozao, Y., Opra, B., Anastasiu, D. C., Sato, Y., and Kobori, N. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding, 2024.

[11] Li, J., Huang, Q., Du, Y., Zhen, X., Chen, S., and Shao, L. Variational abnormal behavior detection with motion consistency. *IEEE Transactions on Image Processing 31* (2022), 275–286.

[12] Lin, H., Deng, J. D., Woodford, B. J., and Shahi, A. Online weighted clustering for real-time abnormal event detection in video surveillance. In *Proceedings of the 24th ACM International Conference on Multimedia* (New York, NY, USA, 2016), MM '16, Association for Computing Machinery, p. 536–540.

[13] Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., and Wei, F. The era of 1-bit llms: All large language models are in 1.58 bits, 2024.

[14] Naphade, M., Chang, M.-C., Sharma, A., Anastasiu, D. C., Jagarlamudi, V., Chakraborty, P., Huang, T., Wang, S., Liu, M.-Y., Chellappa, R., Hwang, J.-N., and Lyu, S. The 2018 nvidia ai city challenge. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (July 2018), vol. 1 of *CVPRW'18*, pp. 53–60.

[15] Nguyen, T. N., and Meunier, J. Anomaly detection in video sequence with appearance-motion correspondence. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 1273–1283.

[16] NVIDIA. Replicator. `https://docs.omniverse.nvidia.com/extensions/latest/ext_replicator.html`, 2023. [Online; accessed 17-July-2024].

[17] Ramachandra, B., Jones, M. J., and Vatsavai, R. R. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 5 (2022), 2293–2312.

[18] Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. Q-BERT: hessian based ultra low precision quantization of BERT. *CoRR abs/1909.05840* (2019).

[19] Szymanowicz, S., Charles, J., and Cipolla, R. Discrete neural representations for explainable anomaly detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), pp. 1506–1514.

[20] To, A. T., Tran, N. M., Ho, T.-B., Ha, T.-L., Nguyen, Q. T., Luong, C. H., Cao, T.-D., and Tran, M.-T. Multi-perspective traffic video description model with fine-grained refinement approach. In *CVPR Workshop* (Seattle, WA, USA, 2024).

[21] Trinh, K. X., Nguyen, N. K., Ngo, B. H., Dinh, V. X., An, H. M., and Dinh, V. Divide and conquer boosting for enhanced traffic safety description and analysis with large vision language model. In *CVPR Workshop* (Seattle, WA, USA, 2024).

[22] Vats, A., and Anastasiu, D. C. Key point-based driver activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2022), pp. 3274–3281.

[23] Wang, Y., Guo, D., Li, S., Camps, O., and Fu, Y. Explainable anomaly detection in images and videos: A survey, 2024.

[24] Zhang, Y., Song, J., Jiang, Y., and Li, H. Online video anomaly detection. *Sensors 23*, 17 (2023).