ATOMS TO EVENTS: CATEGORICAL EVIDENCE COMPOSITION FOR VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly detection (VAD) seeks to identify events that deviate from learned normality. Current Vision–Language Models (VLMs) face significant challenges: anomalies are rare, labels are weak, and visual appearance varies drastically. Mainstream VLMs directly map visual features to events, they overfit to intermediate incidental cues which are present during training and generalize poorly. To address this issue, we propose a categorical view of anomaly understanding. Firstly, an Unsupervised Anomalous Period Detector (UAPD) is proposed to identify abnormal periods. Next, a Category-based Atom Miner (CAM) is proposed to map visual features to learned atoms in video segments, and learn the roles of atoms. In inference, CAM provides role-aware indications to VLM which maps meaningful atoms and visual features to event predictions. This framework harnesses meaningful evidence and preserves the generalization capacity of VLMs. Extensive experiments and ablations show consistent gains over strong vision-only and fine-tuned VLM baselines.

1 INTRODUCTION

Video anomaly detection (VAD) seeks to localize the time spans in long videos where scenes deviate from regular patterns—e.g., violence, accidents, explosions, or other unexpected events. The problem is challenging because anomalies are rare in contrast to the massive scale of surveillance streams, visual conditions vary substantially across cameras and over time due to viewpoint, lighting and other factors, and labels are typically video-level or even abscent. Existing approaches have explored vision-only strategies that rely purely on visual features: prediction-based models that forecast future sequences and flag deviations, reconstruction-based models that assume anomalies reconstruct poorly, representations combining multiple feature types or Multiple Instance Learning (MIL)-based approaches Li et al. (2022); Georgescu et al. (2021); Park et al. (2020); Noghre et al. (2024); Yang et al. (2023); Liu et al. (2021); Huang et al. (2025); Georgescu et al. (2021); Cho et al. (2023); Guo et al. (2023); Liu et al. (2022). Although these approaches are effective, some of them cannot perform well under domain shifts, and may mistake irrelevant variations for true anomalies.

To improve interpretability and zero-shot generalization, recent works leverage Vision–Language Models (VLMs). They use captions, prompts, or pseudo-labels to score abnormal semantics within sliding windows Cao et al. (2024); Chen et al. (2024); Yang et al. (2024a); Micorek et al. (2024); Zhu & Pang (2024); Li et al. (2024); Yang et al. (2024b); Tang et al. (2025). Training-free approaches Zanella et al. (2024) directly caption frames with VLM and summarize how captions change over time. Guidance-driven methods focus VLMs on anomaly evidences by leveraging verbalized questions Ye et al. (2025) or sampling the most suspicious snippets Zhang et al. (2025). These methods are interpretable. However, VLM-based detectors often encode an event as one entangled concept, letting nuisance factors, such as viewpoints and backgrounds, swamp model representations. As a result, the visual features of events are not well distinguished.

To stably attend VLMs to anomaly-relevant semantics without being distracted by variations, challenges lie in the set-theoretic fashion of VLMs. They learn a direct mapping from visual features to event labels by accumulating "relevant" cues, a video is treated as a set of features whose memberships vote for the label. Such membership often admits incidental cues and spurious contexts. We find that meaningful features can be divided into directly relational, inherently relational and counter evidences. In this regard, we propose a categorical, role-aware view: we treat atoms as objects with

roles in events — Direct(support), Indirect (synergy), and Counter(inhibition). The "visual features to events" mapping in set-theoretic VLMs changes to "visual features to atoms to events". The cues without valid roles have no effect. Robust reasoning and auditable explanations are yielded.

REFERENCES

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*, 2024.

Jiankang Chen, Tong Zhang, Wei-Shi Zheng, and Ruixuan Wang. Tagfog: Textual anchor guidance and fake outlier generation for visual out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1100–1109, 2024.

MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12137–12146, 2023.

Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12742–12752. IEEE, 2021.

Chongye Guo, Hongbo Wang, Yingjie Xia, and Guorui Feng. Learning appearance-motion synergy via memory-guided event prediction for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Yuzhi Huang, Chenxin Li, Haitao Zhang, Zixu Lin, Yunlong Lin, Hengyu Liu, Wuyang Li, Xinyu Liu, Jiechao Gao, Yue Huang, et al. Track any anomalous object: A granular video anomaly detection pipeline. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8689–8699, 2025.

Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24. AAAI Press, 2022.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16838–16848, 2024.

Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17081–17092. IEEE, 2022.

Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13588–13597, 2021.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18868–18877, 2024.

Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 995–1004, 2024.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14372–14381. IEEE, 2020.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Yingcong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2025.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. *arXiv preprint arXiv:2407.10299*, 2024a.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14592–14601, 2023.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18899–18908, 2024b.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8679–8688, 2025.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13843–13853, 2025.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17826–17836, 2024.