VADER: Towards Causal Video Anomaly Understanding with Relation-Aware Large Language Models

Supplementary Material

Contents

A. Illustrative Example of Motivation

To better illustrate the motivation behind our work, Figure 1 presents a real-world anomalous event where a dog suddenly attacks a boy walking on the roadside. This example demonstrates the challenges in understanding not only what happens in the scene, but also how it occurs, which requires modeling both causal relationships and dynamic object interactions.

In this case, different methods generate diverse outputs when asked to describe the anomalous segment:

CUVA [1] focuses on basic visual elements but produces a description that does not align with the actual event. While certain textual metrics such as ROUGE are relatively high, the description lacks factual accuracy and fails to capture the core abnormal interaction.

Holmes-VAU [8] includes some relevant context, such as the presence of danger and a dog, but its narrative remains ambiguous. The description does not clearly reflect the cause-effect chain or the temporal progression of the event, limiting interpretability.

VADER (Ours) explicitly models the interactions and causal dynamics between objects, resulting in a coherent description that clearly states who was involved and how the anomaly unfolded. This produces outputs that align well with human judgments, as shown by higher semantic-level and human-evaluation metrics (e.g., BLEURT, UniEval, and mmEval).

The quantitative metrics and qualitative feedback below each output further highlight the importance of deeper reasoning. While lexical overlap metrics alone may not fully capture factual correctness, human-aligned evaluations reflect the clarity and interpretability of the descriptions. VADER achieves the highest ratings by providing a precise and logically structured explanation of the anomalous event.

B. Ablation Study on Sample Mining Hyperparameters

We perform an ablation study on the Gaussian smoothing parameter (σ) and the peak selection threshold (top-k percentile) in our weakly-supervised sample mining pipeline. As shown in Table 1, we report AUC scores on the test set for different combinations of σ and top-k.

Our results reveal two main findings. First, with $\sigma = 2.0$, a top-k percentile of 5% yields the best AUC (72.12). Increasing the threshold to 7% slightly reduces performance, likely due to the inclusion of less informative (noisy) samples. Conversely, a stricter threshold of 3% also degrades performance, suggesting that overly selective sampling yields too few positives for effective training. Second, across all thresholds, $\sigma = 2.0$ consistently outperforms both smaller ($\sigma = 1.0$) and larger ($\sigma = 3.0$) values. We attribute this to improved noise filtering at $\sigma = 2.0$ without excessive smoothing that would obscure salient events.

In summary, $\sigma = 2.0$ and top-$k = 5\%$ strike the best balance between sample quality and quantity, and are used in all subsequent experiments.

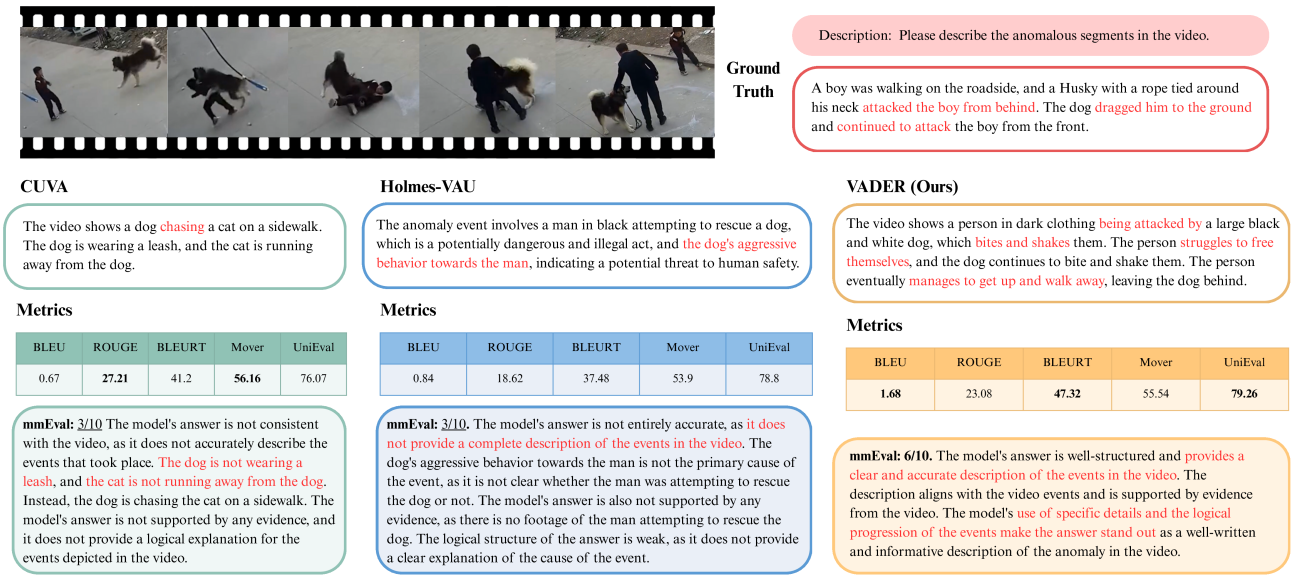

Figure 1. **Illustrative example showing the need for causal and relational modeling.** Given the same video of a dog attacking a boy, CUVA [1] and Holmes-VAU [8] produce incorrect or incomplete descriptions, while VADER captures key interactions and event progression, resulting in accurate and coherent descriptions with higher human-aligned evaluation scores.

Smooth Sigma (σ)	Top-k Percentile (%)		
	3.0%	**5.0%**	**7.0%**
1.0	69.53	70.18	68.91
2.0	70.85	**72.12**	71.64
3.0	68.17	69.25	68.55

Table 1. **Ablation study on sample mining hyperparameters.** Test set AUC (%) for different combinations of Gaussian smoothing parameter (σ) and top-k percentile thresholds used in peak selection. $\sigma = 2.0$ and top-$k = 5\%$ achieve the best performance, highlighting the importance of balancing sample quality and quantity in our weakly-supervised mining strategy.

C. Failure Cases

Figure 2 presents three examples highlighting VADER's tendency to favor visually dynamic events. While VADER accurately captures prominent high-motion activities, it often overlooks the underlying causes of anomalies or neglects subtle contextual cues, such as unattended objects or gradual environmental changes. For instance, in the middle example, VADER correctly identifies the car collision but misses the key causal factor, the vehicle ignoring the traffic signal, resulting in an incomplete explanation of why the anomaly occurred.

Figure 3 presents three examples highlighting VADER's limitation to object-centric reasoning. While VADER accurately captures localized actions and pairwise object interactions, it often fails to represent the broader scene context or collective behaviors. For instance, in the left example, VADER correctly describes individual pedestrians entering and exiting the subway but misses the group of people gathered to watch a street performance, resulting in an incomplete understanding of the true anomaly.

D. Potential Mitigations for Limitations

To address the three primary limitations discussed in Sec. 5, we outline potential mitigation strategies. These strategies aim to enhance VADER's robustness, reduce bias, and broaden its capacity to handle diverse anomaly scenarios.

D.1. Mitigating Dependency on Upstream Modules

VADER relies on the performance of upstream modules, which may cause errors to propagate and impact the final reasoning results. To reduce this dependency and improve the reliability of the pipeline, we consider the following approaches to strengthen upstream components and improve their alignment with downstream tasks:

End-to-End Joint Training. Transform the current pipeline into a jointly trainable framework, allowing gradients from the LLM to update the upstream modules so they can better align with the final reasoning objectives.

Module Robustness Enhancement. Before end-to-end integration, each module can be strengthened individually by training on more diverse data and employing more advanced tracking algorithms, thereby reducing upstream errors and improving overall robustness

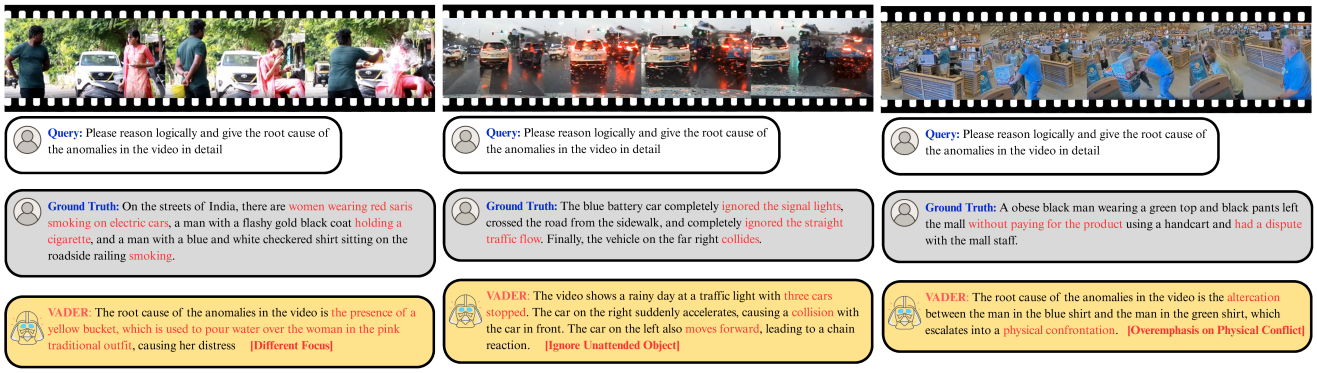

Figure 2. **Examples illustrating VADER's high-motion bias.** VADER tends to focus on visually prominent or dynamic actions, overlooking subtle or context-dependent cues. In the left example, it emphasizes pouring water while ignoring the public smoking. In the middle example, it identifies the car collision but misses the underlying cause, which is the vehicle running the traffic signal. In the right example, it overemphasizes the physical confrontation while neglecting the initial theft that triggered the anomaly.

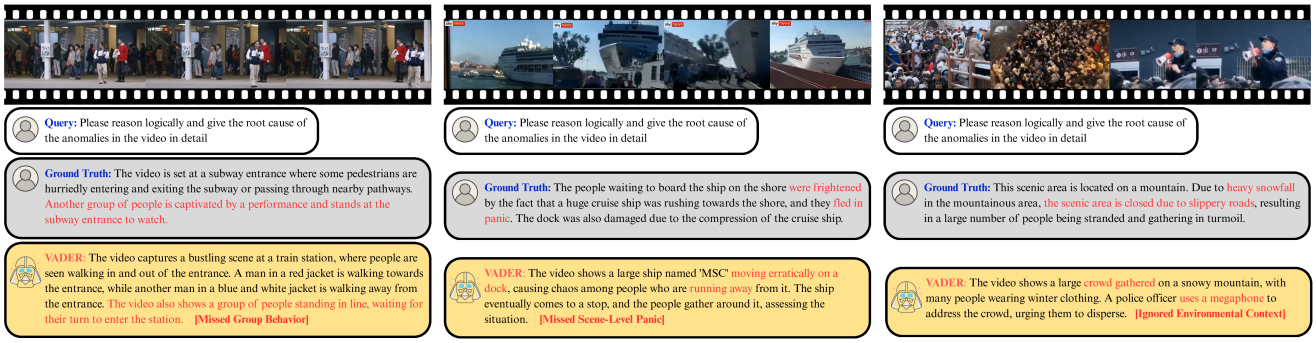

Figure 3. **Examples illustrating VADER's limitation to object-centric reasoning.** VADER focuses on localized pairwise interactions, overlooking group behaviors or scene-level environmental factors. In the left example, it describes individual pedestrian movements but misses the crowd gathered to watch a performance. In the middle example, it identifies the ship's movement but overlooks the collective panic it causes. In the right example, it focuses on the police officer and the crowd but ignores the environmental cause behind the stranded people.

D.2. Reducing Bias Toward High-Motion Events

VADER's reliance on relational volatility can lead to a bias toward anomalies involving strong motion while overlooking subtle anomalies. To address this imbalance, additional cues can be incorporated to complement volatility:

Object State Modeling. Track and interpret object states, such as transitions from "carried" to "stationary" without an owner, to capture static anomalies like abandoned objects.

Global Scene Context. Model typical activity flow within the scene and identify deviations, enabling the detection of subtle anomalies such as loitering or suspicious inactivity.

D.3. Moving Beyond Object-Centric Reasoning

The current design of VADER focuses on pairwise object interactions, limiting its ability to capture scene-level or group-level anomalies. To broaden the scope of anomaly understanding, the following extensions can be considered:

Global Scene Modeling. Add a global feature stream that directly encodes entire frames to detect anomalies not tied to specific objects, such as lighting changes or smoke.

Group-Level Reasoning. Identify and model groups of objects to capture collective behaviors, allowing detection of emergent events like crowd panic or mass movements.

E. Computational Efficiency

We provide a comparison of inference speed between VADER and NVILA [5] on the HAWK benchmark in Table 2. The table reports both the total inference time (in minutes) and frames per second (fps). Although VADER introduces additional modules for relational reasoning and context-aware sampling, it still operates at a practical speed for real-world applications.

Method	Total Time (min)	FPS
NVILA [5]	49.85	33.63
VADER	87.25	19.22

Table 2. Comparison of inference time and fps between NVILA [5] and VADER on the HAWK [7] benchmark.

F. Intermediate Visualization of Module Contributions

Figure 4 illustrates how each module incrementally improves the generated descriptions. The example shows a bustling scene at a train station. The Base Model generates a generic narrative without focusing on key details. CAES helps the model attend to important frames, capturing activities like walking and standing in line. CORE further models object interactions, enabling it to distinguish directional actions and explain relationships between people and their surroundings. This step-by-step progression demonstrates how each module contributes to improving both descriptive accuracy and interpretability.

Base Model: The video captures a bustling scene at a train station platform. The platform is sheltered by a roof supported by white pillars, and the ground is paved with light-colored tiles. A group of people is gathered near the entrance to the station

+CAES: The video features a bustling scene at a train station, where people are seen walking and standing in line, waiting for their train. The crowd is diverse, with individuals dressed in various outfits, such as a man in a red jacket and another in a white and blue jacket.

+CAES+CORE (**VADER**): The video captures a bustling scene at a train station, where people are seen walking in and out of the entrance. A man in a red jacket is walking towards the entrance, while another man in a blue and white jacket is walking away from the entrance. The video also shows a group of people standing in line, waiting for their turn to enter the station.

Figure 4. **Intermediate visualization** showing how CAES and CORE improve descriptions. CAES helps focus on key frames, while CORE models object interactions for more interpretable outputs.

G. Implementation Details

Anomaly intervals are detected using an adaptive threshold with the 97th percentile computed per video. For each interval, pre- and post-event contexts are determined by the 95th and 85th percentiles of the score slope over a 5-frame window, with a maximum context window of 30 frames. We uniformly sample 4 frames from each context, 8 from the event, and fill to 64 frames with background frames.

For relational analysis, object association combines cosine similarity of appearance embeddings (weight 0.8) and IoU, with matching solved by the Hungarian algorithm [3] and a maximum track age of 15 frames. Relational volatility at each timestep is measured as the maximum L2 distance between all co-tracked relation pairs in adjacent frames, followed by Gaussian smoothing with a standard deviation of 2.0. The top 5% of volatility peaks are used as positives.

The Relational Dynamic Encoder is a two-layer MLP trained with triplet margin loss [6] with margin 0.5 and semi-hard negative mining [6] with pool size 30. The encoder training uses Adam (learning rate 1×10^{-4}) with StepLR for 50 epochs.

For LLM fine-tuning, we adopt NVILA [4] as backbone, updating only the projector and LoRA [2] adapters while freezing all other parameters. The learning rate is set to 2×10^{-5} with cosine schedule and a warm-up ratio of 0.03.

References

[1] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18793–18803, 2024. 1, 2

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[3] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955. 4

[4] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 4

[5] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134, 2025. 3, 4

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 815–823, 2015. 4

[7] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. In *Neural Information Processing Systems (NeurIPS)*, 2024. 4

[8] 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*, pages 13843–13853, 2025. 1, 2