Harnessing Large Language Models for Training-free Video Anomaly Detection

Supplementary Material

In this supplementary material, we first provide the exact form of the prompts employed in our method and then we present additional experimental analyses. Specifically, we first present the impact of the task-related priors in prompting the anomaly scores on XD-Violence [36]. We then present the impact of captioning models, *i.e*. different variants of BLIP-2 models, for the VAD performance of our method on both XD-Violence [36] and UCF-Crime [24] datasets. Finally, we ablate the hyperparameters in constructing temporal windows to justify our design choice. Moreover, we describe the limitations and broader social impacts of our work, and we showcase additional qualitative results that demonstrate temporal summaries and the detection results. More qualitative results in the form of videos can be conveniently accessed on the project website at https://lucazanella.github.io/lavad/.

A. Prompts

The prompts utilized in our approach serve distinct purposes. The contextual prompt $\texttt{P}_C$ provides priors to the LLM for VAD. In line with the findings of our ablation studies presented in Tab. 4 and in Tab. 5, this prompt differs for UCF-Crime [24] and XD-Violence [36]. For UCF-Crime, the prompt is structured as: "*If you were a law enforcement agency, how would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities?*". In contrast, for XD-Violence, the prompt has the form: "*How would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious or potentially criminal activities?*".

The prompt $\texttt{P}_F$ provides guidance to the LLM for the desired output format, aimed at facilitating automated text parsing. This prompt remains consistent across both datasets and is defined as follows: "*Please provide the response in the form of a Python list and respond with only one number in the provided list below [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] without any textual explanation. It should begin with '[' and end with ']'.*".

Lastly, the prompt $\texttt{P}_S$ is employed to obtain a temporal summary S_i for each frame $\mathbf{I}_i$. The prompt is formulated as follows: "*Please summarize what happened in few sentences, based on the following temporal description of a scene. Do not include any unnecessary details or descriptions.*".

B. Additional analyses

Task priors in the context prompt. In Tab. 5 we present the impact of different priors in the context prompt $\texttt{P}_C$, i.e. impersonation and anomaly priors, on XD-Violence [36]. This follows the same ablation design as presented in Tab. 4 in the main manuscript for UCF-Crime, with the priors added in the same way for both datasets. As shown in Tab. 5, for videos within XD-Violence, incorporating the anomaly prior (Row 2) improves the average precision (AP) by $+1.67\%$ compared to using only the base context prompt (Row 1). Conversely, introducing impersonation (Row 3) degrades the AP by -1.51% compared to not using it (Row 1). Videos in XD-Violence originate from various sources, including CCTV cameras, movies, sports, and games. The effectiveness of the impersonation prior might be limited to CCTV camera videos, given that the surveillance domain is more closely associated with the concept of *"law enforcement agency"* which is utilized for impersonation. Finally, combining both priors (Row 4) leads to improved performance compared to not utilizing any of them, primarily due to the positive impact of the anomaly prior.

Table 5. Results of LAVAD on XD-Violence with different priors in the context prompt when querying the LLM for anomaly scores.

ANOMALY PRIOR	IMPERSONATION	AP (%)	AUC (%)
✗	✗	60.34	84.42
✓	✗	**62.01**	**85.36**
✗	✓	58.83	84.50
✓	✓	60.78	85.26

Impact of different BLIP-2 models. As captioners, we consider different BLIP-2 [14] models and their ensemble for both UCF-Crime [24] and XD-Violence [36], and we present the results in Tables 6 and 7, respectively.

In Tab. 6, the most effective strategy for UCF-Crime videos is employing an ensemble of all BLIP-2 models (Row 6). This involves generating captions for all frames in a video using all BLIP-2 models and relying on the vision-language model (VLM) to identify the semantically closest captions for each frame. The effectiveness of the ensemble might be attributed to the challenges posed by UCF-Crime videos. These videos, characterized by low-resolution CCTV footage, often lead captioning models to hallucinate scene descriptions. For instance, it is common to encounter captions, such as *"a person riding a skateboard down a road"* when the image only depicts a road in the absence of any specific event. The ensemble approach, by allowing the selection from a larger set of candidates, increases the likelihood of choosing more correct captions and filtering incorrect ones.

For XD-Violence, as shown in Tab. 7, utilizing the captions generated by *flan-t5-xxl* (Row 3) yields the best average

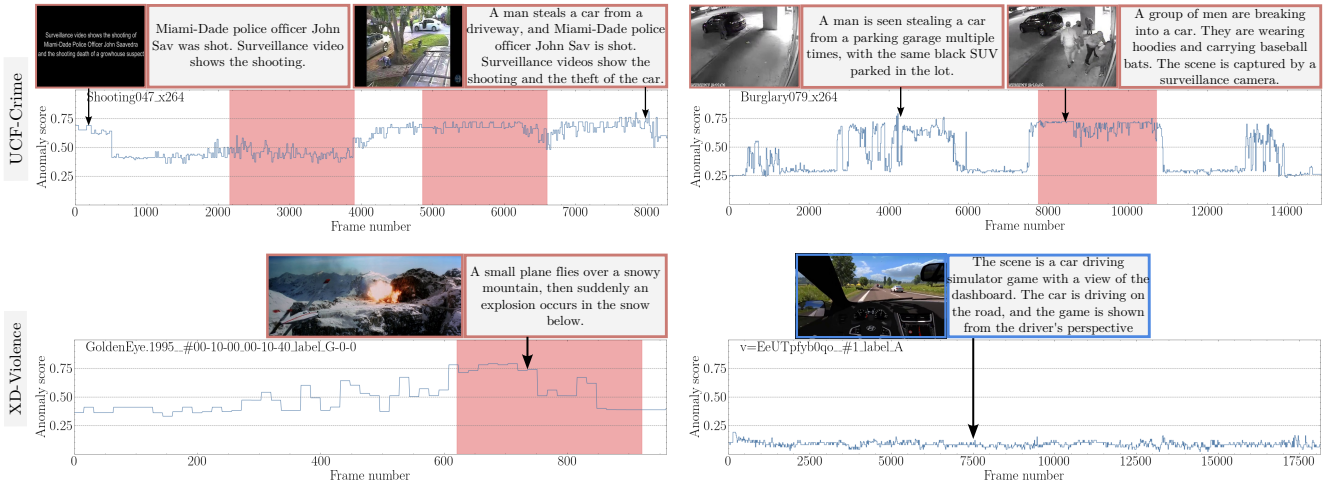

Figure 7. We showcase qualitative results obtained by LAVAD on four test videos, including two videos (top row) from UCF-Crime and two videos from XD-Violence (bottom row). For each video, we plot the anomaly score over frames computed by our method. We display some keyframes alongside their most aligned temporal summary (blue bounding boxes for normal frame predictions and red bounding boxes for abnormal frame predictions), illustrating the relevance among the predicted anomaly score, visual content, and description. Ground-truth anomalies are highlighted.

Table 6. Results of LAVAD on UCF-Crime with different BLIP-2 model variants in our Image-Text Caption Cleaning technique.

BLIP-2					AUC
FLAN-T5-XL	FLAN-T5-XL-COCO	FLAN-T5-XXL	OPT-6.7B	OPT-6.7B-COCO	(%)
✓	✗	✗	✗	✗	74.19
✗	✓	✗	✗	✗	74.49
✗	✗	✓	✗	✗	74.38
✗	✗	✗	✓	✗	75.50
✗	✗	✗	✗	✓	73.94
✓	✓	✓	✓	✓	**80.28**

Table 7. Results of LAVAD on XD-Violence with different BLIP-2 model variants in our Image-Text Caption Cleaning technique.

BLIP-2					AP	AUC
FLAN-T5-XL	FLAN-T5-XL-COCO	FLAN-T5-XXL	OPT-6.7B	OPT-6.7B-COCO	(%)	(%)
✓	✗	✗	✗	✗	61.09	85.16
✗	✓	✗	✗	✗	57.41	82.78
✗	✗	✓	✗	✗	**62.01**	**85.36**
✗	✗	✗	✓	✗	56.55	82.42
✗	✗	✗	✗	✓	54.71	82.93
✓	✓	✓	✓	✓	59.62	84.90

precision (AP). Other BLIP-2 variants for XD-Violence may provide captions that prioritize foreground objects, potentially overlooking background elements constituting anomalies (*e.g.* a vehicle enveloped in smoke on a busy street), yet better aligning with the VLM's representation of the video frames. Hence, when employing the ensemble of BLIP-2 models (Row 6), captions that specifically highlight elements constituting anomalies are not chosen as the semantically closest captions to video frames in the cleaning step, with a negative impact on the anomaly scoring phase.

Temporal window's duration and number of sampled frames. In Tab. 8, we evaluate the impact of varying the duration of the temporal window (T) and the number of sampled frames (N), which is used to query the LLM for the temporal summary S_i. Specifically, the temporal window duration T determines the time interval, while the number of sampled frames N determines the number of captions. First, we conduct experiments by adjusting the duration T to 2.5, 5, 10, and 20 seconds, while maintaining $N = 10$. The 10-second temporal window yields the highest AUC score (Row 3). This is in line with the fact that ImageBind [6] is trained with video clips of 10 seconds.

Table 8. Results of LAVAD on UCF-Crime with different combinations of temporal window duration (T) and number of sampled frames per window (N).

T(s)	N	AUC (%)
2.5	10	79.33
5	10	78.10
10	10	**80.28**
20	10	79.24
10	5	77.48
10	20	74.45

Subsequently, we maintain the temporal window's duration T at 10 seconds and vary the number of frames from 5 to 10 and 20. Notably, using 10 frames (Row 3), *i.e.* 1 frame every second, is the optimal choice within this experiment. Balancing the number of captions per snippet presents a trade-off with the quality of the summary. Too many captions may overwhelm with excessive and non-diverse content, while too few captions may result in limited coverage of the content.

C. Qualitative results

In Fig. 7, we present additional qualitative results demonstrating the effectiveness of our proposed LAVAD in detecting anomalies within a set of UCF-Crime [24] and XD-Violence [36] test videos. The figure showcases keyframes along with the most semantically similar temporal summaries. For example, in the video *Shooting047* (Row 1, Column 1), LAVAD assigns a high anomaly score when the video is labeled abnormal. However, it also assigns a high anomaly score during the initial and final segments, despite these parts being labeled as normal. This discrepancy arises because the video begins with text describing the subsequent content, leading the LLM to attribute a high anomaly score. In the final part, our method correctly identifies abnormality as the frame depicts a person on the ground who has been shot. In the video *Burglary079* (Row 1, Column 2), there is a false abnormal instance. This occurs because the temporal summary associated with that frame incorrectly suggests the presence of a man stealing a car. In reality, the video depicts a man behaving suspiciously near the car, leading to a wrong interpretation by the captioning module. In the XD-Violence videos (Row 2), an anomaly caused by an explosion is correctly detected (Row 2, Column 1), while a normal video is consistently predicted as normal for more than $17,500$ frames (Row 2, Column 2).

D. Limitations

We identify two main limitations of our work. Firstly, our method fully relies on pre-trained models from VLMs and LLMs, thus its performance greatly depends on i) how well the captioning model describes the visual content, ii) how reliable the LLM is when generating the anomaly scores, and iii) how aligned the multi-modal encoders are when processing videos from various domains. Secondly, our anomaly scores are primarily obtained via prompting LLMs. Although we conducted experiments investigating different prompting strategies, a systematic understanding of LLM prompting for VAD requires a community effort.

E. Broader Societal Impacts

While our work pioneers the technical aspect of leveraging LLMs for detecting anomalies in videos, there exist open ethical challenges for a broader concern. VAD systems are mostly applied to safety-related contexts, for private use or public interests. Prior to any deployment, it is crucial to first investigate the behaviors of LLM-based methods, mitigating any potential bias in LLMs and improving explainability. Our work serves as the first technical exploration of leveraging LLMs for training-free VAD, proving it as a competitive alternative. This is a necessary step to increase the awareness of the community on these important topics.