Week 6

07/03/2025

Multi-modal Large Language Models for Video Anomaly Detection

Lead：Dr. Tian

Eric Li

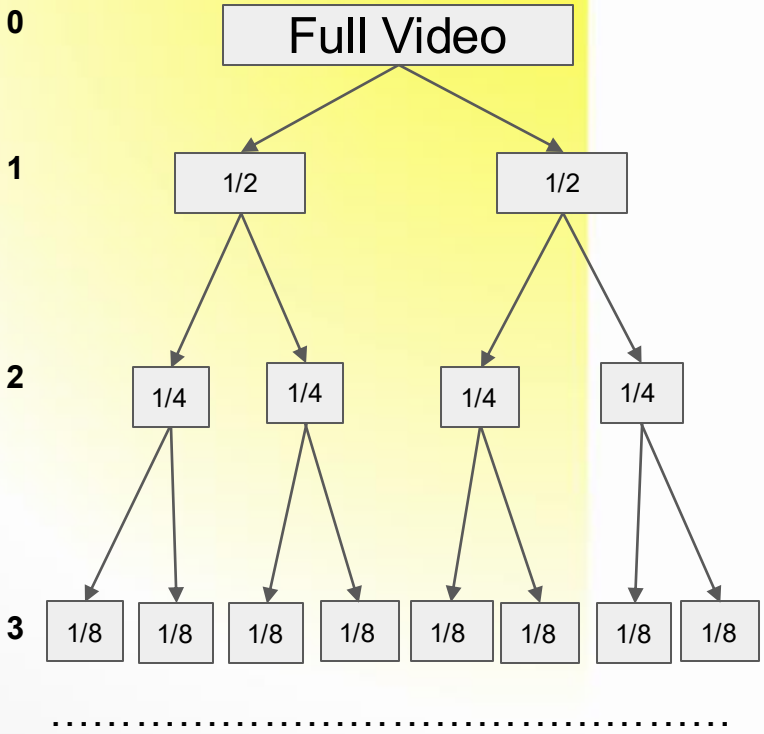
Goal:

Reference: Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E. (2024). *Harnessing Large Language Models for Training-free Video Anomaly Detection*. CVPR 2024.

- To create a new **training-free LVLM** with enhanced Video Anomaly Detection (VAD) capabilities.

(Currently, the highest **ROC AUC** achieved by training-free LVLMs on the UCF-Crime dataset is around 0.8)

Our current model:



Keep splitting in half. Eventually, we will get all frames at the end.

Execution Logic:

1. Splitting each video clip from last level into half
2. Apply Video LLM to obtain the summary of each video clip at each level.
3. For each frame, we use LLaMA2 to aggregate the summaries of all levels of video clips that contain that frame, generating a final summary.
4. Focusing on the caption with the contextual information that the Final Summary provided, ask LLaMA2 to give each frame an anomaly score.

Summary Prompt: You are itself with the summarization expert. Here are many summarys, now combine them into a new, concise summary, keeping all abnormal or criminal behaviors.

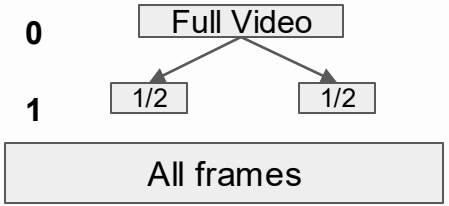
Score Prompt: You are an expert security analyst. The following is a combined summary of video segments that contain the current frame. It may describe one or more abnormal events. Focusing on the caption with the summary, respond **only** with one number in this exact Python list format: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. Your response MUST be a single number in this list format and nothing else.

What I did this week

Last week, I used Video-LLaVA to generate summaries for each video clip of the model on the left.

Based on its poor performance on long videos, I replaced it with **LLaVA-OneVision** and make a new dataset with 48 short videos (half abnormal, half normal) from ucf_crime dataset.

- A small demo of partial of our model:

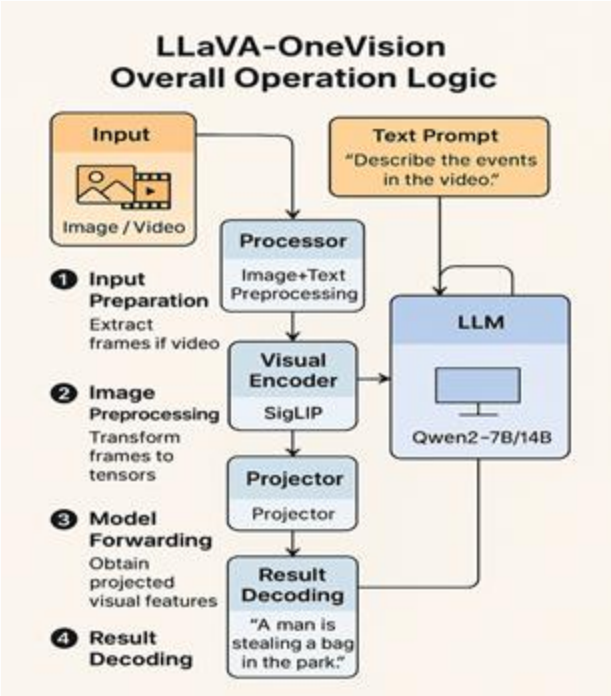


I conducted four experiments to test if contextual information

can help LLM score more accurately:

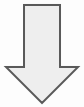
By providing LLaMA2 with different levels of information to score each frame.

1. Only the **caption** of each frame
2. **full-video summary** & **caption**
3. **full-video summary** & **half-video summary** & **caption**
4. **full-video summary** & **half-video summary** & **10-second video summary** & **caption**

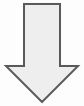


Current Difficulties:

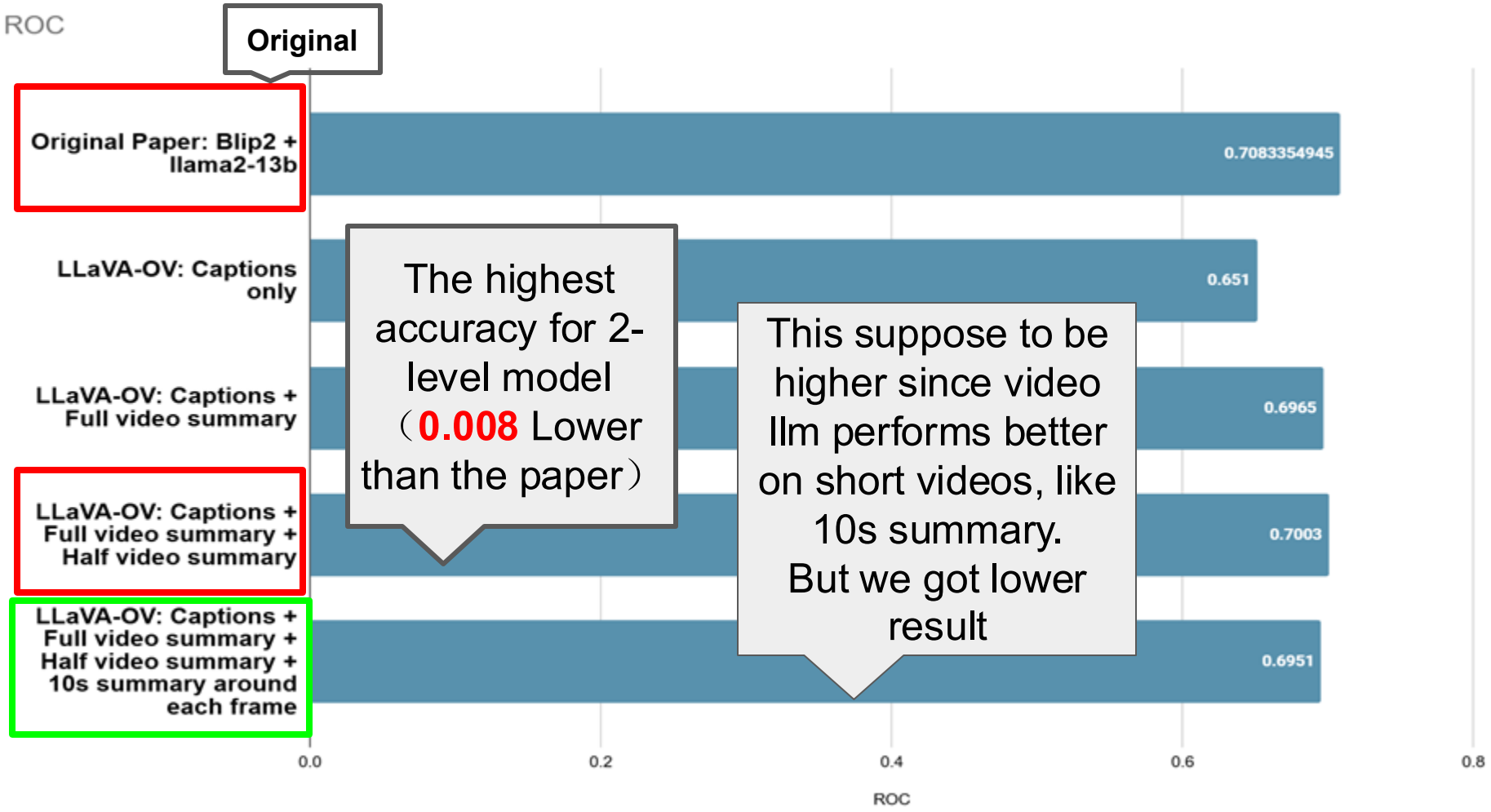
Both Video-LLaVA and LLaVA-oneVison don’t perform well on long videos. Thus, the summaries of these full abnormal videos, half abnormal videos don’t include any anomaly behaviors.



Mislead LLM to give lower scores even if some captions describe some abnormal behavior.



Therefore, I think the previous experiments are not meaningful, as they were not based on the complete implementation of our model and also failed to reflect whether contextual information is helpful for LLM-based scoring. **So, I have implemented the complete code for the entire model.** It is currently running, and since each video takes over 20 minutes to process, I expect to get the results tonight.


ROC
Original
Original Paper: Blip2 + llama2-13b
LLaVA-OV: Captions only
LLaVA-OV: Captions + Full video summary
LLaVA-OV: Captions + Full video summary + Half video summary
LLaVA-OV: Captions + Full video summary + Half video summary + 10s summary around each frame
0.7083354945
0.651
0.6965
0.7003
0.6951
The highest accuracy for 2-level model （0.008 Lower than the paper）
This suppose to be higher since video llm performs better on short videos, like 10s summary. But we got lower result
0.0
0.2
0.4
0.6
0.8
ROC

Future plan

1. Conducting an ablation study by removing the full-video, half-video, or any other levels video summaries to see if the model performance would improve. (This is because current video LLMs do not perform well when dealing with full or overly long videos.)

1. To be discussed.

Thank you!