Week 5

06/26/2025

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.

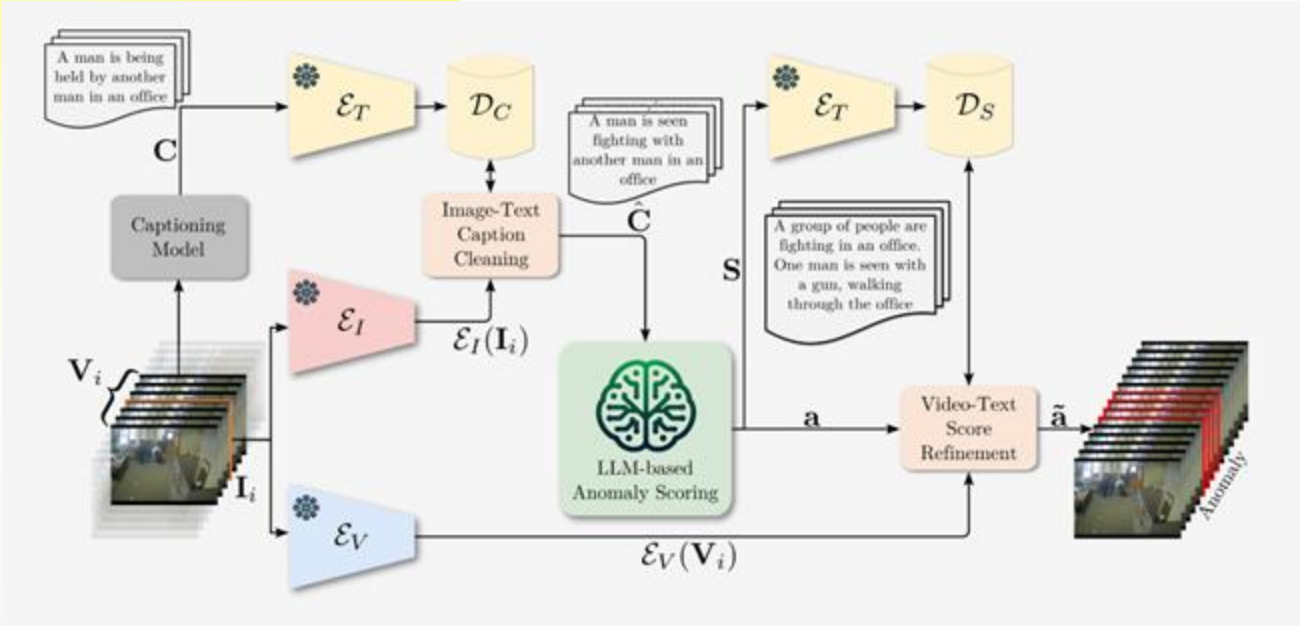
- To create a new **training-free LVLM** with enhanced Video Anomaly Detection (VAD) capabilities.

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

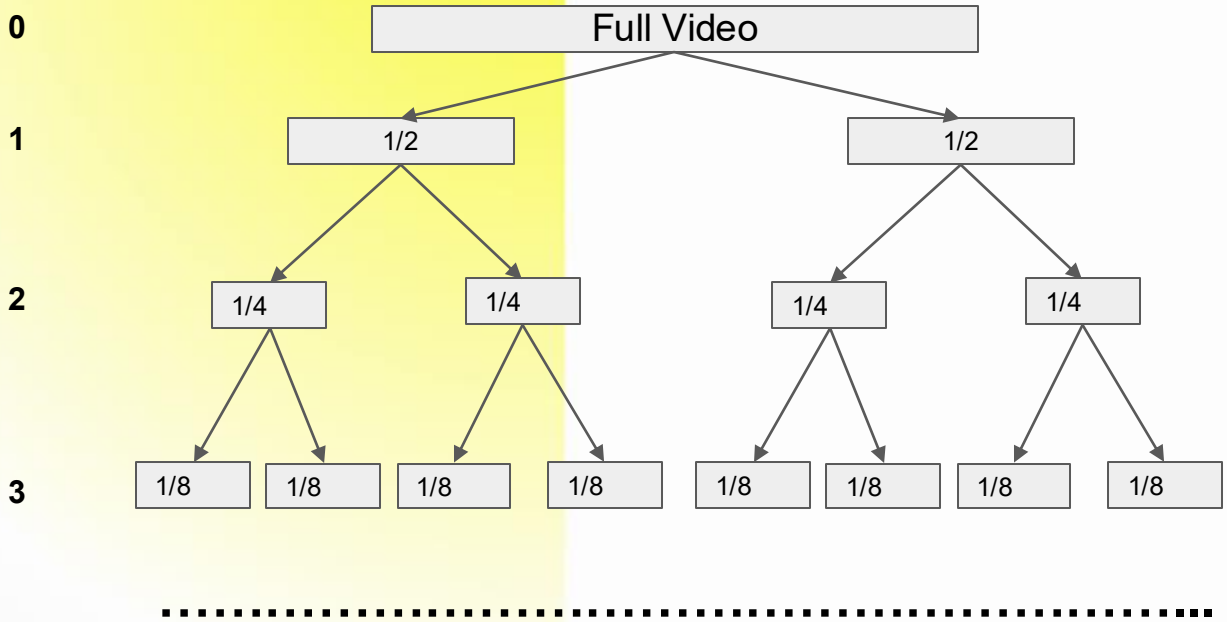
What I’ve finished

- Reproduce the paper:

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



Our current model:



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

Steps:

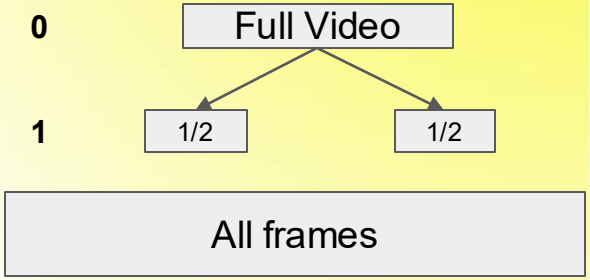
- Splitting each video clip from last level into half
- Obtain the summary of each video clip at each level.
- Scoring: When the LLM assigns a score to each frame, it will take into account all the video clip summaries from previous levels that include that frame.

Advantage:

After understanding more contextual information about each frame, the model can assign anomaly scores to each of them more accurately.

What I did this week

- A small demo of partial of our model:

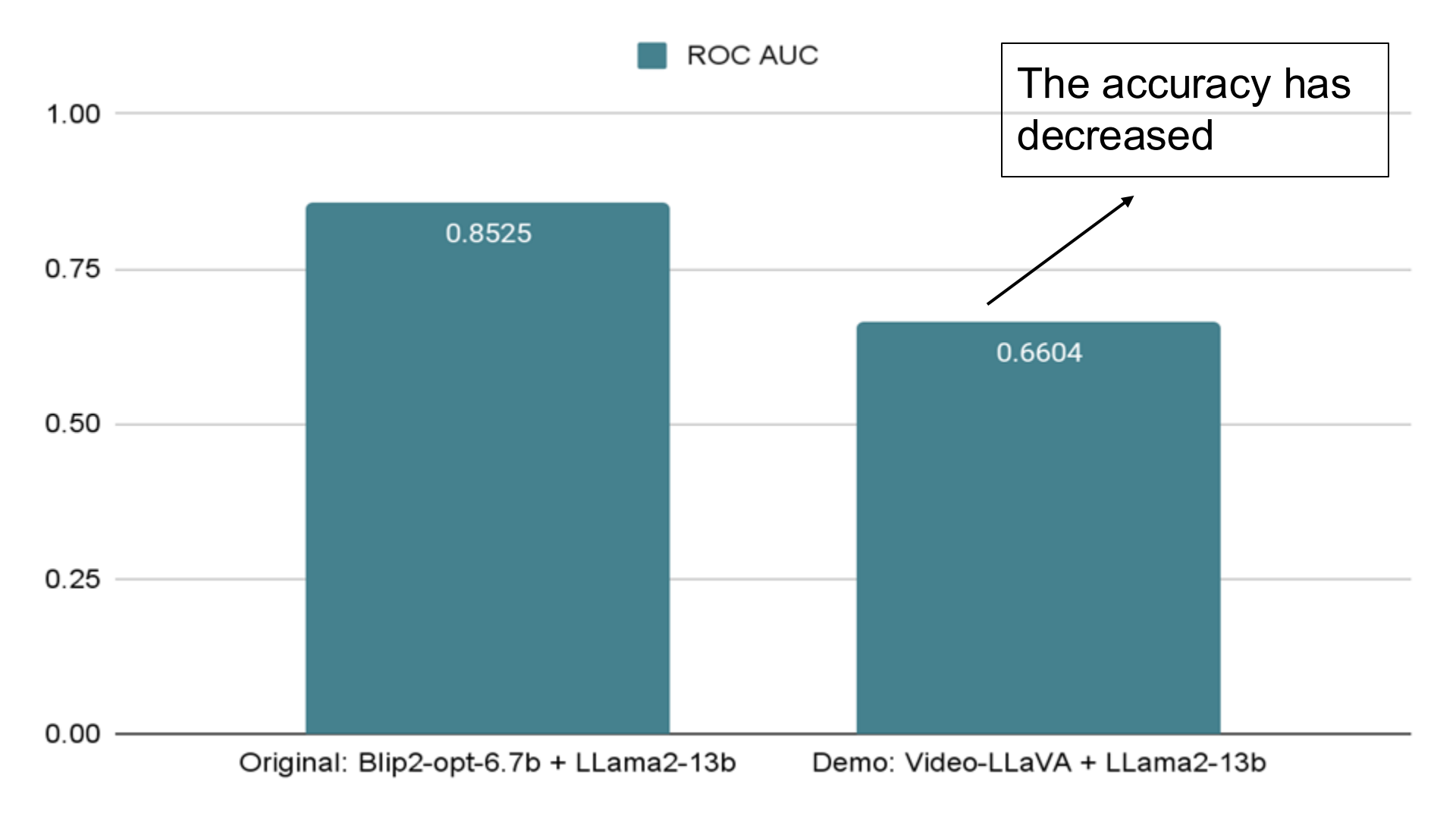


Prompt:
"If you were a law enforcement agency, how would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities? Focus primarily on the frame-level caption. The half-video summary, and full-video summary are provided for context only and may not always be correct."

1. Use Video-LLaVA 1.0 to generate a summary of the entire video, as well as summaries of two halves obtained by splitting the full video in half.

2. Apply LLaMA2-13B to score each frame: while focusing on the frame's own caption, it also takes into account the two related summaries.

3. Calculate ROC AUC


ROC AUC
The accuracy has decreased
1.00
0.75
0.50
0.25
0.00
0.8525
0.6604
Original: Blip2-opt-6.7b + LLama2-13b
Demo: Video-LLaVA + LLama2-13b

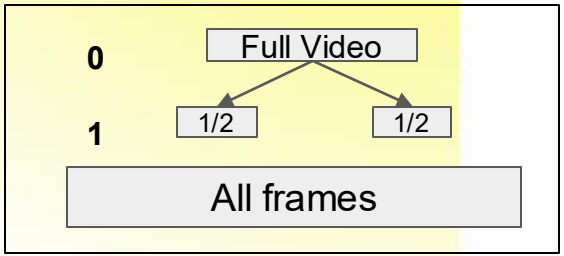
Current Difficulties:

1. Video-LLaVA has already been updated to version 1.5, but I only have access to version 1.0, which selects eight frames evenly from the video regardless of its length.Thus, its ability to summarize the entire video is not good—especially for videos containing abnormal behavior, where the generated summaries can’t include any mention of the anomalies.

2. LLaMA2 has a relatively low token limit, which makes it impossible to provide all the summaries at once when scoring each frame. As a result, we can't directly input the summaries at every level into the model at once during the scoring process for a single frame.

Future plan:

- To address the current weakness of the Video-LLaVA model in summarizing videos, I plan to let LLaMA2-13B read and understand all the frame-level captions to generate a more detailed summary of the entire video instead. Then, I will incorporate this new summary into the demo model I'm currently using to see how it performs.



- Try implementing a complete model to see how it performs, or consider creating an entirely new one from scratch.

Thank you!