Week 8

07/17/2025

Multi-modal Large Language Models for Video Anomaly Detection

Lead : Dr. Tian

Eric Li

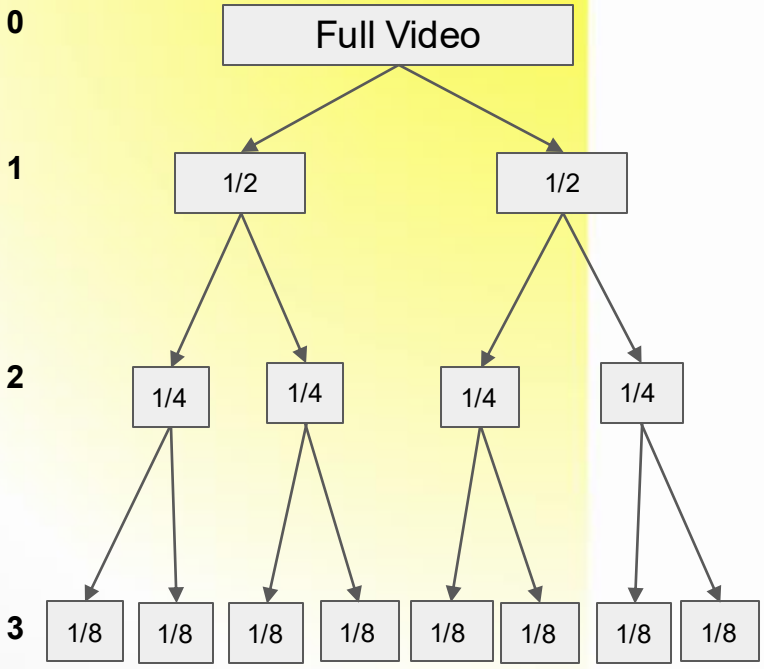
Goal:

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.

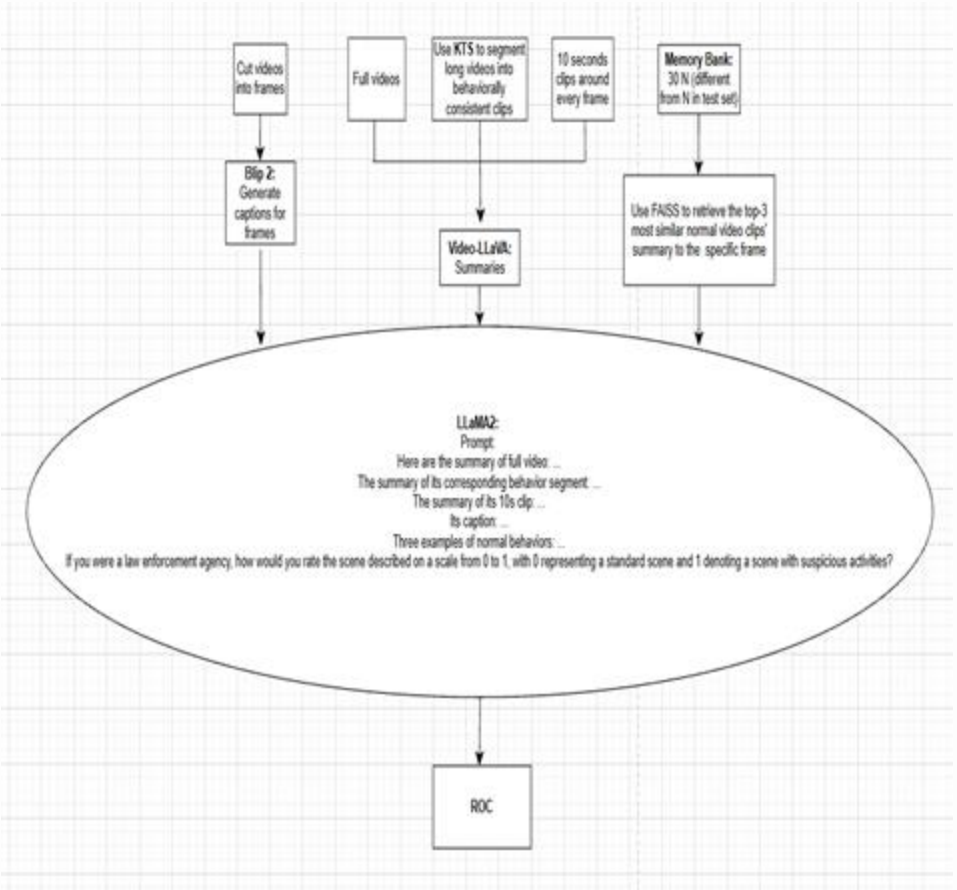
(Currently, the highest **ROC AUC** achieved by training-free LVLMs on the UCF-Crime dataset is around 0.8)

Our model 1.0:



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

Our model 1.1:

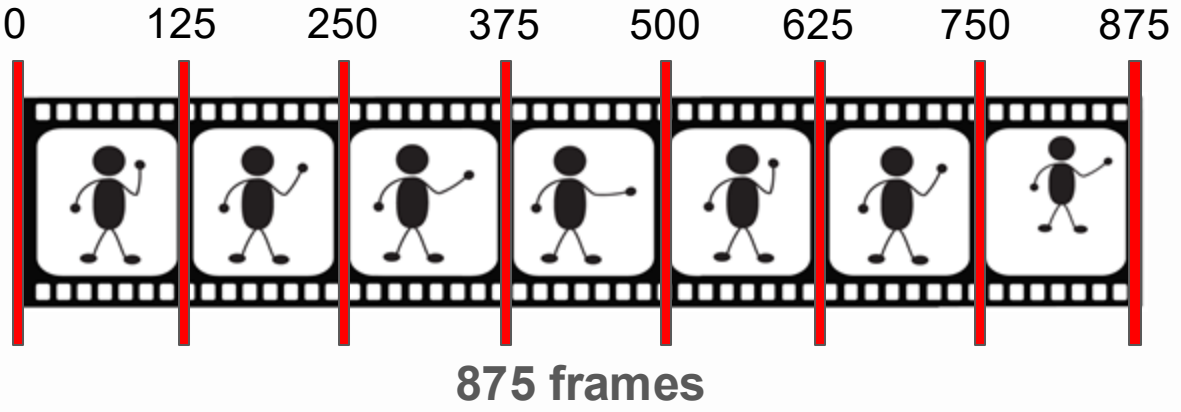


Reasons → Low Results

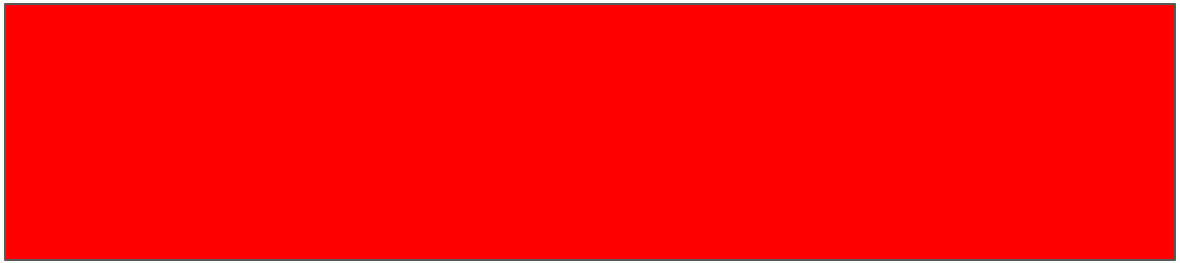
When using LLaVA-OneVision to do summarizations, I used an incorrect frame sampling method.

The poor video summaries **MISLED** the LLM when it assigned anomaly scores to each 16 frames.

What I did: Uniform frame sampling



New Way: Sampling every frames to generate the summaries



Our model 1.2:

Our previous design logic:

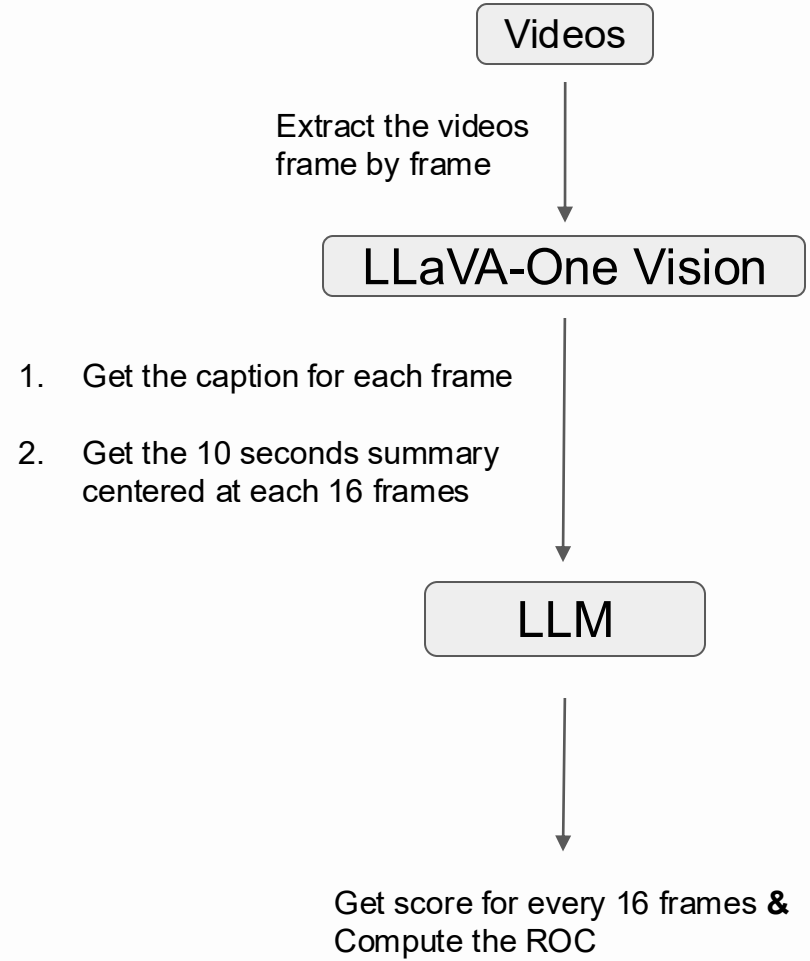
If we can provide the LLM with more information, it can generate more accurate scores.

→ Our past 2 models have complex structures.

Our latest design logic:

Start to build our model from the most fundamental structure & find the optimal solution at each step.

→ Create a better model than the paper



Results with different LLM:

(ROC got by the paper: 0.78)

LLM	ROC
LLaVA-One Vision	0.55
LLaMA2-13b	Waiting for GPU access
LLaMA3-8b	Waiting for GPU access

Future plan

1. After we got the result of our new model with llama2-13b & llama3, we will decide what to adjust or add to our model next.

Thank you!