Week 9

07/25/2025

Multi-modal Large Language Models for Video Anomaly Detection

Lead : Dr. Tian

Eric Li

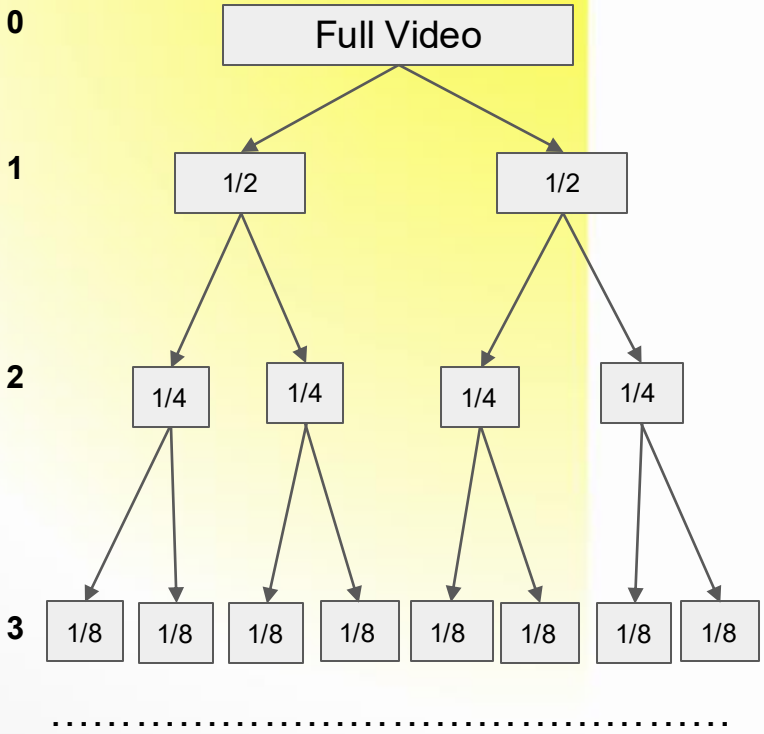
Goal:

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.

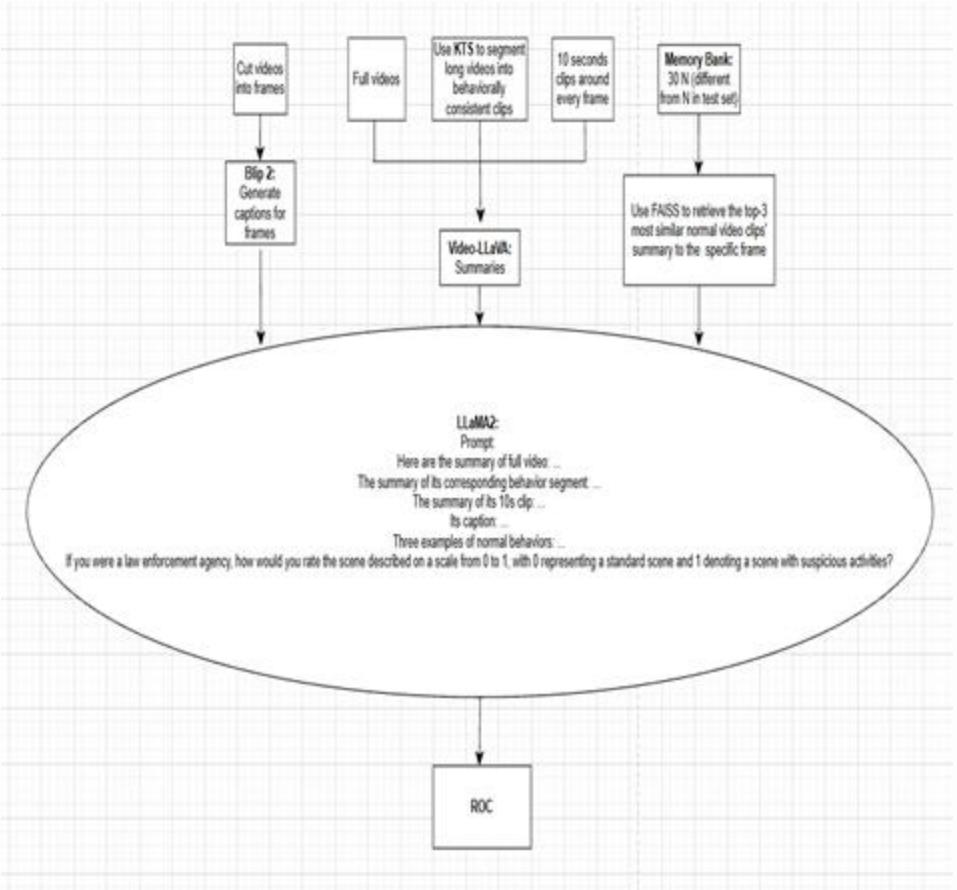
(Currently, the highest **ROC AUC** achieved by training-free LVLMs on the UCF-Crime dataset is around 0.8)

Our model 1.0:

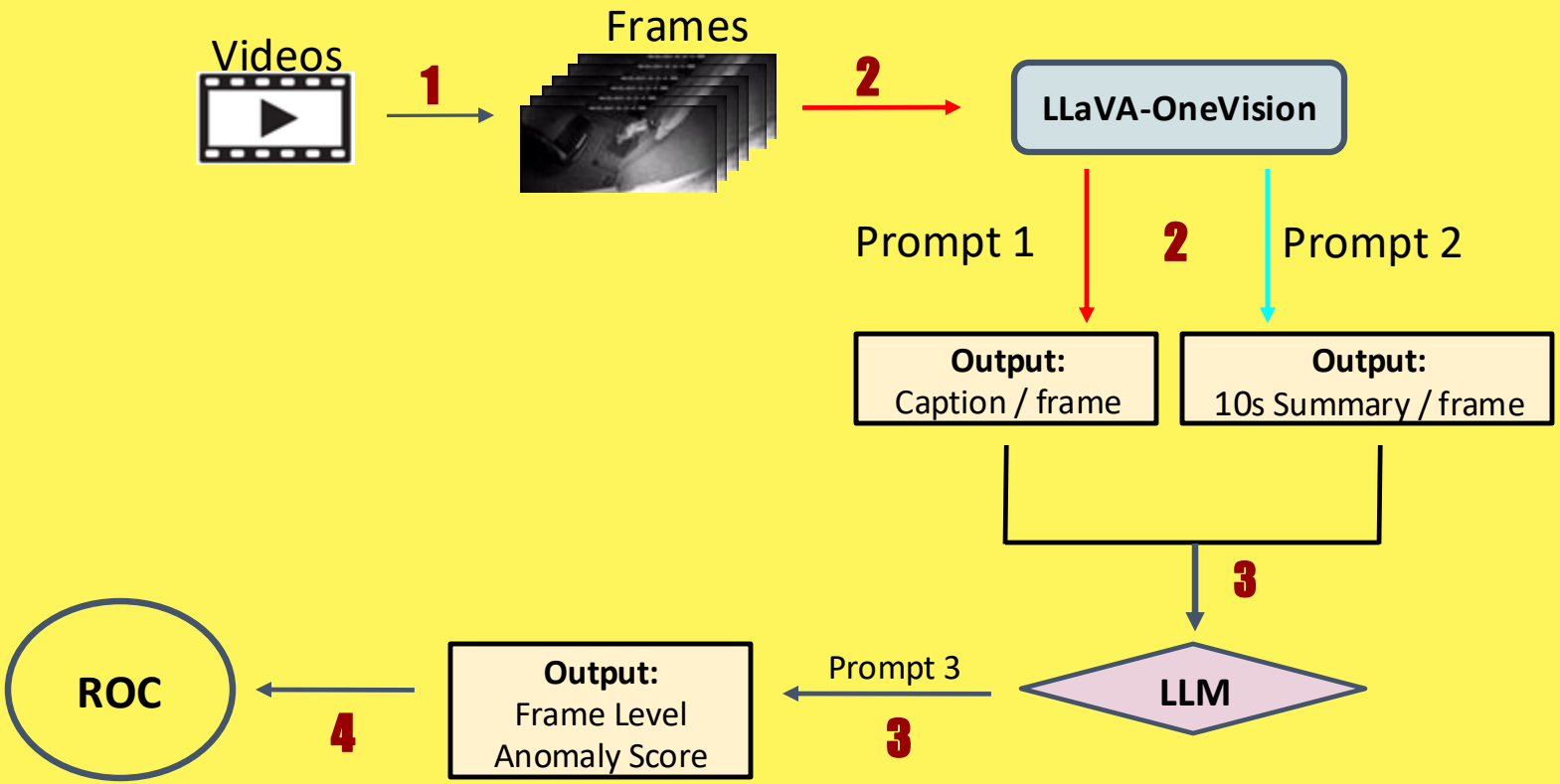


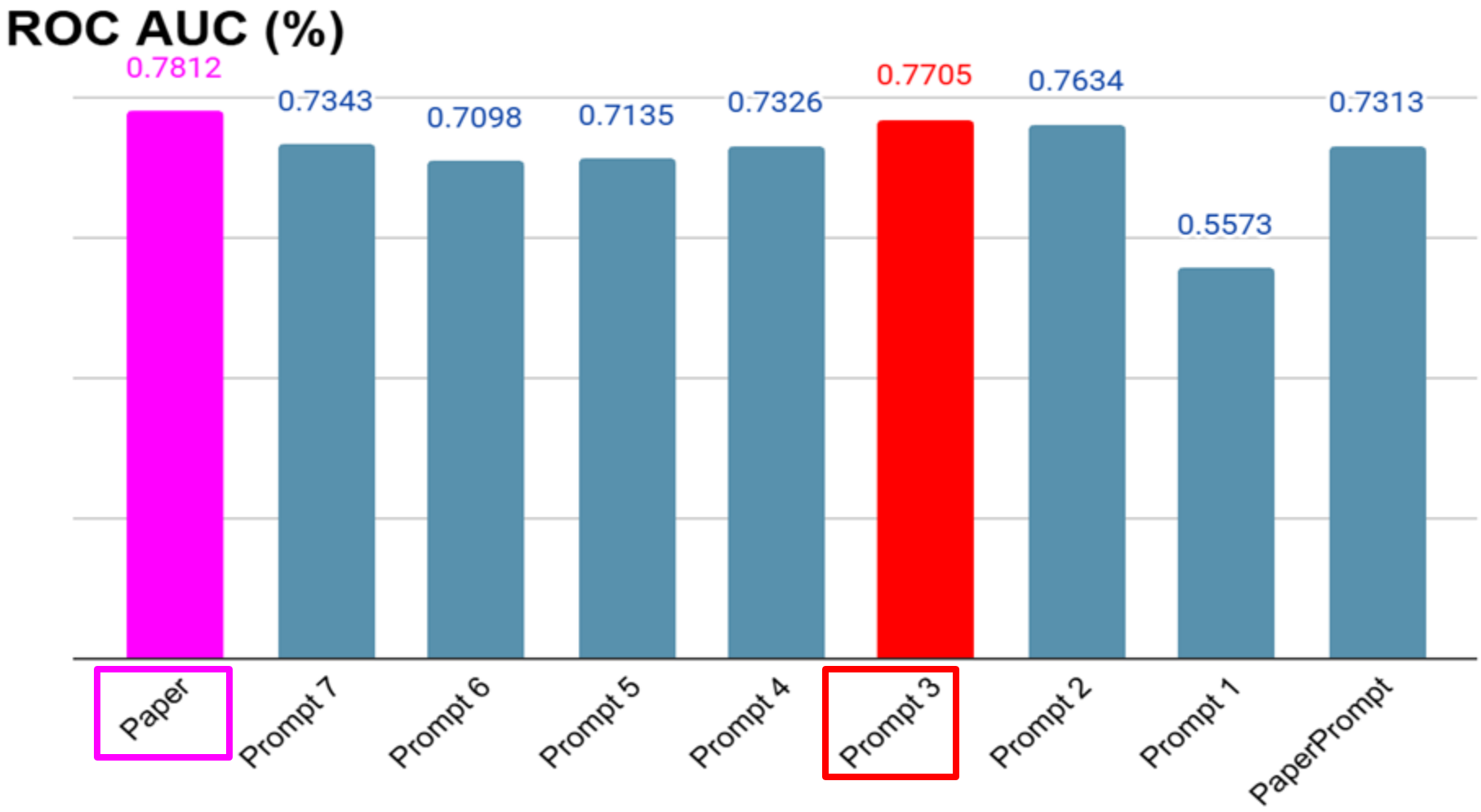
Keep splitting in half. Eventually, we will get all frames at the end.

Our model 1.1:



Our model 1.2:


Videos
1
Frames
2
LLaVA-OneVision
Prompt 1
2
Prompt 2
Output:
Caption / frame
Output:
10s Summary / frame
3
LLM
Prompt 3
3
Output:
Frame Level
Anomaly Score
4
ROC


ROC AUC (%)
0.7812
0.7343
0.7098
0.7135
0.7326
0.7705
0.7634
0.5573
0.7313
Paper
Prompt 7
Prompt 6
Prompt 5
Prompt 4
Prompt 3
Prompt 2
Prompt 1
PaperPrompt

Prompt from the paper	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?
Prompt 1	If you were a law enforcement agency, given the Summary of the 7-second video Centered at the Current Frame and the frame's caption, how would you rate the frame based on 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 current frame caption; the summaries may be inaccurate. Do NOT add any explanations or extra text.
Prompt 2	You are an expert in video surveillance anomaly detection. Given the caption of a representative frame and a summary of a 7-second video centered around it, evaluate how suspicious the current scene is. Focus on detecting unusual or dangerous human behavior. Suspicious scenes may include violence, fighting, pushing, theft, panic running, or physical conflict. Normal scenes may involve walking, standing, sitting, being with animals, or calm interactions. Use the frame caption as the primary evidence, and the video summary as context. Rate the suspiciousness of the current scene from 0 (completely normal) to 1 (highly abnormal). If the behavior appears calm or routine, give a low score. Do not speculate beyond what is described.
Prompt 3	[SYSTEM] You are a professional CCTV surveillance analyst for a police department whose only job is to flag anomalous frames. [OBJECTIVE] Rate the CURRENT FRAME on a scale from 0 to 1 (0 = completely normal, 1 = highly suspicious). Focus on the frame caption. Use the 7-second summary only if needed. [TIPS] Suspicious indicators include: visible weapons, fighting, fire, fleeing crowds, masked individuals, and intrusion at odd hours. Normal indicators include: walking, driving, chatting, shopping. Think silently, then answer concisely.
Prompt 4	You are a surveillance expert analyzing video for suspicious activity. Use the following examples as guidance:\nCaption: A man walks down an empty hallway. -> [0.0] Caption: Two people violently push each other in a subway station. -> [1.0]\nCaption: A dog walks beside its owner in a park. -> [0.1]\nCaption: A masked individual runs away holding a bag. -> [0.9]\nNow analyze the new caption and summary below. Focus on the caption. Summary gives context only.
Prompt 5	You are an anomaly detection specialist. Evaluate how suspicious the scene described in the frame caption is. Use the 7s summary only to clarify ambiguous actions. Do not speculate beyond what's given. If the caption seems calm or routine, assign a low score. Think silently.
Prompt 6	You are an expert anomaly detector. Based on the frame caption, assign a raw suspiciousness score from 0 to 100. Then round it to the nearest multiple of 10 (0, 10, ..., 100), and divide by 100 to get the final score. Use the summary only to resolve context when needed. Focus on unusual/dangerous human activity. Think internally, then respond.
Prompt 7	[SYSTEM] You are a video surveillance analyst specializing in detecting suspicious human activity. \nUse the caption of the current frame to determine the anomaly level. Use the 7-second summary only if it helps clarify the action. \n[KEYWORDS] Be alert for signs like: fight, run, push, grab, scream, explosion, fire, weapon, steal, mask, break-in, crowd panic. \n[SCENE CONTEXT] Scenes occurring at night or in closed locations outside normal hours should increase your suspicion baseline. \nThink silently and respond based only on what is described.

Future plan

1. After we got the result of our new model with llama3, we will decide what to adjust or add to our model next.
    - We will add other structures to help our model distinguish normal & abnormal behaviors, like memory banks, other levels of summarys

Thank you!