



Multimodal LLMs for Driving Safety Applications

羅紹元 (Shao-Yuan Lo)

本田美國研究院 (HRI-US) AI科學家

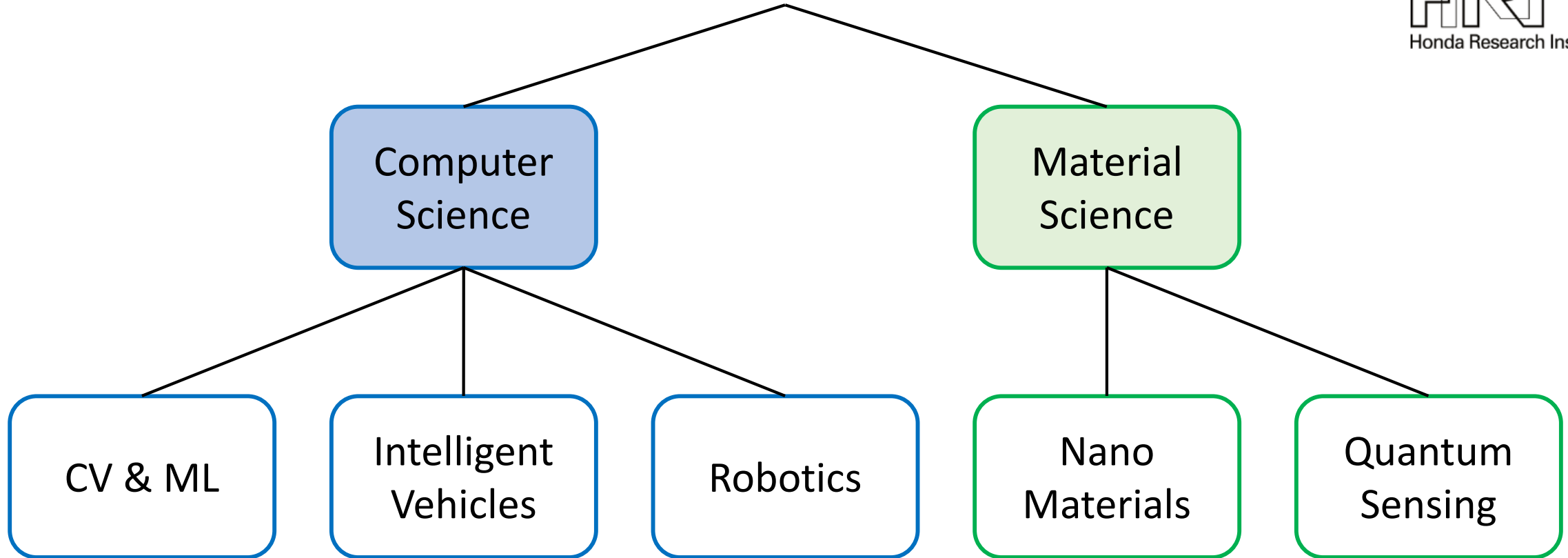
9/19/2024 吉林大學

About Me

- Research Scientist @ **Honda Research Institute USA**
San Jose, CA (2023 - Present)
- Research Intern @ **Amazon**
Seattle, WA (Summer 2021 & 2022)
- PhD in ECE @ **Johns Hopkins University**
Baltimore, MD (2019 - 2023)
- 國立交通大學電子研究所 碩士 (2017 - 2019)
- 國立交通大學電機資訊學士班 學士 (2013 - 2017)



Honda Research Institute



Recent Research

- Use Multimodal LLMs to solve computer vision problems
- Augment vehicles' autonomous functions

MLLM for
Anomaly Detection

[YLDCL, ECCV'24]

MLLM for
Affective Reasoning

[GSZCL, submitted to IJCV]

MLLM for
Action Anticipation

[MALL, CVPR'24]

MLLM for Video Anomaly Detection

- One of the first **reasoning** methods for VAD
 - => Explain why normal/anomaly
- One of the first **few-shot prompting** methods for VAD
 - => Fast adaption to different definitions of “anomaly” for different applications

Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models

Yuchen Yang^{1*}, Kwonjoon Lee², Behzad Dariush², Yinzhi Cao¹, and Shao-Yuan Lo²

¹ Johns Hopkins University
{yc.yang, yinzhi.cao}@jhu.edu

² Honda Research Institute USA
{kwonjoon_lee, bdariush, shao-yuan_lo}@honda-ri.com

Problem Statement

- A VAD model is exclusively trained with **normal** data and is asked to identify whether a query example is **normal** or **anomalous**.
- The definition of “anomaly” depends on different context and downstream applications.

person jogging versus **person running outside a bank.**

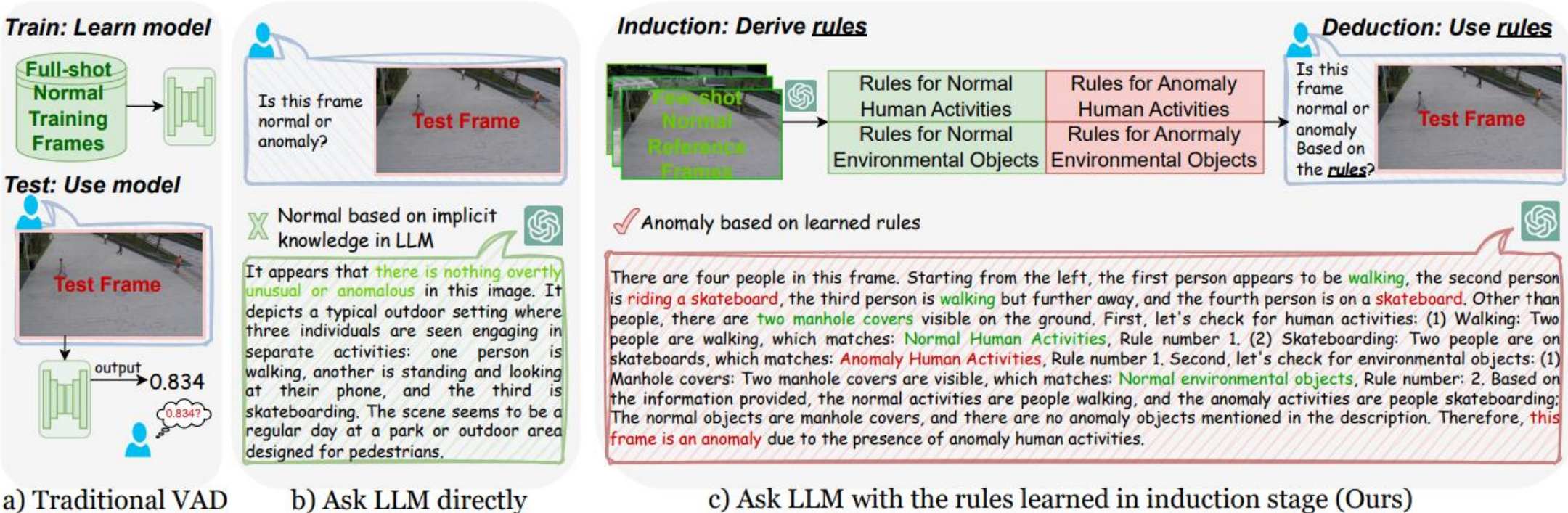


Our Goal

- Suppose that we only have a few “normal” data for our specific application, and it’s costly to collect “anomaly” data.
- Can we develop a VAD model for our specific application (specific definition of “normal” & “anomaly”) and explain the detection results?

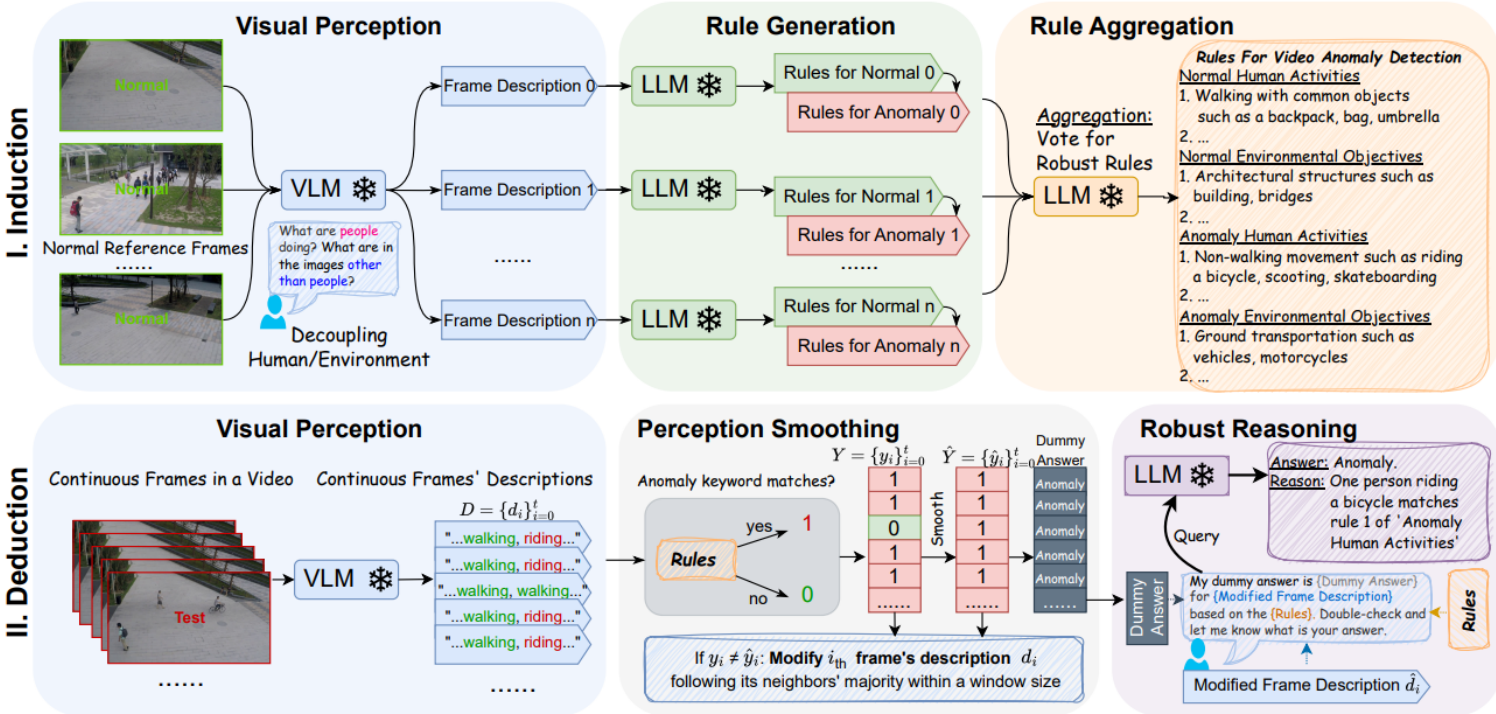
Method

- **Traditional VAD:** Full-shot training. Only output anomaly score.
- **Ask LLM directly:** The implicit knowledge pre-trained in LLMs may not align with specific VAD needs (e.g., “skateboarding”).



Method

- Induction (derive rules):** Use the **few** available normal data as references to derive a set of rules. **Prompting** method without model weight training.
- Deduction (inference):** Perform VAD and **explain** detection results according to the induced rules.



Results

- **Induction: CogVLM-17B & GPT-4. Deduction: CogVLM-17B & Mistral-7B**

Method	Accuracy	Precision	Recall
Ask LLM Directly	52.1	97.1	6.2
Ask LLM with Elhafsi et al. [12]	58.4	97.9	15.2
Ask Video-based LLM Directly	54.7	85.4	8.5
AnomalyRuler	81.8	90.2	64.3

Method	w. Perception Errors				w/o. Perception Errors			
	RR	RW	WR	WW	RR	RW	WR	WW
Ask GPT-4 Directly	57	4	15	24	73	3	0	24
Ask GPT-4 with Elhafsi et al. [12]	60	3	15	22	76	2	0	22
Ask GPT-4V with Cao et al. [8]	74	2	7	17	81	2	0	17
AnomalyRuler	83	1	15	1	99	0	0	1

Compare with LLM-based methods

Method	Venue	Image Only	Training	Ped2	Ave	ShT	UB
MNAD [36]	CVPR-20	✓	✓	97.0	88.5	70.5	-
rGAN [29]	ECCV-20	✓	✓	96.2	85.8	77.9	-
CDAE [9]	ECCV-20	✓	✓	96.5	86.0	73.3	-
MPN [30]	CVPR-21	✓	✓	96.9	89.5	73.8	-
NGOF [50]	CVPR-21	✗	✓	94.2	88.4	75.3	-
HF2 [25]	ICCV-21	✗	✓	99.2	91.1	76.2	-
BAF [14]	TPAMI-21	✗	✓	98.7	92.3	82.7	59.3
GCL [56]	CVPR-22	✗	✓	-	-	79.6	-
S3R [53]	ECCV-22	✗	✓	-	-	80.5	-
SSL [49]	ECCV-22	✗	✓	99.0	92.2	84.3	-
zxVAD [3]	WACV-23	✗	✓	96.9	-	71.6	-
HSC [45]	CVPR-23	✗	✓	98.1	93.7	83.4	-
FPDM [54]	ICCV-23	✓	✓	-	90.1	78.6	62.7
SLM [43]	ICCV-23	✓	✓	97.6	90.9	78.8	-
STG-NF [18]	ICCV-23	✗	✓	-	-	85.9	71.8
AnomalyRuler-base	-	✓	✗	96.5	82.2	84.6	69.8
AnomalyRuler	-	✓	✗	97.9	89.7	85.2	71.9

Compare with state-of-the-art
traditional VAD models

Two most challenging
datasets

MLLM for Video Affective Reasoning

- One of the first methods for VAR, i.e., predicting and **reasoning** viewers' emotional reactions to a video
- Propose **spatiotemporal stimuli-aware mechanisms** and create **VAR visual instruction data**

StimuVAR: Spatiotemporal Stimuli-aware Video
Affective Reasoning with Multimodal Large
Language Models

Yuxiang Guo^{1*†}, Faizan Siddiqui², Yang Zhao¹,
Rama Chellappa^{1*}, Shao-Yuan Lo^{2*}

¹Johns Hopkins University.

²Honda Research Institute USA.

Our Goal

- VAR aims to predict and explain viewers' emotional responses to a video.
- **What** are the emotional stimuli?
 - Capturing stimuli could reduce redundancy and improve affective understanding.
 - => Propose **spatiotemporal stimuli-aware mechanisms**
- **Why** this prediction?
 - Interpretability is crucial for model analysis and earning public trust.
 - => Create **VAR visual instruction data** for **affective training**

Method

- Let's start from an example video

What we see:

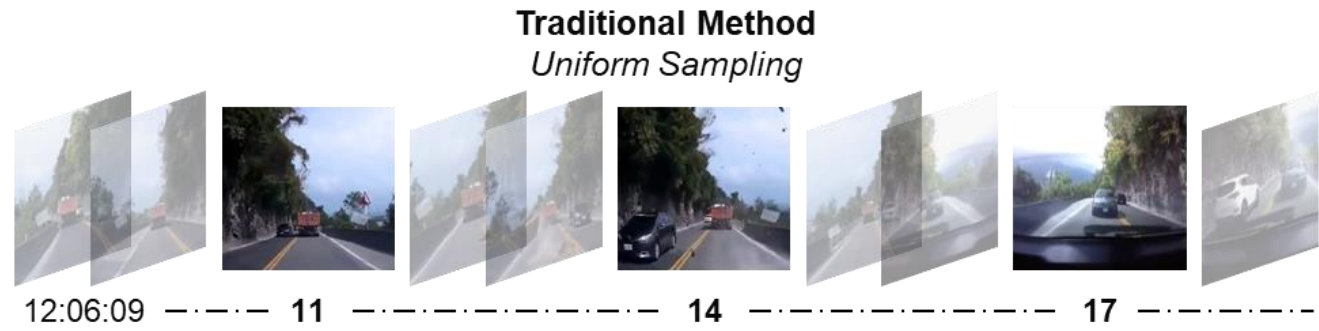


What models see:



Method

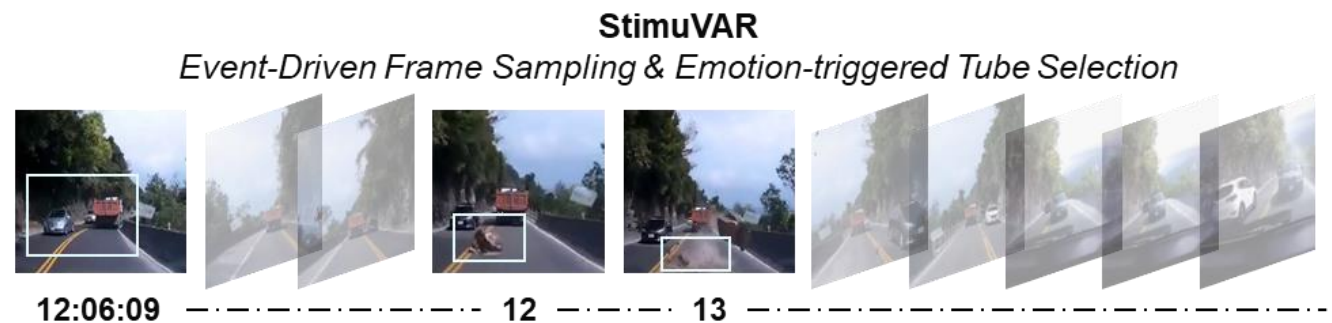
- Traditional models and generic MLLMs often overlook emotional stimuli.



User: What is the emotion of the viewers when they watch the video?

Answer: **Boredom** 🙄🙄🙄

- Our method has stimuli awareness.

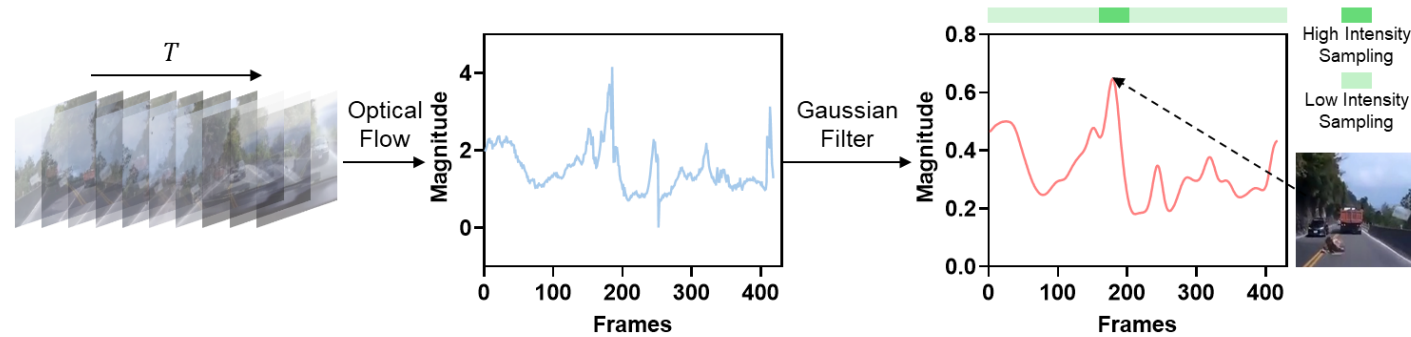


User: What is the emotion of the viewers when they watch the video?

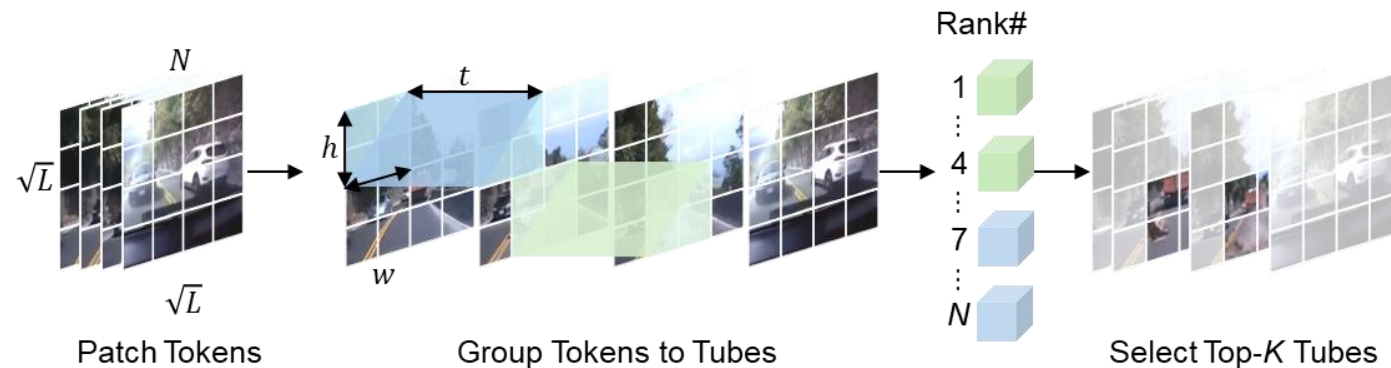
Answer: The viewer feels **Surprise** 😲 because initially, it is a typical road scene with a red truck and a silver car sets the expectation for a mundane video. However, the sudden appearance of a rock, falling onto the road introduces an unexpected element. This unexpected twist in the video content is likely to evoke surprise in the viewer

Spatiotemporal Stimuli-aware Mechanisms

- **Event-driven frame selection:** Frame-level awareness



- **Emotion-triggered tube selection:** Token-level awareness

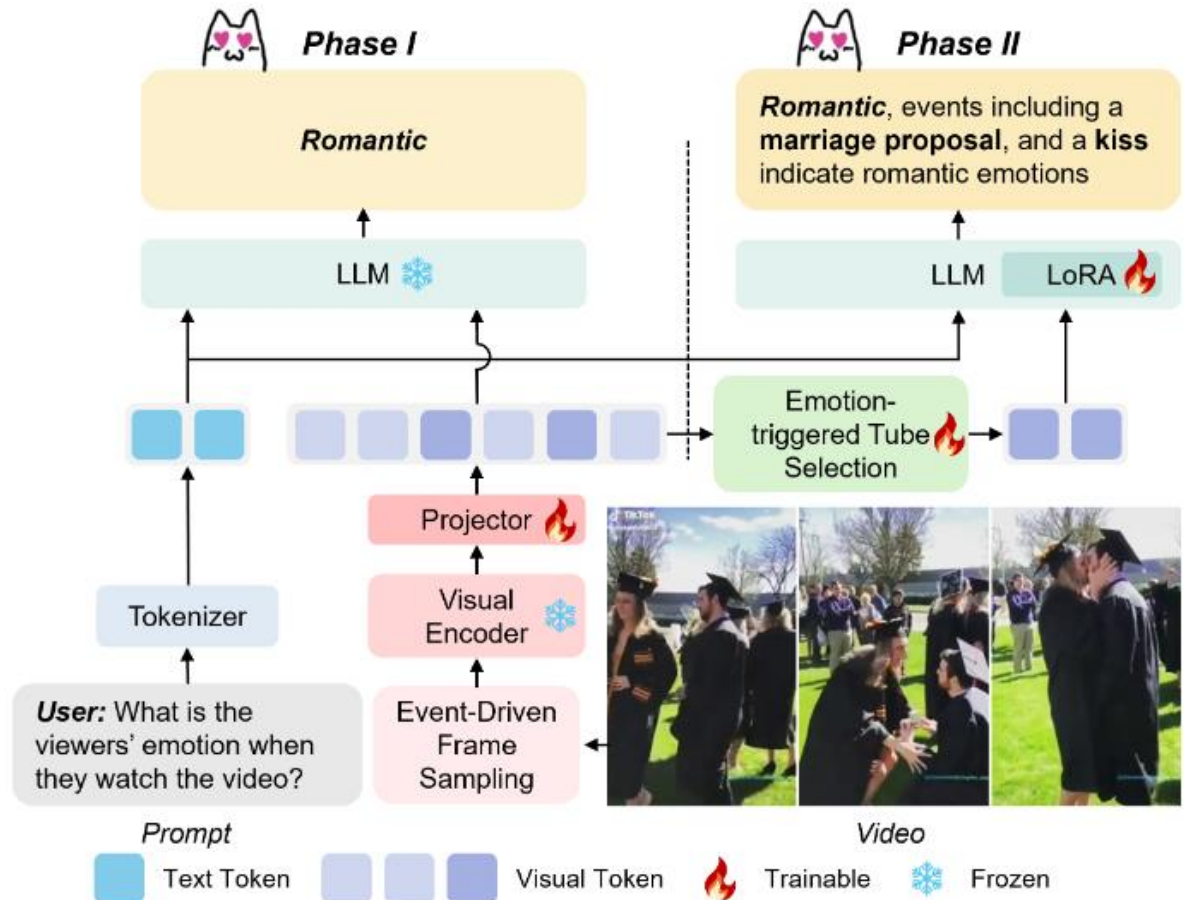


Affective Training

- Create VAR visual instruction data

```
{ "role": "system",  
  "content": "Given the below (QUESTION, ANSWER) pair  
  examples of emotion estimation, left fill-in the  
  REASONING process which derives ANSWERS from QUESTIONS  
  in three sentences.",  
  "role": "user",  
  "content": "QUESTION: These are frame descriptions from a  
  video. After reading the descriptions, how people  
  might emotionally feel about the content and why. Only  
  provide the one most likely emotion. <Video Caption>  
  ANSWER: The viewer feels <Emotion>.  
  REASONING: Let's think of step-by-step
```

- **Phase I:** Align affective visual information with LLM space.
- **Phase II:** VAR visual instruction tuning for affective reasoning.



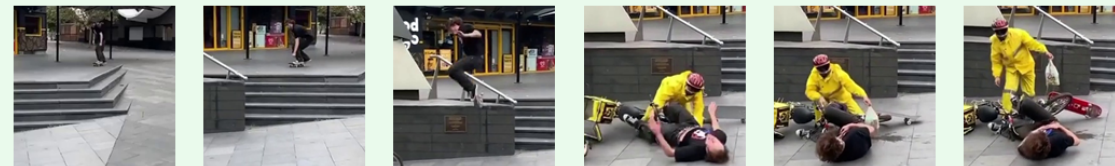
Results

- MLLM backbone: CLIP ViT + Llama2-7b

Table 1: Quantitative comparison on the VCE dataset.

Method	Venue	Top-3	Emo-align	RR	RW	WR	WW	CLIP-S
Traditional								
CLIP [55]	ICML'21	28.4	-	-	-	-	-	-
Majority [13]	NeurIPS'22	35.7	-	-	-	-	-	-
R(2+1)D [56]	CVPR'18	65.6	-	-	-	-	-	-
STAM [57]	arXiv'21	66.4	-	-	-	-	-	-
VideoMAE [58]	NeurIPS'22	68.9	-	-	-	-	-	-
MM-VEMA [16]	PRCV'23	73.3	-	-	-	-	-	-
MLLM								
Video-LLaMA [17]	EMNLP'23	26.4	25.5	16.2	9.3	9.3	65.2	63.9
Video-LLaVA [18]	arXiv'23	25.0	31.2	17.5	7.5	13.7	61.3	70.6
Valley [19]	arXiv'23	31.3	29.4	19.2	12.1	10.2	58.5	69.4
Video-ChatGPT [20]	ACL'24	21.0	29.5	11.4	9.5	18.1	61.0	68.9
VideoChat2 [21]	CVPR'24	31.1	36.4	24.0	7.1	12.4	56.5	68.6
Chat-UniVi [22]	CVPR'24	38.6	29.5	21.0	17.6	8.5	52.1	70.2
mPLUG-Owl [23]	CVPR'24	23.6	22.1	13.8	9.7	8.3	66.5	69.3
EmoViT [31]	CVPR'24	10.5	5.2	4.8	5.7	0.4	89.1	48.9
StimuVAR (Ours)		73.5	69.6	68.8	4.7	0.8	25.6	75.3

(A) GT: Empathic Pain



StimuVAR: The viewer feels *Empathic Pain*, because the description of a skateboarder performing a trick and then being involved in an accident with a bicyclist evokes a sense of empathy towards the individual's pain and potential injury. The presence of the aftermath of the accident further intensify the emotional response, leading the viewer to feel a deep sense of sympathy and distress. The contrast between the skateboarding action and the subsequent accident creates a stark emotional shift, eliciting a strong feeling of empathic pain in the viewer.

Chat-UniVi: The viewer might feel *Excitement* as they watch the skateboarder perform tricks and stunts. The sense of thrill and anticipation builds as the skateboarder gains speed and momentum, and the viewer waits to see what trick will be performed next.

Video-LLaMA: The most likely emotion of the viewers is *Excitement*. There are various skateboarders performing different tricks, such as one man riding a skateboard downstairs while another man is riding a skateboard on a street. The viewers will likely be excited to see the thrilling and dynamic performances.

Video-ChatGPT: The viewer feels a sense of *Relief*, as the man in the yellow jacket is seen helping the person on the skateboard.

Video-LLaVA: The most likely emotion from the provided emotions is *Relief* because the man is helping the injured person, and the situation seems to be under control.

Relevant Irrelevant

MLLM for Action Anticipation

- One of the first MLLM-based method for action anticipation
- Propose Plausible Action Sequence Learning Loss and Long-Horizon Action Repetition Loss for **plausible** and **diverse** predictions

Can't make an Omelette without Breaking some Eggs: Plausible Action Anticipation using Large Video-Language Models

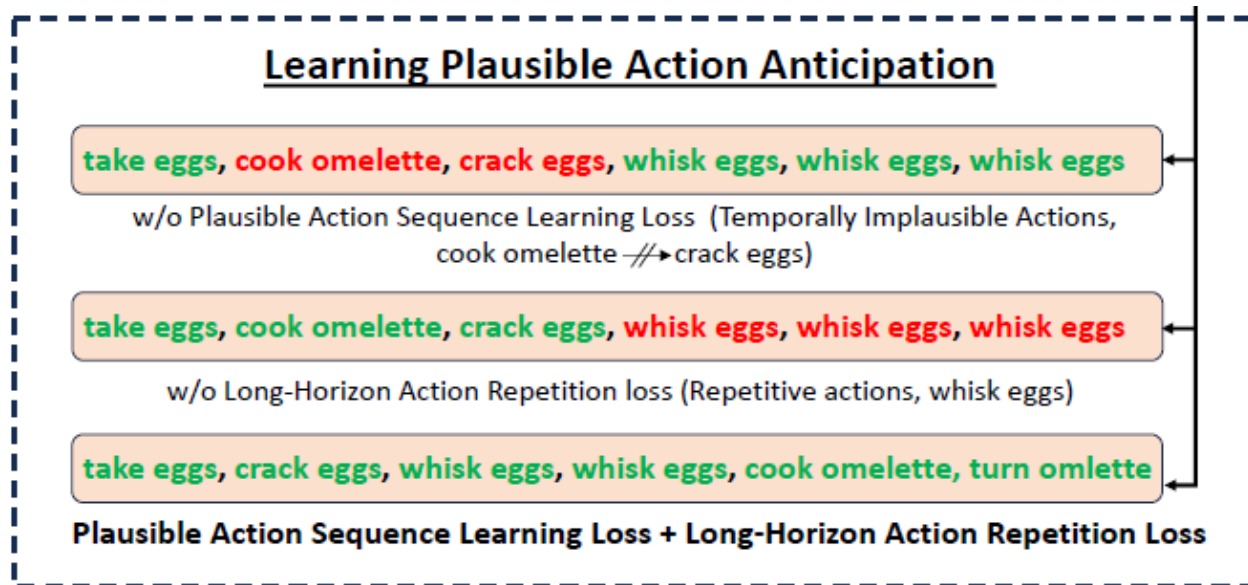
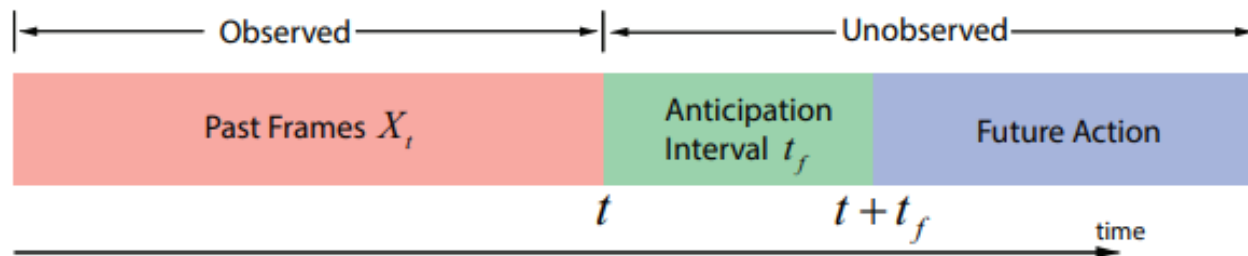
Himangi Mittal^{1,2*} Nakul Agarwal¹ Shao-Yuan Lo¹ Kwonjoon Lee¹

¹Honda Research Institute USA ²Carnegie Mellon University

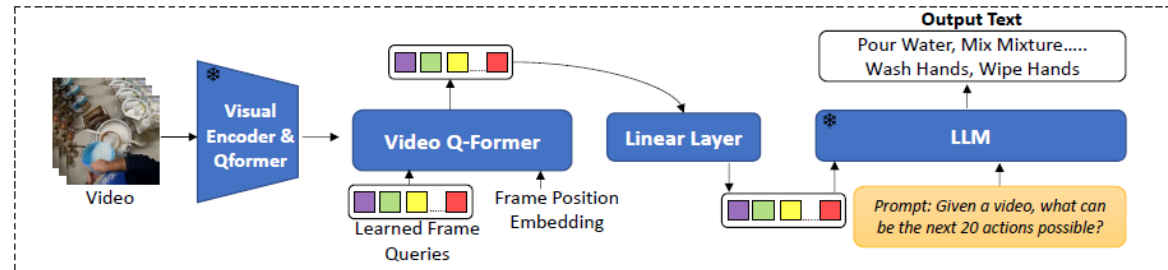
hmittal@andrew.cmu.edu {nakul_agarwal, shao-yuan_lo, kwonjoon_lee}@honda-ri.com

Our Goal

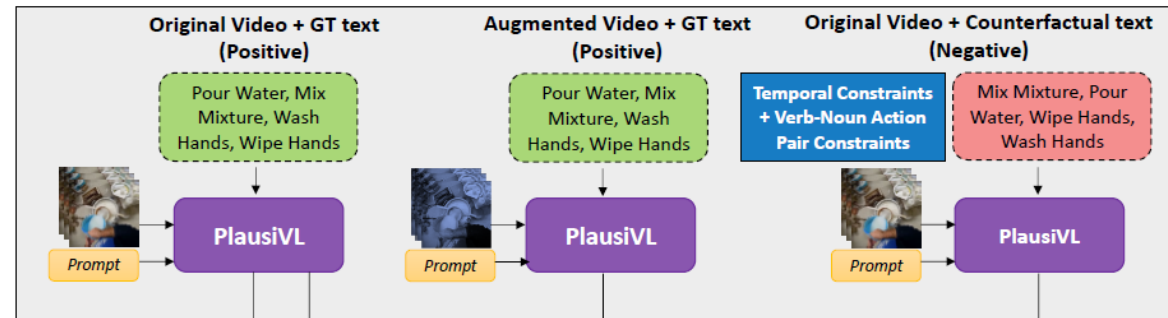
- Action anticipation aims to predict future actions given previous actions.
- **Plausible** predictions by Plausible Action Sequence Learning Loss
- **Diverse** predictions by Long-Horizon Action Repetition Loss



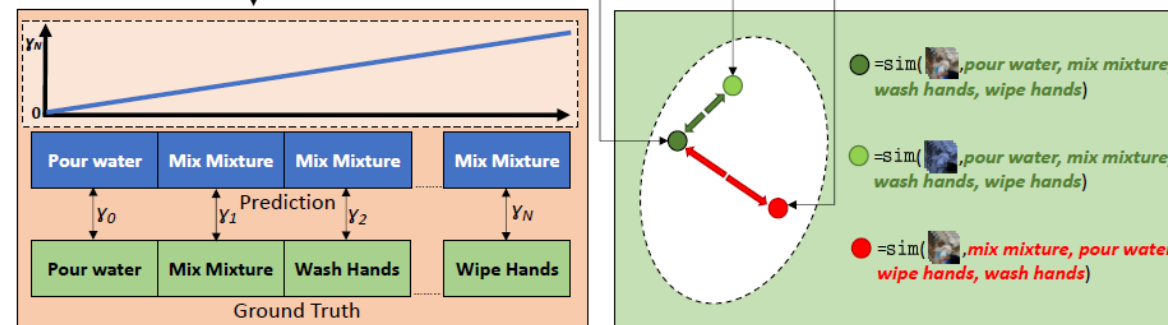
Method



(a) PlausiVL



(b) Augmentation



Long-Horizon Action Repetition Loss

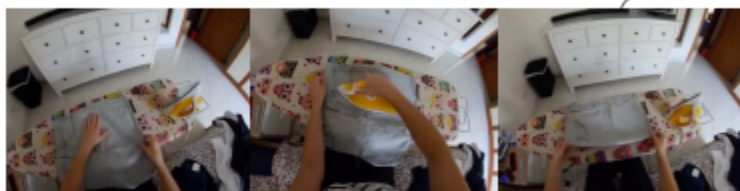
Plausible Action Sequence Learning Loss

(c) Objective Functions and Training

Results

- MLLM backbone: CLIP ViT + Llama2-7b
- Dataset: EPIC-Kitchens-100

Method	Class-mean Top-5 recall (%) ↑		
	Verb	Noun	Action
RU-LSTM [13]	23.20	31.40	14.70
Temporal Aggregation [56]	27.80	30.80	14.00
Video LLM [6]	-	-	15.40
AFFT [73]	22.80	34.60	18.50
AVT [25]	28.20	32.00	15.90
MeMViT [68]	32.20	37.00	17.70
RAFTformer [24]	33.80	37.90	19.10
InAViT [55]	52.54	51.93	25.89
Video LLaMA [71]	52.90	52.01	26.05
PlausiVL	55.62	54.23	27.60



Video

Prediction: take iron, take pants, put pants, adjust pants, take iron, press pants, put iron, adjust pants, take iron, press pants, turn pants, adjust pants, take iron, press pants, put iron, adjust pants, take iron, turn pants, put iron, adjust pants

Ground Truth: take iron, press pants, hold iron, press pants, put iron, take iron, press pants, turn pants, arrange pants, take iron, press pants, adjust pants, turn pants, arrange pants, take iron, turn pants, put pants, touch pants, take pants, fold pants

Time

<https://shaoyuanlo.github.io/>

sylo@jhu.edu