Anomaly Detection in the Era of Multimodal Large Language Models

Shao-Yuan Lo

Research Scientist @ Honda Research Institute USA

4/15/2025 @ UC Santa Cruz

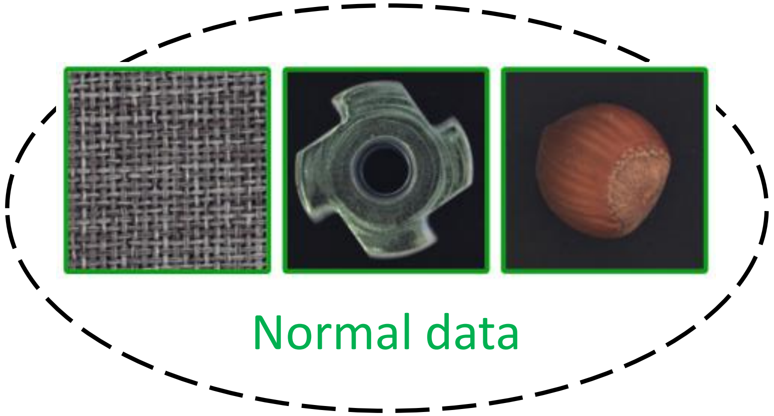
About Me

- Research Scientist @ **Honda Research Institute USA**
  San Jose, CA (2023 - Present)
- Research Intern @ **Amazon**
  Seattle, WA (Summer 2021 & 2022)

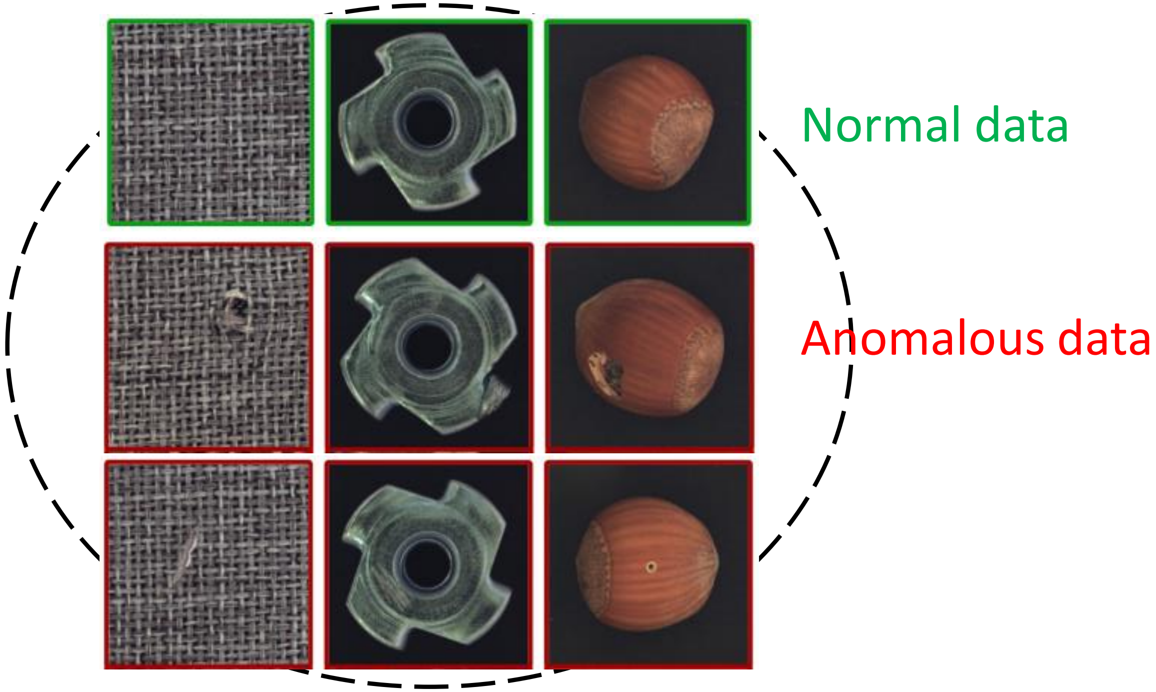
- PhD in ECE @ **Johns Hopkins University**
  Baltimore, MD (2019 - 2023)
- MS in EE @ **National Chiao Tung University**
  Taiwan (2017 - 2019)
- BS in EECS @ **National Chiao Tung University**
  Taiwan (2013 - 2017)

What is Anomaly Detection?

- **Problem definition**: An AD model is exclusively trained with **normal** data and is asked to identify whether a query example is **normal** or **anomalous**.
- **Motivation**: Anomalies are often rare and long-tailed, so they are costly to collect.
- Example:
  - **Normal data**: Flawless objects
  - **Anomalous data**: Defects



Training set



Test set

Image source:
[Bergmann et al. CVPR'19]

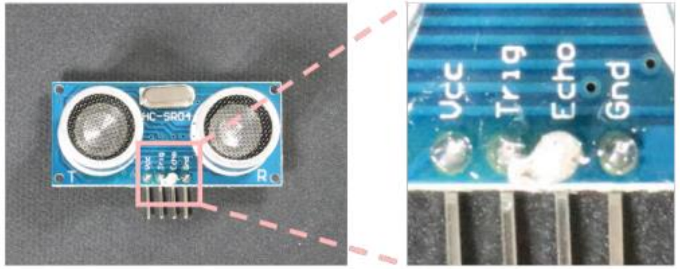
Visual Anomaly Detection: Images and Videos

- Visual AD is a long-established problem in computer vision.
- Given its practical significance, AD has been widely deployed in various applications.

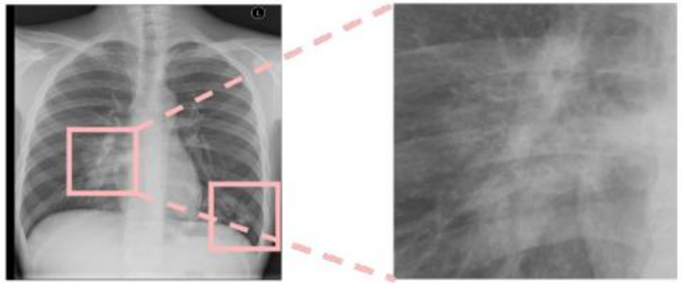
Video Anomaly Detection (VAD)

Security surveillance

Image Anomaly Detection (IAD)



Industrial image inspection



Medical image diagnosis

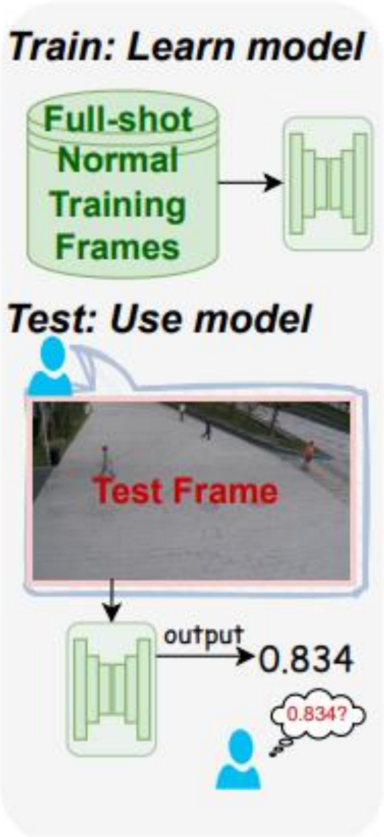
Image source: [Bergmann et al. CVPR’19] [Bao et al. CVPR’24] [Liu et al. CVPR’18]

Conventional Learning-based AD Approaches

- **Full-shot training**: An AD model is trained by a large amount of normal data to learn normal patterns

- **Output format**: Anomaly scores -> Thresholding

- **Metrics**: AUROC (area under ROC curve)



The Era of Multimodal Large Language Models

ChatGPT

Claude

deepseek

InternVL

MISTRAL AI_

Gemini

Qwen-VL

How can AD benefit from MLLMs?

What breakthroughs can MLLMs bring to this long-established vision problem?

Remain underexplored!

Anomaly Detection in the Era of MLLMs

	Conventional		**MLLMs**
How to learn	Full-shot training	→	Few-shot / Zero-shot
What to output	Detection only (anomaly scores)	→	Detection + Reasoning

How to learn — MLLMs: Few-shot / Zero-shot

- Prompting pipelines
- Model adaptability

What to output — MLLMs: Detection + Reasoning

- Instruction tuning data
- New metrics
- New benchmarks

Anomaly Detection in the Era of MLLMs

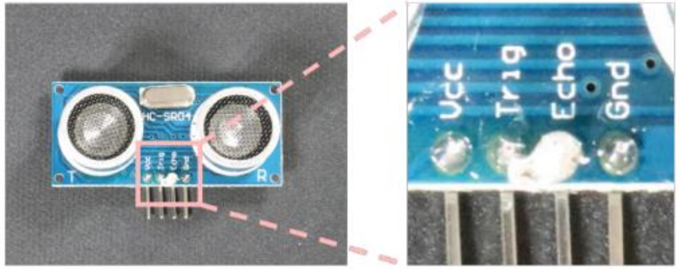
ECCV 2024

CVPR 2025

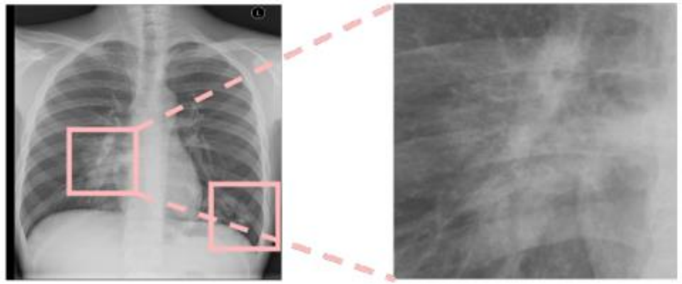
Video Anomaly Detection (VAD)

Security surveillance

Image Anomaly Detection (IAD)



Industrial image inspection



Medical image diagnosis

Image source: [Bergmann et al. CVPR’19] [Bao et al. CVPR’24] [Liu et al. CVPR’18]

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

Yuchen Yang[1*], Kwonjoon Lee[2], Behzad Dariush[2], Yinzhi Cao[1], and Shao-Yuan Lo[2]

[1] Johns Hopkins University
{yc.yang, yinzhi.cao}@jhu.edu
[2] Honda Research Institute USA
{kwonjoon_lee, bdariush, shao-yuan_lo}@honda-ri.com

ECCV 2024

- 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

Problem Statement

- **Assumption**: We only have a few **normal** data for our specific application, and it's costly to collect **anomaly** data.
- **Challenge**: The definition of "**anomaly**" **depends on different context and downstream applications**.
- **Goal**: Develop a VAD model for our specific application (specific definition of "**normal**" & "**anomaly**") and explain the detection results.

person jogging versus person running outside a bank.

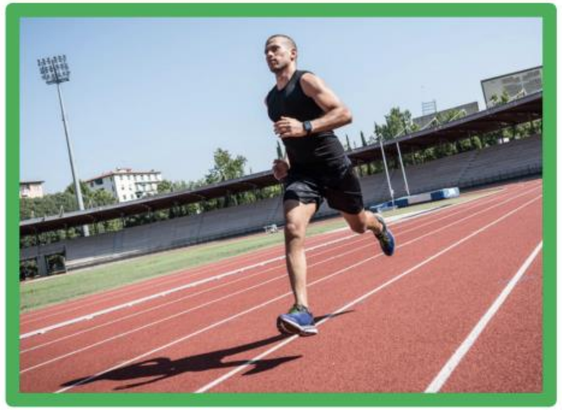
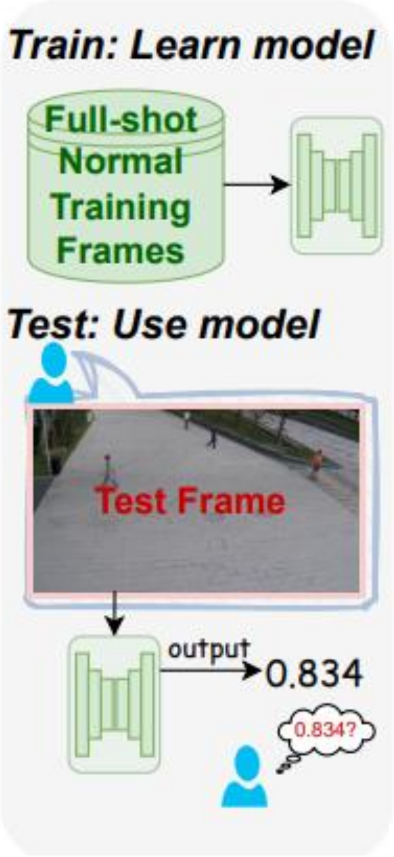


Image source: [Radu Tudor Ionescu. CVPR'23 Tutorial]

Conventional Learning-based Approaches

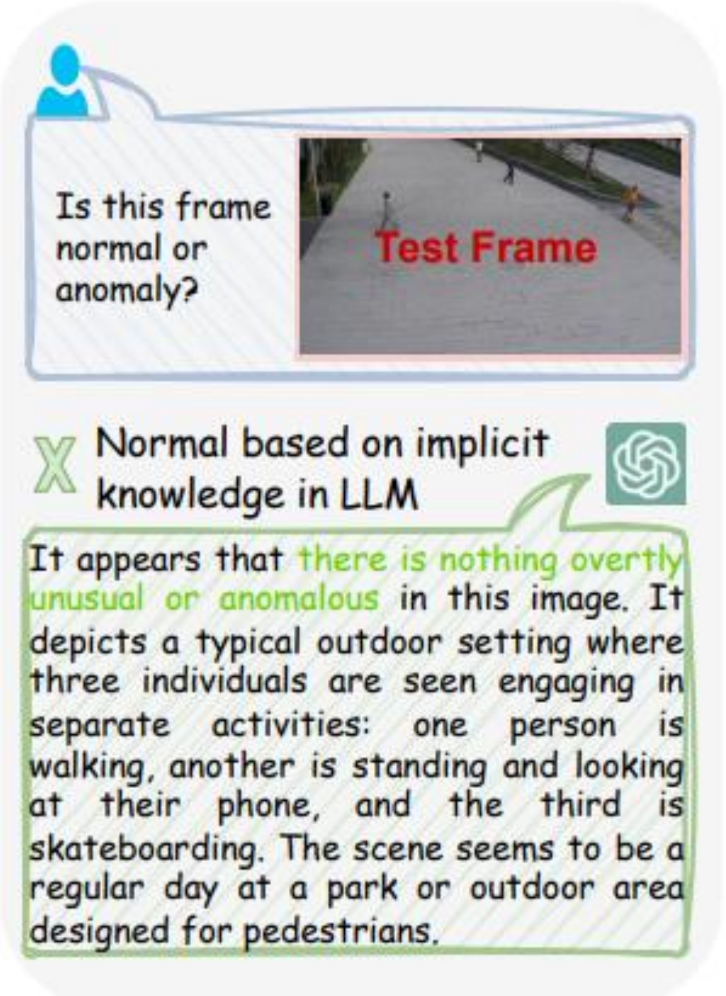
- **Full-shot training**: A VAD model is trained by a large amount of normal data to learning normal patterns
- **Output format**: Anomaly scores -> Thresholding
- **Metrics**: AUROC (area under ROC curve)



Query MLLMs Directly

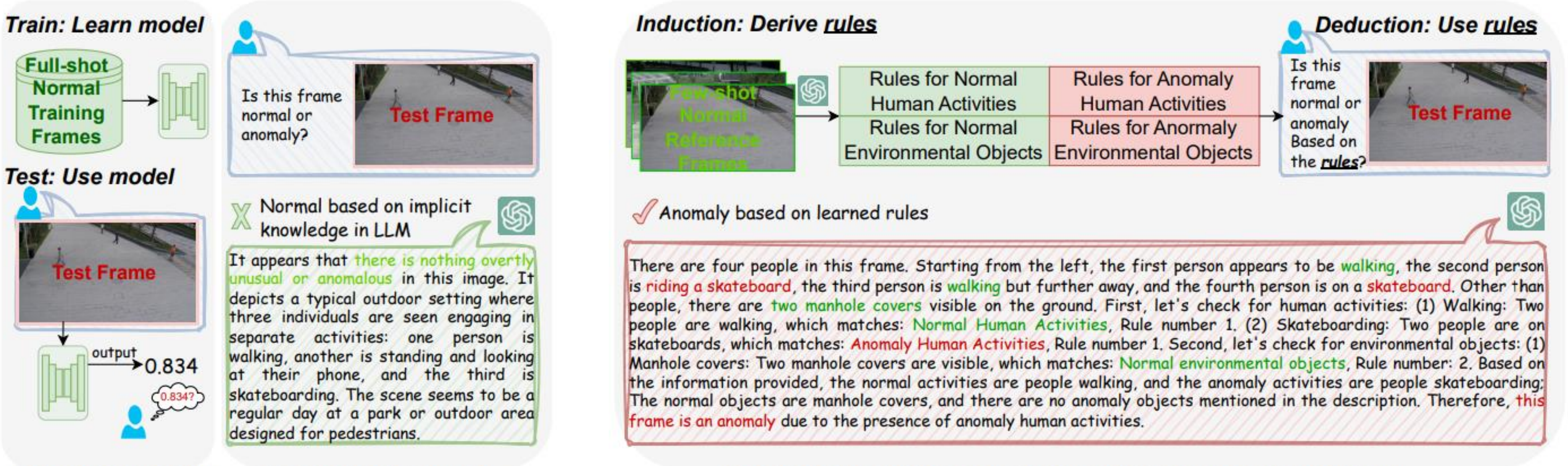
- The implicit knowledge pre-trained in MLLMs may not align with specific VAD needs.

- Here GPT-4V mistakenly treats "skateboarding" as normal.



Method

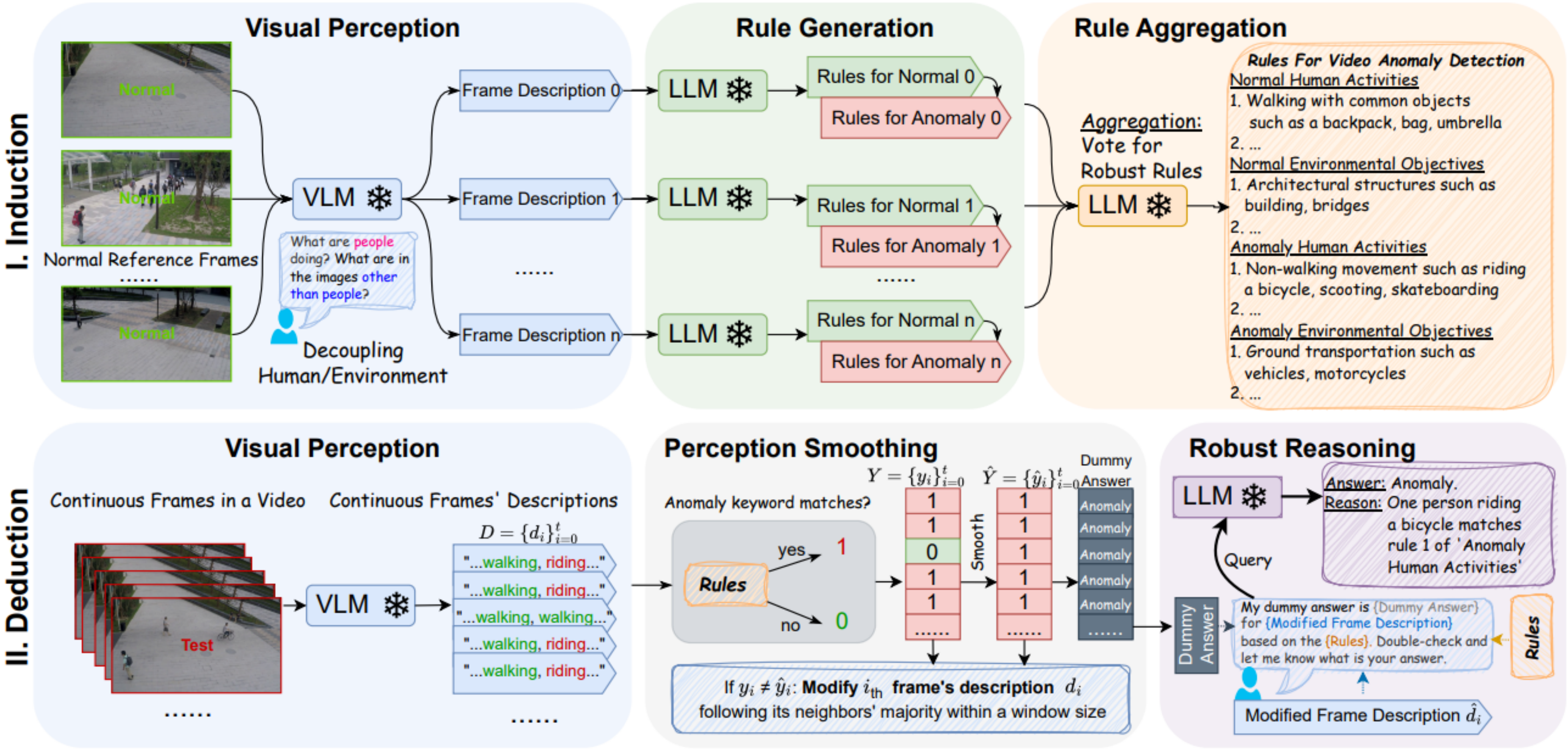
- **Induction**: Learn rules from few-shot normal reference frames
  - Few-normal shot prompting (no training needed)
- **Deduction**: Detect anomalies based on the rules
  - Correctly identifying "skateboarding" as an anomaly



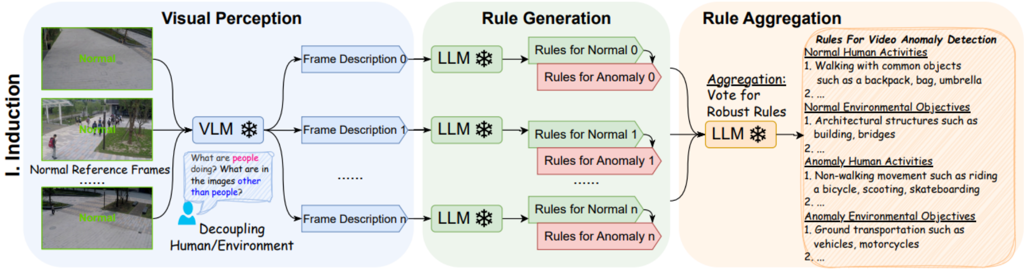
a) Traditional VAD b) Ask LLM directly c) Ask LLM with the rules learned in induction stage (Ours)

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.



Example (Induction)



A few normal reference frames

```
**Rules for Anomaly Human Activities:**
1. Using any non-walking movement such as riding a bicycle or scooter,
    skateboarding
2. Running or jumping
3. Moving with the usual speed such as very fast or in a hurry
4. Person lying or bending down on the ground
5. Aggressive or unusual movements such as fighting or pushing
6. Loitering around the trash bin
```

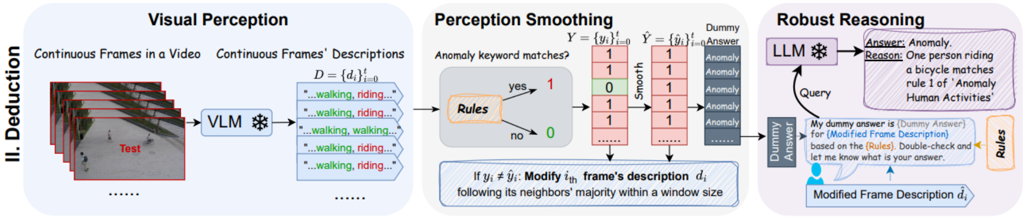
```
**Rules for Anomaly Environmental Objects:**
1. Missing, removed, or ajar manhole covers
2. Ground transportation such as vehicles, vans, bicycles.
3. Unattended bags or objects (security threat)
4. Unusual objects (vehicles, large unattended luggage, etc.)
5. Unauthorized posters or signs
```

```
**Rules for Normal Human Activities:**
1. Walking alone or with another person
2. Standing and using a mobile device
3. Walking on the sidewalk
4. Walking towards or away from a building
5. Entering a building
6. Standing near a trash bin
7. Waiting at traffic lights
```

A set of rules

```
**Rules for Normal Environmental Objects:**
1. Manhole covers on the ground or street
2. Trees along the street or Plants present
3. Trash bin in vicinity or placed beside the street
4. Posters on glass door
```

Example (Deduction)



Query samples

$\hat{y}_i^*$ = The third person is riding a bicycle, which matches rule number 1 under Anomaly Human Activities.

The bicycle matches rule number 2 under Anomaly Environmental Objects.

Answer: anomaly.

$\hat{y}_j^*$ = The human activities present are two people walking. The rules for normal human activities include walking alone or with another person. So, the rule matching is:
1. Walking alone or with another person (normal)

The environmental objects present are two manhole covers on the ground, a trash bin, and some trees and plants. So, the rules matching are:
1. Manhole covers on the ground or street (normal)
2. Trees along the street or Plants present (normal)
3. Trash bin in vicinity or placed beside the street (normal)

Answer: normal.

Output reasoning

Comparison with Conventional VAD Models

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

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**

Two most challenging datasets

Comparison with LLM-based Approaches

- **Induction**: CogVLM-17B & GPT-4
- **Deduction**: CogVLM-17B & Mistral-7B
- **Metrics**: Accuracy/precision/recall; Doubly-right (RR/RW/WR/WW)

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

Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models

Jiacong Xu[1*] Shao-Yuan Lo[2] Bardia Safaei[1] Vishal M. Patel[1] Isht Dwivedi[2]
[1]Johns Hopkins University [2]Honda Research Institute USA
{jxu155, bsafaei1, vpatel36}@jhu.edu {shao-yuan_lo, idwivedi}@honda-ri.com

CVPR 2025 (Highlight)

- One of the **first zero-shot and reasoning** methods for IAD
  - => Explain why normal/anomaly
- One of the **first instruction tuning datasets and benchmarks** for IAD reasoning
  - => Allow us to train and evaluate models for their AD reasoning capabilities

Problem Statement

Complex reasoning

In-the-Wild Example: Graffitied Road Sign	
	
User	Describe the anomaly in the image.
Anomaly-OV	The anomaly in the image is the graffiti on the highway sign, which includes black spray paint marks and scribbles over the word 'Seattle' and other parts of the sign.
User	What is the reason for the anomaly?
Anomaly-OV	The graffiti is likely the result of vandalism, where individuals have used spray paint to deface public property for personal expression or as a form of protest.
User	How does the anomaly affect the functionality of the object?
Anomaly-OV	The graffiti affects the sign's functionality by potentially obstructing the visibility of important information, such as the word 'Seattle,' which could confuse drivers and lead to navigational errors.
User	How to avoid the anomaly in the future?
Anomaly-OV	To avoid such anomalies, regular maintenance and cleaning of signs should be implemented. Additionally, community awareness programs can help reduce vandalism by promoting respect for public property.

Anomaly detection

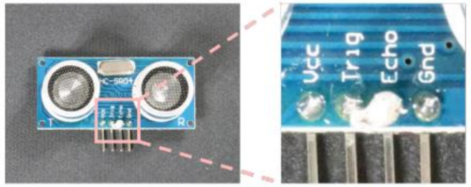
PCB Example: Excessive Solder	
	
User	Are there any anomalies visible in the image?
GPT-4o	The image shows an HC-SR04 ultrasonic sensor. There don't appear to be any visible anomalies with the sensor itself. The components and pins seem intact and properly aligned.
LLaVA-OV	The image shows an ultrasonic sensor module, specifically the HC-SR04 model. There are no visible anomalies in the image; it appears to be a standard representation of this type of sensor module.
Ours	Yes, there is a noticeable anomaly in the image: a short circuit on the PCB, indicated by a white, crusty residue near the 'Echo' pin.

Table 3. Anomaly-OV presents more accurate anomaly detection.

Basic reasoning

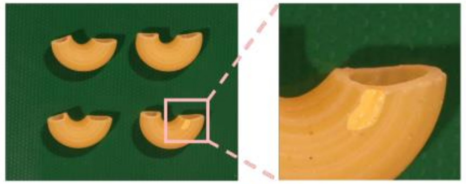
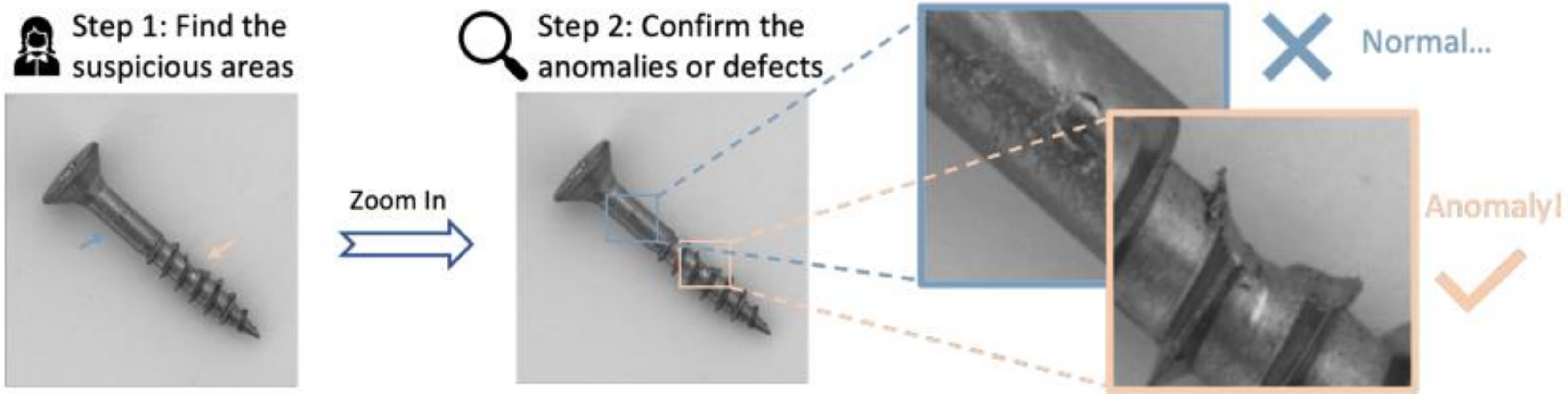
Macaroni Example: Yellowish Spot	
	
User	Do you see any anomalies in the image?
GPT-4o	The image shows four pieces of elbow macaroni on a green background. The anomaly is that the macaroni pieces are not whole; they are cut in half.
LLaVA-OV	The image shows four pieces of pasta, specifically macaroni shells, arranged on a green textured surface. The pasta appears to be uniformly colored and shaped, with no visible defects or anomalies.
Ours	Yes, there is an anomaly in the image. The bottom right pasta piece has a noticeable yellowish discoloration on its surface.

Table 5. Anomaly-OV presents more precise anomaly reasoning.

Method

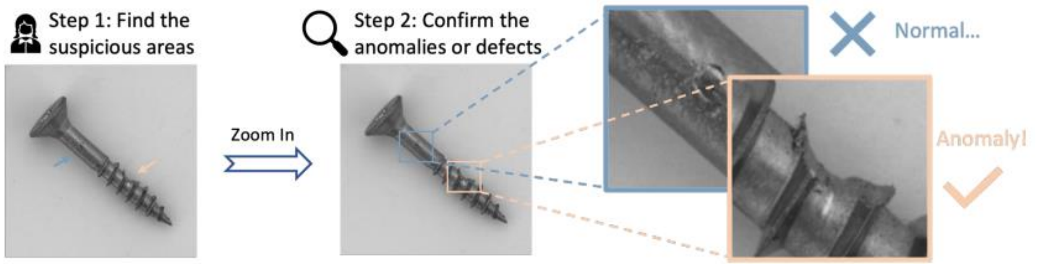
- Two-step inspection for improving detection
- Simulate human's visual anomaly inspection

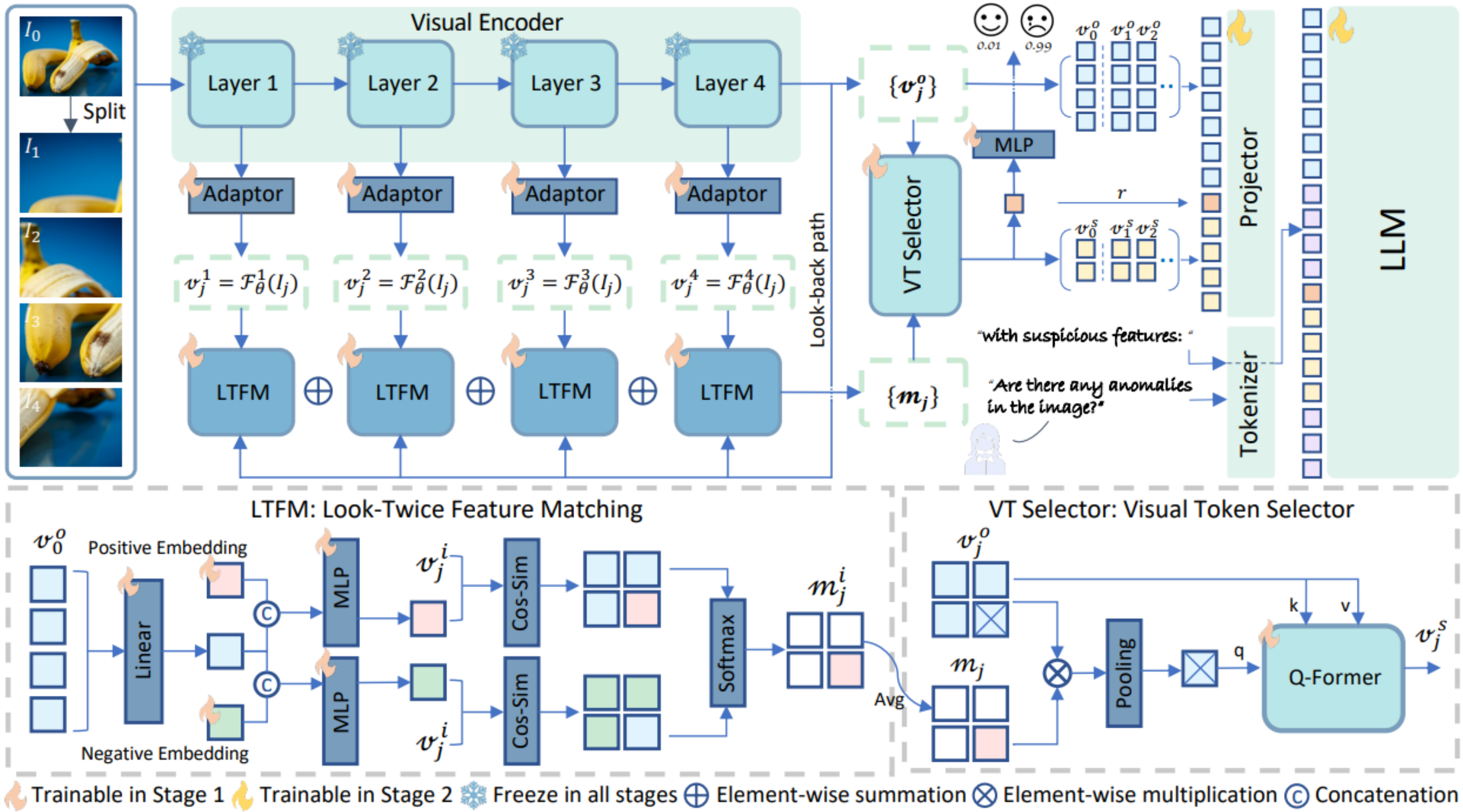


Method

- **MLLM backbone**: Visual encoder + tokenizer + projector + LLM
- **LTFM**: (Step 1) Find suspicious areas
- **VT Selector**: (Step 2) Confirm anomalies

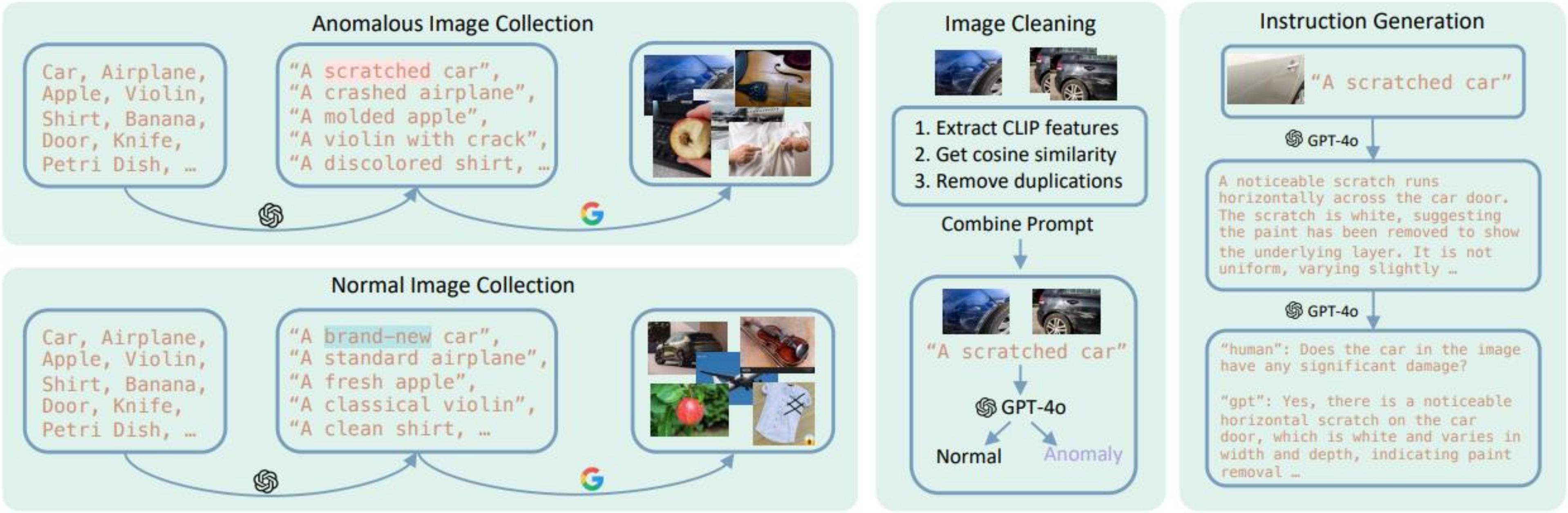
- **Training stage 1**: Train the anomaly expert (LTFM + VT Selector)
- **Training stage 2**: Visual instruction tuning (projector + LLM)





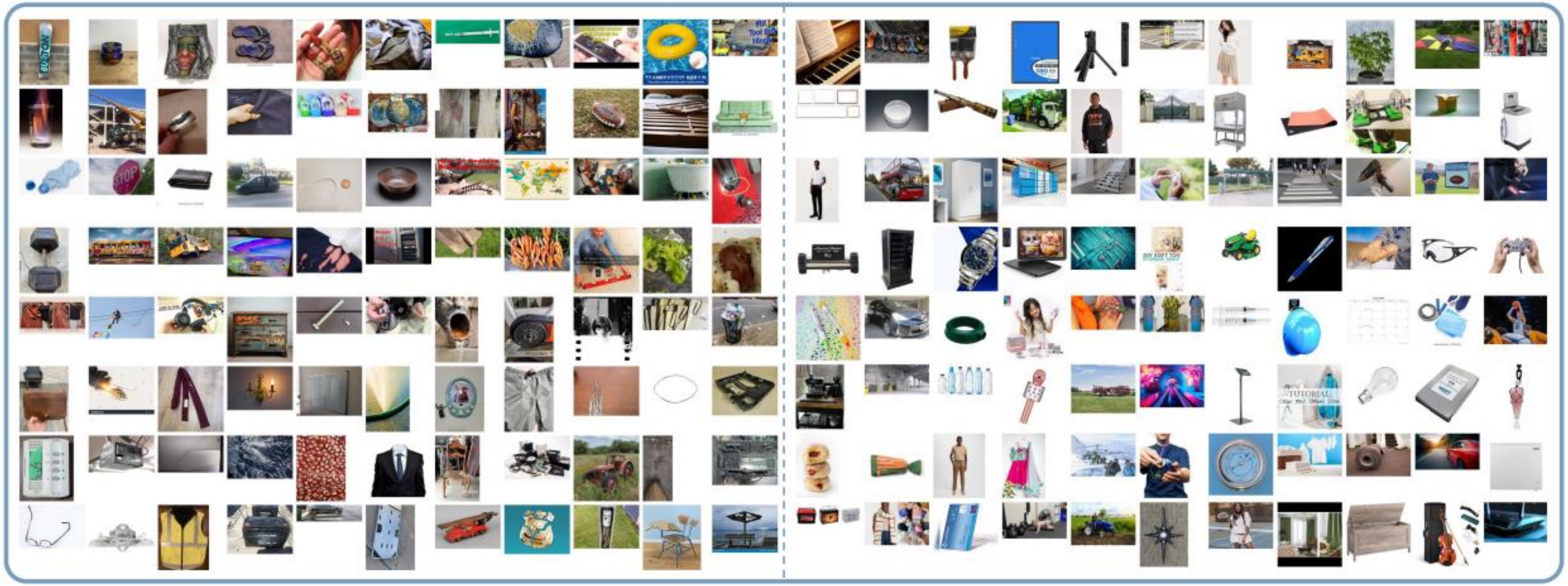
Instruction Tuning Dataset

- Visual instruction tuning (VIT) for improving reasoning
- An automatic pipeline for constructing AD VIT datasets
- **WebAD: 72k, 380 classes**
- MVTec AD [Bergmann et al. CVPR'19] : 5k, 15 classes



Instruction Tuning Dataset

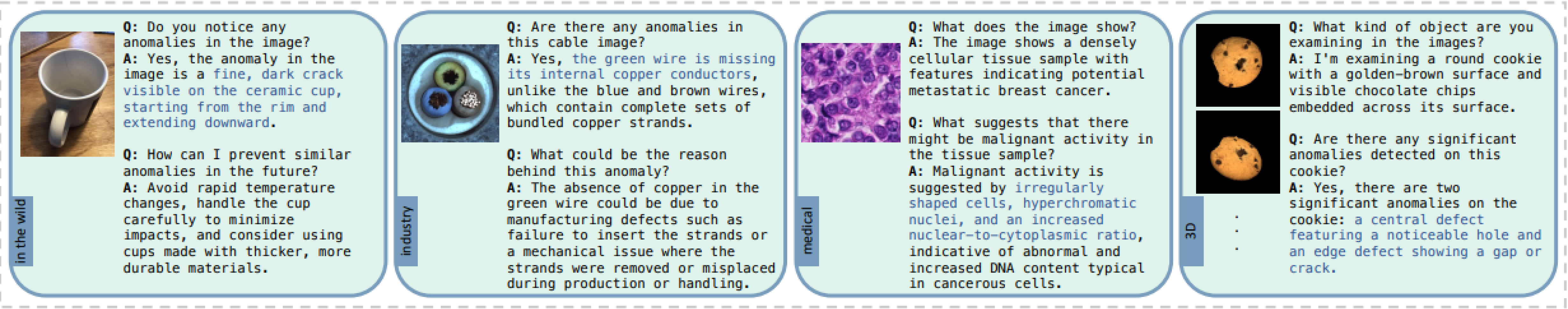
- **WebAD: 72k, 380 classes**



Anomalous Normal

Instruction Tuning Dataset

- **Anomaly-Instruct-125k**
- Four types of images



WebAD-72k	MVTec AD	BMAD	MVTec-3D AD
[Collected by us via Google Image Search]	[Bergmann et al. CVPR'19]	[Bao et al. CVPR'24]	Real3D-AD
			Anomaly-ShapeNet

[Bergmann et al. 2022] [Liu et al. NeurIPS'23] [Li et al. CVPR'24]

Reasoning Benchmark

- **VisA-D&R** (761 normal + 1000 anomaly samples)
- **Detection metrics**: Accuracy/precision/recall/F1-score
- **Reasoning metrics**: ROUGE-L, Sentence-BERT, GPT-score
- **Q1 & Q2**: Basic reasoning
- **Q3 & Q4**: Complex reasoning

```
Detection:
Q: Are there any defects for the object in the image?
Please reply with 'Yes' or 'No'.

Reasoning:
Q1: Do you observe any anomalies in the image?
Q2: Can you describe the anomalies you observed?
Q3: What is the potential cause for the anomalies?
Q4: How can such anomalies be prevented in the future?
```

Detection Results of Zero-Shot IAD Approaches

- MLLM backbone: LLaVA-OV [Li et al. 2024]
- Metric: AUROC

Model	Industrial Defects						Medical Anomalies			Average
	MVTec AD	VisA	AITEX	ELPV	BTAD	MPDD	BrainMRI	HeadCT	Br35H	
CLIP [73]	74.1	66.4	71.0	59.2	34.5	54.3	73.9	56.5	78.4	63.1
CoOp [108]	88.8	62.8	66.2	73.0	66.8	55.1	61.3	78.4	86.0	70.9
WinCLIP [38]	91.8	78.8	**73.0**	74.0	68.2	63.6	92.6	90.0	80.5	79.2
APRIL-GAN [11]	86.2	78.0	57.6	65.5	73.6	73.0	89.3	89.1	93.1	78.4
AnoVL [19]	<u>92.5</u>	79.2	<u>72.5</u>	70.6	80.3	68.9	88.7	81.6	88.4	80.3
AnomalyCLIP [110]	91.5	82.1	62.2	<u>81.5</u>	88.3	<u>77.0</u>	90.3	<u>93.4</u>	94.6	84.5
AdaCLIP [6]	89.2	<u>85.8</u>	64.5	79.7	<u>88.6</u>	76.0	**94.8**	91.4	**97.7**	<u>85.3</u>
Ours	**94.0**	**91.1**	72.0	**83.0**	**89.0**	**81.7**	<u>93.9</u>	**97.6**	<u>95.5</u>	**88.6**

Detection and Reasoning Results of MLLMs

- MLLM backbone: LLaVA-OV
- LLaVA-OV-0.5B*: Fine-tuned on our Anomaly-Instruct-125k

Model	Anomaly Detection				Low-level Reasoning			Complex Reasoning	
	Accuracy	Precision	Recall	F1-score	ROUGE-L	SBERT	GPT-Score	SBERT	GPT-Score
GPT-4V [71]	0.68	0.90	0.49	0.55	0.16	0.65	3.31	0.77	5.64
GPT-4o [72]	0.70	0.83	0.71	0.68	0.24	0.71	**4.84**	0.81	**6.89**
Qwen2-VL-2B [87]	0.65	0.87	0.55	0.59	0.22	0.55	1.94	0.74	4.26
Qwen2-VL-7B [87]	0.76	<u>0.91</u>	0.69	0.75	0.25	0.61	3.09	0.68	4.62
InternVL-2-8B [13]	0.74	0.78	0.81	0.76	0.23	0.73	3.69	0.80	5.08
InternVL-2-26B [13]	0.73	0.86	0.66	0.68	0.21	**0.74**	4.13	0.80	5.49
IXC-2.5-7B [101]	0.72	0.88	0.63	0.67	0.21	0.58	2.45	0.77	5.14
LLaVA-OV-0.5B [44]	0.54	0.70	0.19	0.28	0.20	0.63	2.54	0.81	4.34
LLaVA-OV-7B [44]	0.71	**0.95**	0.56	0.63	0.24	0.66	3.57	0.79	5.44
LLaVA-OV-0.5B*	0.71	0.77	<u>0.84</u>	0.76	0.31	0.70	3.69	0.82	5.31
Anomaly-OV-0.5B	**0.79**	0.86	0.83	<u>0.82</u>	<u>0.33</u>	0.72	3.87	<u>0.83</u>	5.67
Anomaly-OV-7B	**0.79**	0.83	**0.86**	**0.83**	**0.34**	<u>0.73</u>	<u>4.26</u>	**0.84**	<u>6.34</u>

Summary

ECCV 2024

Video Anomaly Detection (VAD)

CVPR 2025

Image Anomaly Detection (IAD)

	Conventional		**MLLMs**
How to learn	Full-shot training	→	Few-shot / Zero-shot
What to output	Detection only (anomaly scores)	→	Detection + Reasoning

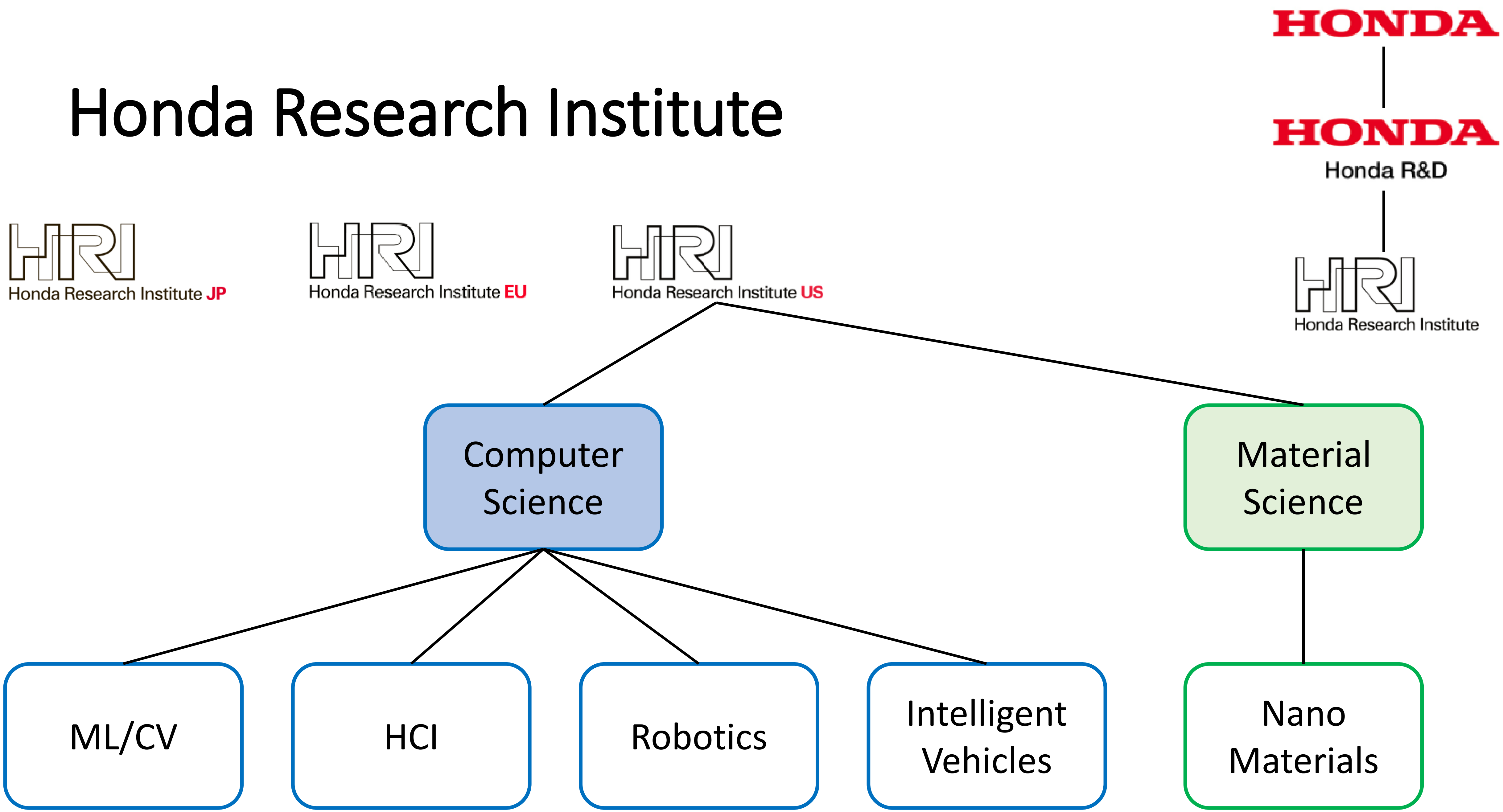
Few-shot / Zero-shot:

- Prompting pipelines
- Model adaptability

Detection + Reasoning:

- Instruction tuning data
- New metrics
- New benchmarks

Honda Research Institute



Current Research in HRI ML/CV Team

- **Goal**: Adapt general-purpose MLLMs to domain experts for applications on vehicles or robots.

- MLLM for anomaly detection
- MLLM for action anticipation
- MLLM for affective understanding
- MLLM for Theory-of-Mind
- Data-efficient MLLM fine-tuning

MLLM for Action Anticipation

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

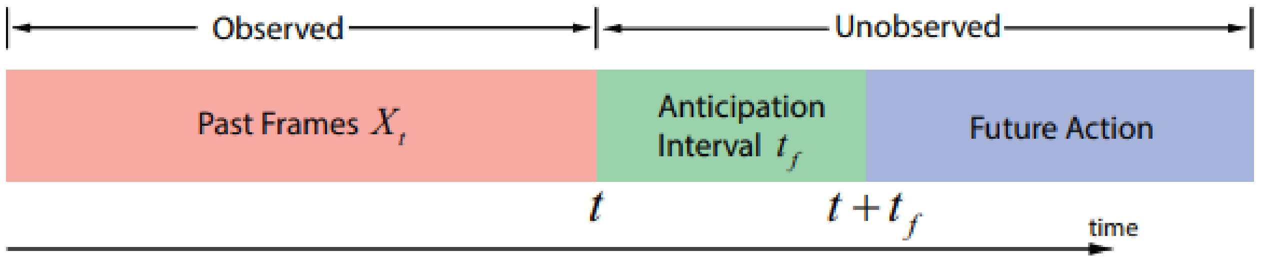
Himangi Mittal[1,2*] Nakul Agarwal[1] Shao-Yuan Lo[1] Kwonjoon Lee[1]
[1]Honda Research Institute USA [2]Carnegie Mellon University
hmittal@andrew.cmu.edu {nakul_agarwal, shao-yuan_lo, kwonjoon_lee}@honda-ri.com

CVPR 2024

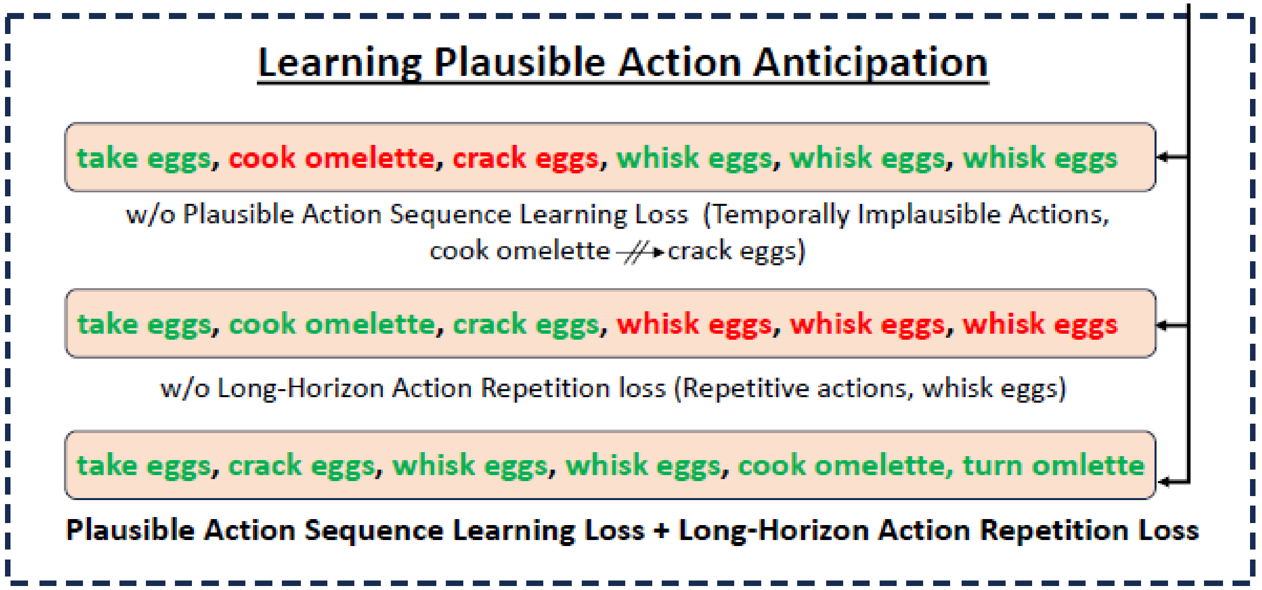
- One of the first MLLM-based method for action anticipation
- Propose Plausible Action Sequence Learning Loss and Long-Horizon Action Repetition Loss for **plausible** and **diverse** predictions

Goal

- Action anticipation aims to predict future actions given previous actions.

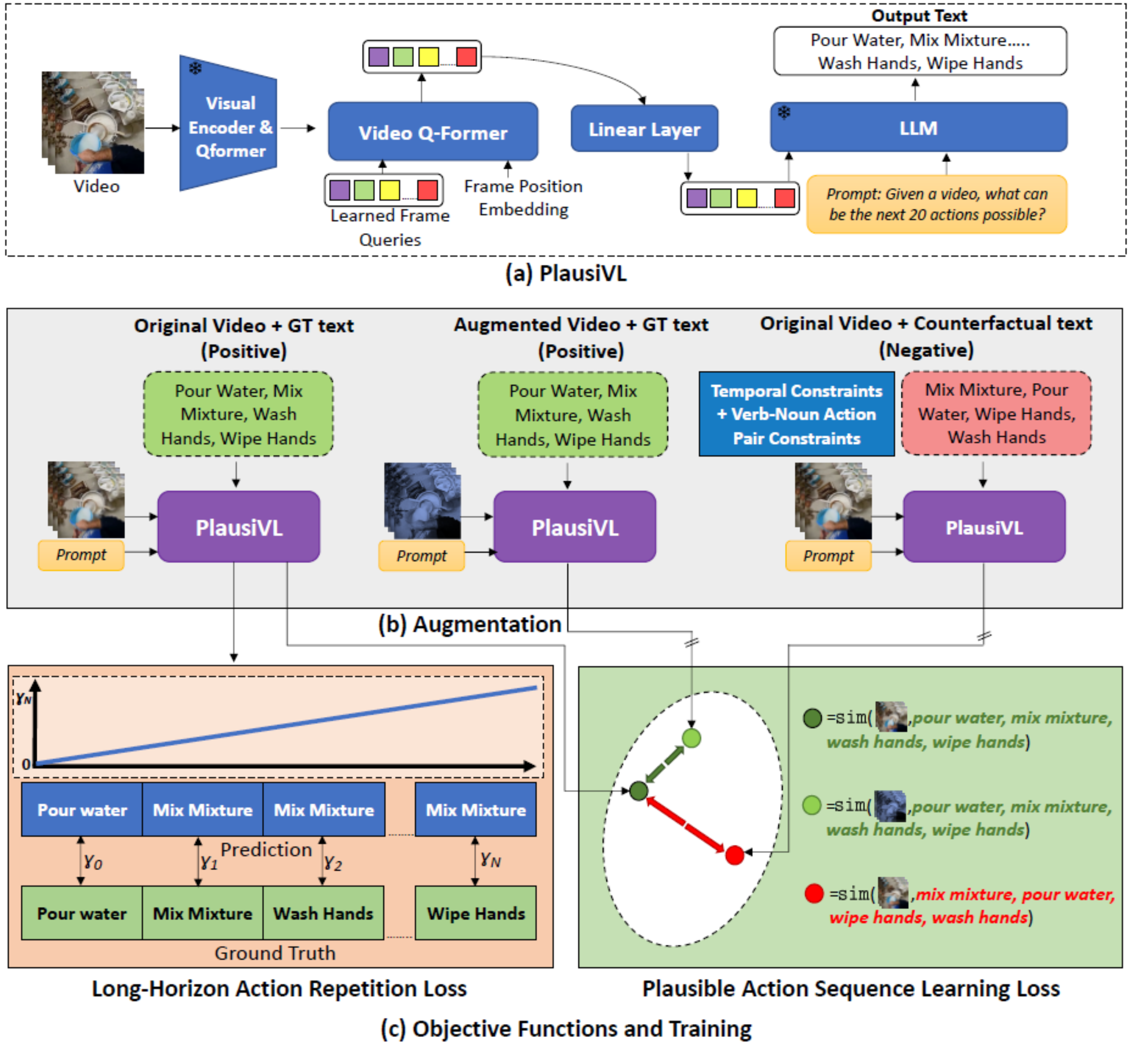


- **Plausible** predictions by Plausible Action Sequence Learning Loss
- **Diverse** predictions by Long-Horizon Action Repetition Loss



Method

- **Plausible action sequence learning loss**: Help models differentiate between plausible and not plausible action sequences.
- **Long-horizon action repetition loss**: Put a higher penalty on the actions that are more prone to repetition over a longer temporal window.



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**

MLLM for Video Affective Reasoning

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

Yuxiang Guo[1*†], Faizan Siddiqui[2], Yang Zhao[1], Rama Chellappa[1*], Shao-Yuan Lo[2*]

[1]Johns Hopkins University.
[2]Honda Research Institute USA.

- 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**

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

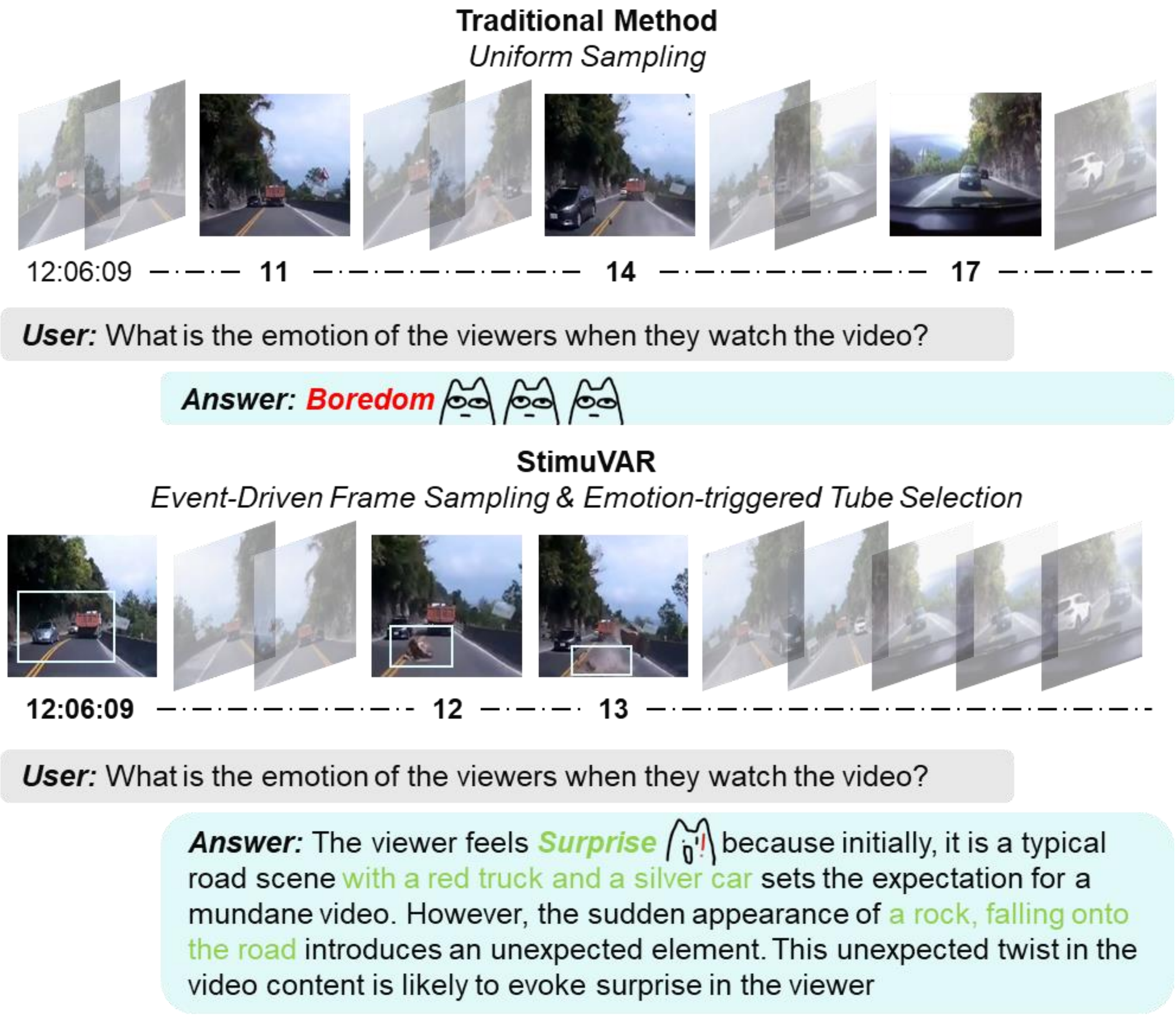
- Let's start from an example video

What we see:

What models see:

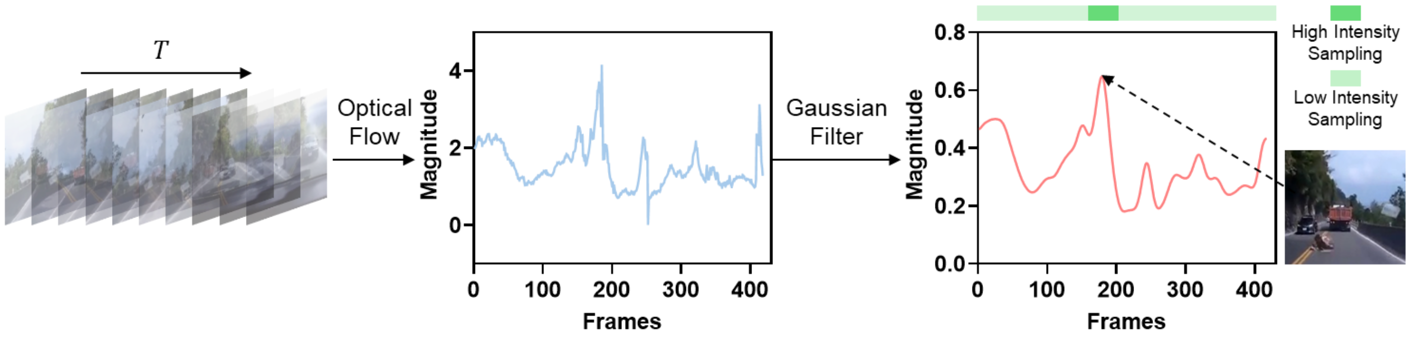
Method

- **Traditional models** and **generic MLLMs** often overlook emotional stimuli.
- Our method has stimuli awareness.

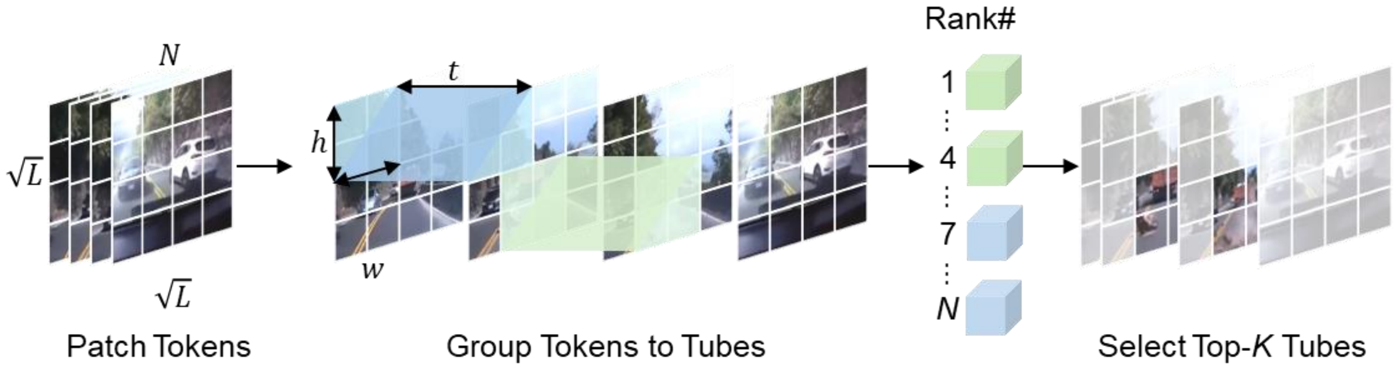


Spatiotemporal Stimuli-aware Mechanisms

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



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

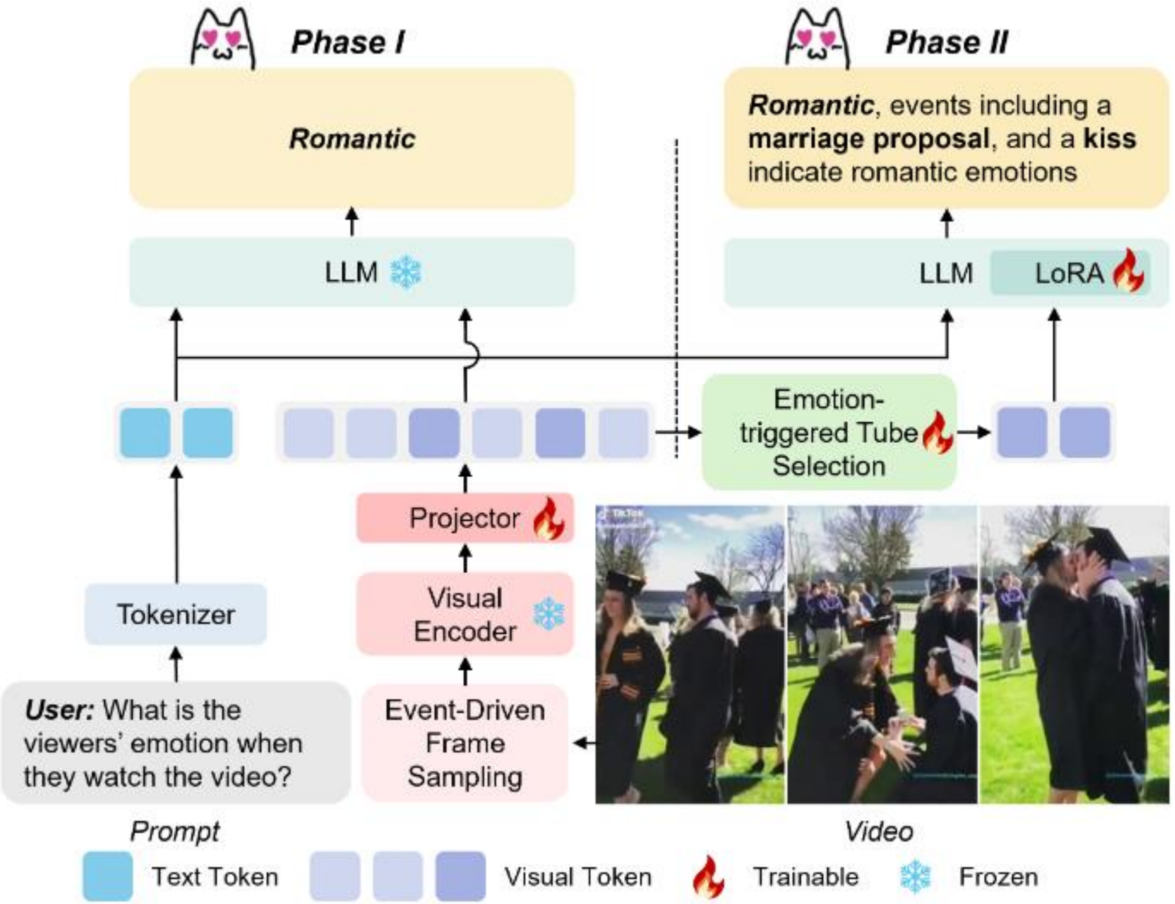


Affective Training

- 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.



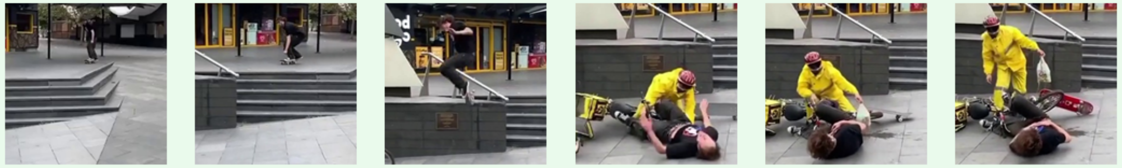
Results

- 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 Theory-of-Mind

Overcoming Multi-step Complexity in Multimodal Theory-of-Mind Reasoning: A Scalable Bayesian Planner

Chunhui Zhang, Sean Dae Houlihan, Kwonjoon Lee, Nakul Agarwal, Zhongyu Ouyang, Soroush Vosoughi, Shao-Yuan Lo

- An **analysis-style** paper for Multimodal Theory-of-Mind (MMToM), a new topic.
- Scaling MMToM on larger language models (LMs) **without increasing training costs**.

What is Theory of Mind?

- Theory of Mind (ToM) is the ability to **understand other people's mental states**, such as thoughts, emotions, intentions, and beliefs.
- **Machine ToM** aims to replicate this human's innate ability in AI agents.

[He et al. EMNLP-Findings'23]

Multimodal ToM, a New Topic

MMToM-QA: Multimodal Theory of Mind Question Answering

Chuanyang Jin[1], Yutong Wu[2], Jing Cao[3], Jiannan Xiang[4],
Yen-Ling Kuo[5], Zhiting Hu[4], Tomer Ullman[2], Antonio Torralba[3], Joshua Tenenbaum[3], Tianmin Shu[6]
[1]NYU, [2]Harvard, [3]MIT, [4]UCSD, [5]UVA, [6]JHU
ACL 2024

Outstanding Paper Award

VIDEO INPUT

TEXT INPUT

What's inside the apartment: … The kitchen is equipped with a microwave, eight cabinets, … Inside the microwave, there is a cupcake. There is a wine glass and an apple on one of the kitchen tables. There are water glasses, a bottle wine, a condiment bottle, and a bag of chips in inside the cabinets. …
Actions taken by Emily: Emily is initially in the bathroom. She then walks to the kitchen, goes to the sixth cabinet, opens it, subsequently closes it, and then goes towards the fourth cabinet.

QUESTION

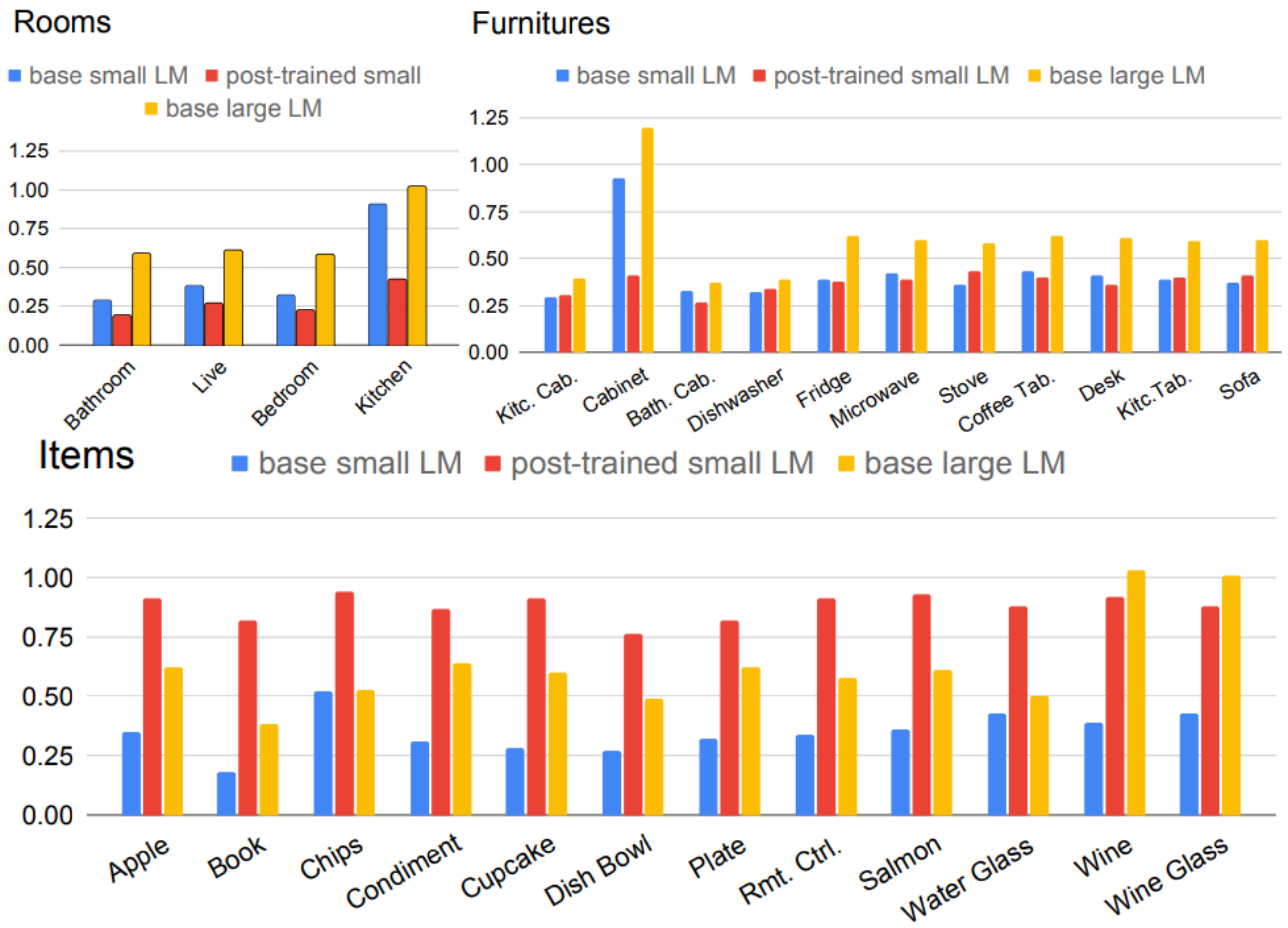
Which one of the following statements is more likely to be true?
(a) Emily has been trying to get a cupcake. ✔ (b) Emily has been trying to get a wine glass. ✖

- However, MMToM training is expensive, e.g., **12 GPU hours for Llama2-7B**.
- How can we efficiently scale MMToM on larger LMs, e.g., **Llama3.1-405B**?

Model Behaviors

- **Base Small LM** vs. **Post-trained Small LM** vs. **Base Large LM**

- 3 levels of concept granularity: **rooms**, **furniture**, and **items**



Weak-to-Strong Control

- **Post-trained Small LM** is better aligned with requirements for specific ToM scenarios.
- **Base Large LM** has better general world knowledge and reasoning.

- Transfer the post-trained alignment from Small LM to Large LM.
- Adapt Large LM's ToM behaviors by training Small LM only.

$$\text{Logits}_{\text{large aligned}} = \text{Logits}_{\text{large}} \times \left(\frac{\text{Logits}_{\text{small aligned}}}{\text{Logits}_{\text{small base}}} \right)$$

Results

- **Dataset**: MMToM-QA. **Metric**: Accuracy.

LM	config	belief inference 1.1	1.2	1.3	avg.	goal inference 2.1	2.2	2.3	2.4	avg.	all
Llama-3.1	8B-zero-shot	88.00	72.00	91.00	83.67	65.33	62.67	22.67	54.67	51.33	65.19
	8B-post-trained	90.00	71.00	93.00	84.67	69.33	72.00	62.67	72.00	69.00	75.71
	70B-zero-shot	85.00	63.00	93.00	80.33	72.00	76.00	16.00	61.33	56.33	66.62
	70B-post-trained	91.00	69.00	**95.00**	85.00	69.33	80.00	29.33	69.33	62.00	71.86
	405B-zero-shot	86.00	70.00	90.00	82.00	73.33	78.67	21.33	66.67	60.00	69.43
	70B-ours	90.00	74.00	93.00	85.67	**74.67**	77.33	70.67	76.00	74.67	79.38
	405B-ours	**92.00**	**76.00**	93.00	**87.00**	73.33	**80.00**	**76.00**	**78.67**	**77.00**	**81.29**

Data-efficient MLLM fine-tuning

Filter Images First, Generate Instructions Later:
Pre-Instruction Data Selection for Visual Instruction Tuning

Bardia Safaei[1*], Faizan Siddiqui[2], Jiacong Xu[1], Vishal M. Patel[1], Shao-Yuan Lo[2]
[1]Johns Hopkins University, [2]Honda Research Institute USA
{bsafaei1, jxu155, vpatel36}@jhu.edu {faizan_siddiqui, shao-yuan_lo}@honda-ri.com

CVPR 2025 (Highlight)

- A new data selection paradigm for visual instruction tuning (VIT): **Pre-instruction** data selection
- Reduces not only **VIT runtime** but also **instruction generation cost**

What is Visual Instruction Tuning?

- A standard way to train fine-tune a MLLM
- Require a large amount of visual instruction tuning data (e.g., LLaVA-1.5 uses 665K data, covering 10 vision tasks)
- Use GPT to generate visual instructions from images

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

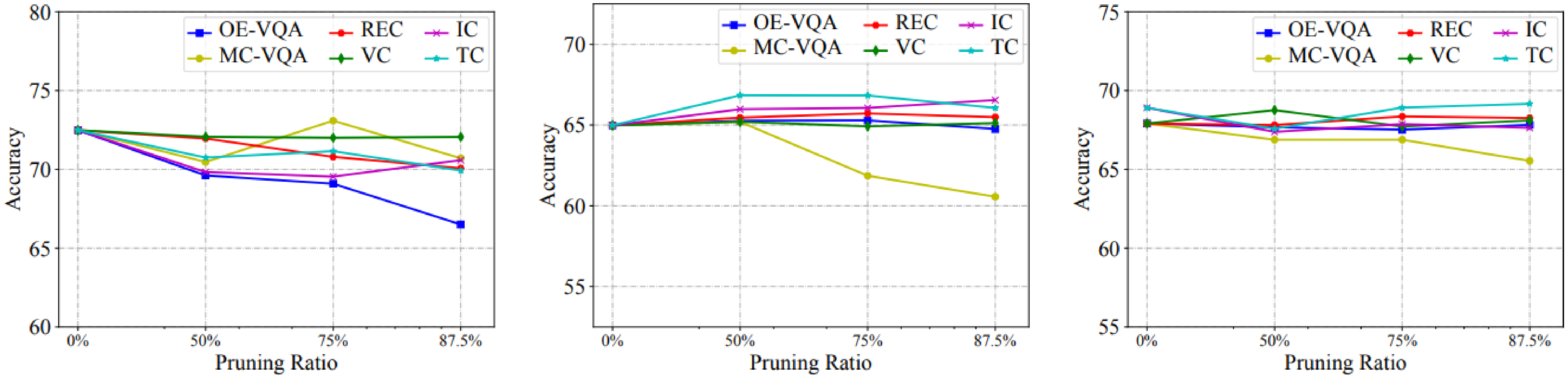
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

[Liu et al. NeurIPS'23]

Visual Instruction Data are Redundant

- Selecting data in advance can reduce training cost

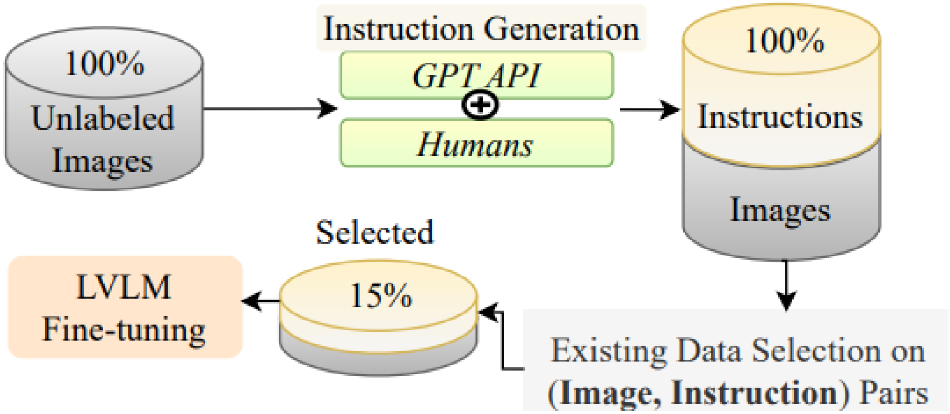


(a) MME-P (b) MMBench (c) ScienceQA (Image)

[Liu et al. arXiv'24]

Visual Instruction Data are Redundant

- Existing data selection method assumes that instructions are already generated.



Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

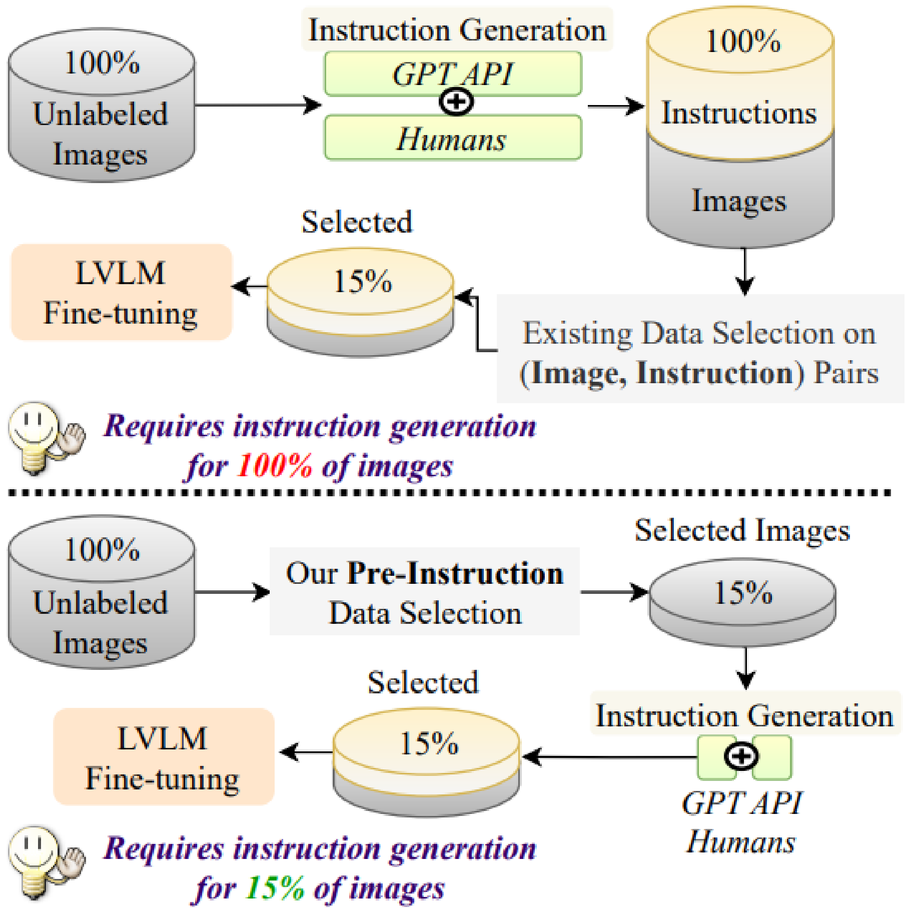
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

[Liu et al. NeurIPS'23]

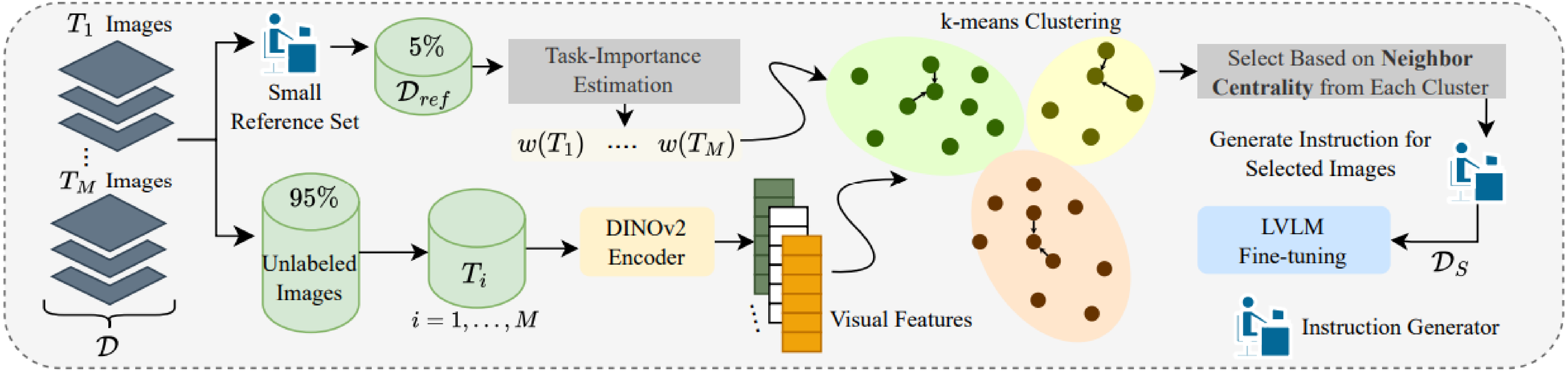
Goal

- Generating instruction is also expensive!
- Can we select essential data before instruction generation?



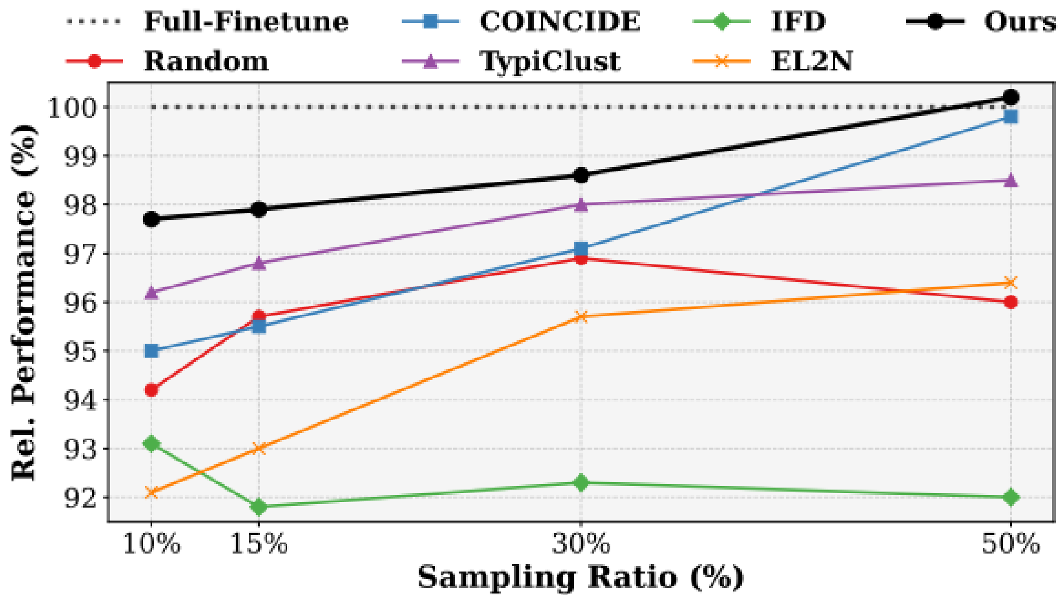
Method

- **Task-importance estimation**: Determine data proportion for each task.
- **Cluster-based selection**: Select representative data within each task.



Results

- Experiments on LLaVA-1.5



Methods	Selection Cost	Finetuning Cost	Inst. Gen. Cost	Total Cost	**Rel. (%)**
Full Finetune	–	76.0 GPU-hr	$100\% \cdot C$	76.0 GPU-hr + $100\% \cdot C$	100
Self-Filter [5]	73.5 GPU-hr	11.0 GPU-hr	$100\% \cdot C$	84.5 GPU-hr + $100\% \cdot C$	88.8
COINCIDE [15]	55.5 GPU-hr	11.0 GPU-hr	$100\% \cdot C$	66.5 GPU-hr + $100\% \cdot C$	95.5
PreSel (Ours)	9.0 GPU-hr	11.0 GPU-hr	$15\% \cdot C$	**20.0 GPU-hr + 15% $\cdot C$**	**97.9**

HRI hires research interns!
(next hiring will be around October)
https://usa.honda-ri.com/intern-positions

About 5 scientists in the ML/CV Team

CVPR 2025	x3
ICLR 2025	x1
NeurIPS 2024	x2
ECCV 2024	x3
CVPR 2024	x2

Candidates with at least one first-authored paper at a top conference have a stronger chance.

Honda Research Institute US

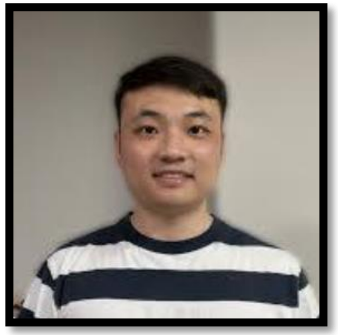
Johns Hopkins University

Yuchen Yang

Kwonjoon Lee

Behzad Dariush

Yinzhi Cao



Jiacong Xu

Bardia Safaei

Vishal M. Patel

Isht Dwivedi

Email: shao-yuan_lo@honda-ri.com
Webpage: https://shaoyuanlo.github.io