AI Safety and Beyond:
Robustness, Monitoring, and Alignment

羅紹元 (Shao-Yuan Lo)

Research Scientist @ Honda Research Institute USA

11/4/2024 @ NTUEE

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 (2013 - 2017)

Prof. Vishal M. Patel

Prof. Hsueh-Ming Hang

AI Safety Matters!!

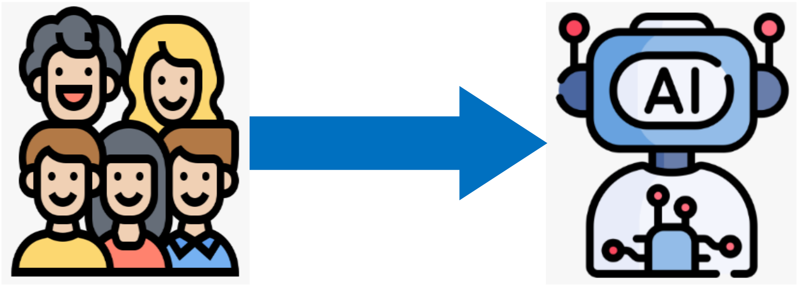
AI is becoming increasingly integrated into human society.

However, AI also brings considerable **risks**, and AI safety research has **not** kept pace with its rapid advancement.

AI safety research ensure AI's **positive impact on humanity** and enables us to **unlock AI's full potential** safely.

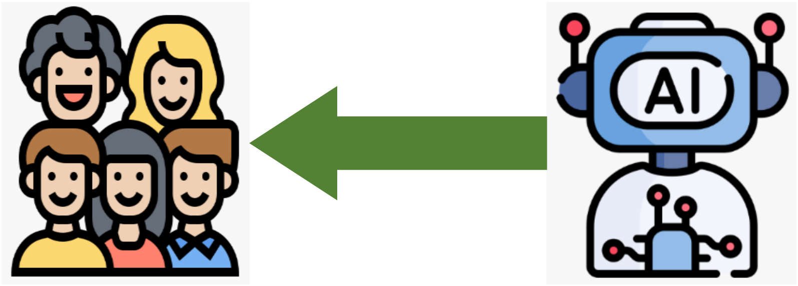
AI Safety Scope

Robustness



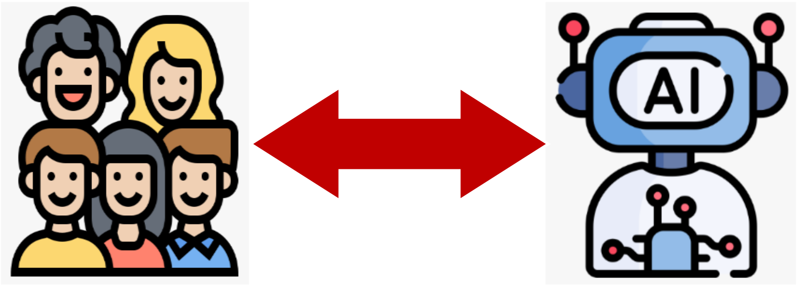
Make AI safe

Monitoring



Use AI to make
human society safe

Alignment



Foster harmonious
human-AI interactions

My Research in AI Safety

Robustness

Adversarial Robustness

[LOP, T-PAMI'22]
[LP, T-IP'21]
[LP, ICIP'21]
[LVP, ICIP'21]
[LP, ACCV'22]
[LP, AVSS'24]
[LP, AVSS'21]

Domain Generalization

[LOCGP, CVPR'23]
[LWTZPK, IROS'22]

Monitoring

Anomaly Detection

[YLDCL, ECCV'24]
[XLPD, 2024]

Behavior Forecast

[GALLJ, CVPR'24]
[MALL, CVPR'24]

Alignment

Theory of Mind

[ZHLAOHL, 2024]

Learning Alignment

[GSZCL, 2024]
[SSPL, 2024]

Robustness

Robust AI

- **Adversarial Robustness**
  - **Novel attacks** [Lo & Patel, AVSS’21]
  - **Empirical defenses** [Lo et al. ICIP’21] [Lo & Patel, ICIP’21]
  - **Generalizable defenses** [Lo & Patel, IEEE T-IP’21]
  - **Defenses for less explored tasks** [Lo et al. IEEE T-PAMI’22]
- **Domain Generalization**
  - **Unsupervised domain adaptation** [Lo et al. IROS’22]
  - **Source-free domain adaptation** [Lo et al. CVPR’23]
- Others

Intersection of AR & DG

- **Adversarial defense for domain adaptation** [Lo & Patel, ACCV’22]
- **Adversarial defense via domain adaptation** [Lo & Patel, AVSS’24]

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2022

Adversarially Robust One-class Novelty Detection

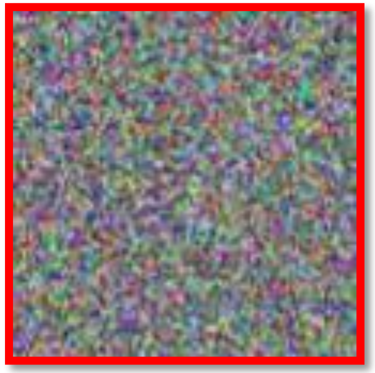
Shao-Yuan Lo, *Student Member, IEEE,* Poojan Oza, *Student Member, IEEE,*
and Vishal M. Patel, *Senior Member, IEEE*

- We find that **image classification**-based methods do not work well on the **novelty detection** task due to the **unique property of this task**.
- We propose the **first** adversarially robust methods for novelty detection.
- We establish a **solid evaluation benchmark** and **comprehensive baseline results**.

Recall: Adversarial Examples

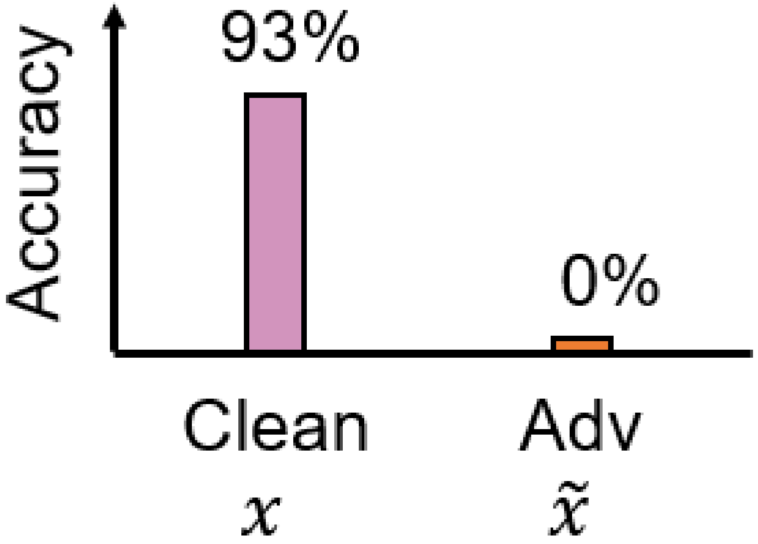
- Deep networks are **vulnerable** to adversarial examples.

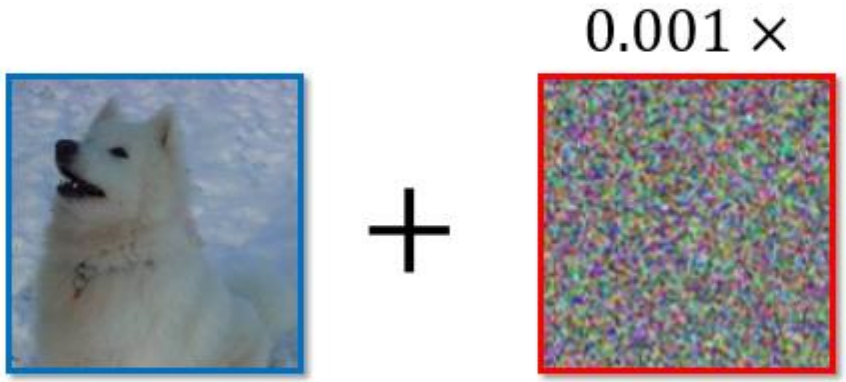
$f_\theta($  $) = $ "$\color{blue}{Dog}$"

$f_\theta($  $+\ 0.001 \times$  $) = $ "$\color{red}{Cat}$"

Recall: Adversarial Examples

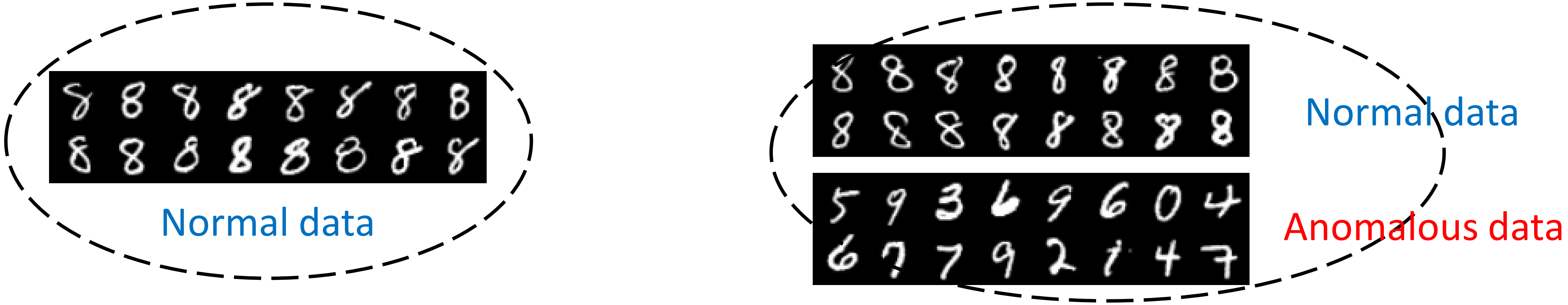
- Dataset: CIFAR-10
- Network: ResNet-50





Recall: One-class Novelty Detection

- One-class novelty detection model is trained with examples of **a particular class** and is asked to identify whether a query example belongs to the same known class.
- Example:
  - **Known class** (normal data): 8
  - **Novel classes** (anomalous data): 0-7 & 9 (the rest of classes)

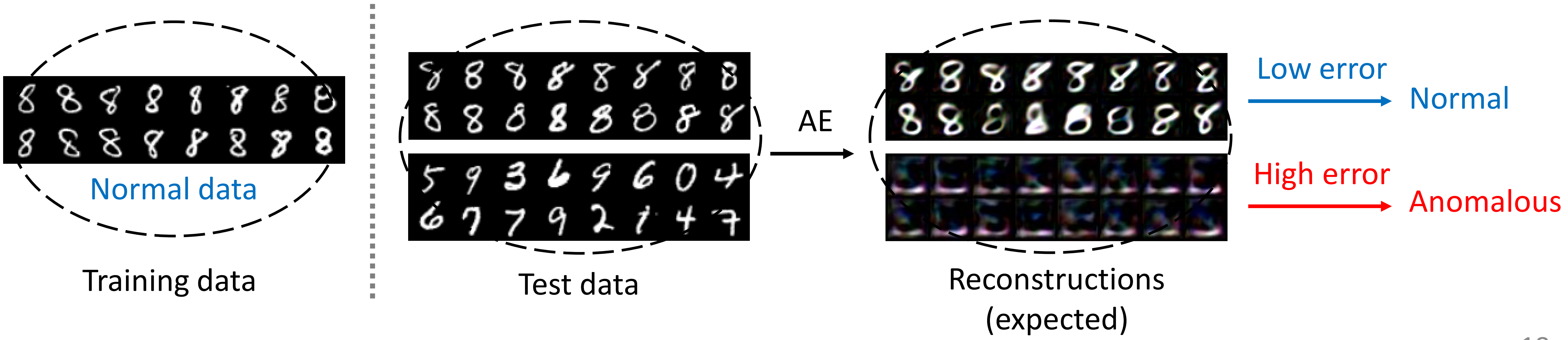


Training data

Test data

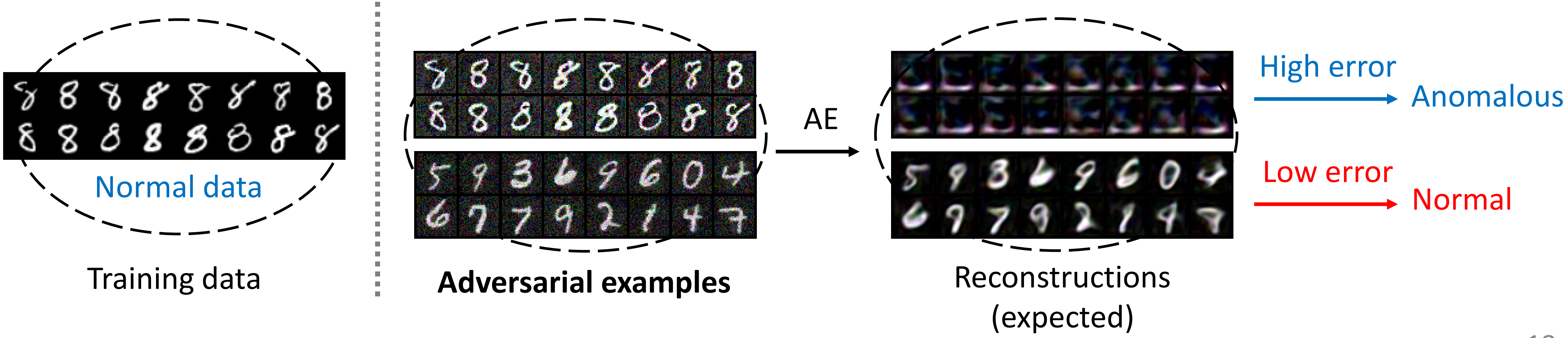
Recall: One-class Novelty Detection

- Most recent advances are based on the **autoencoder** architecture.
- Given an autoencoder that learns the distribution of the known class, we expect that the **normal data** are reconstructed accurately while the **anomalous data** are not.



Attacking One-class Novelty Detection

- How to generate adversarial examples against a novelty detector?
- If a test example is **normal**, **maximize** the reconstruction error.
- If a test example is **anomalous**, **minimize** the reconstruction error.

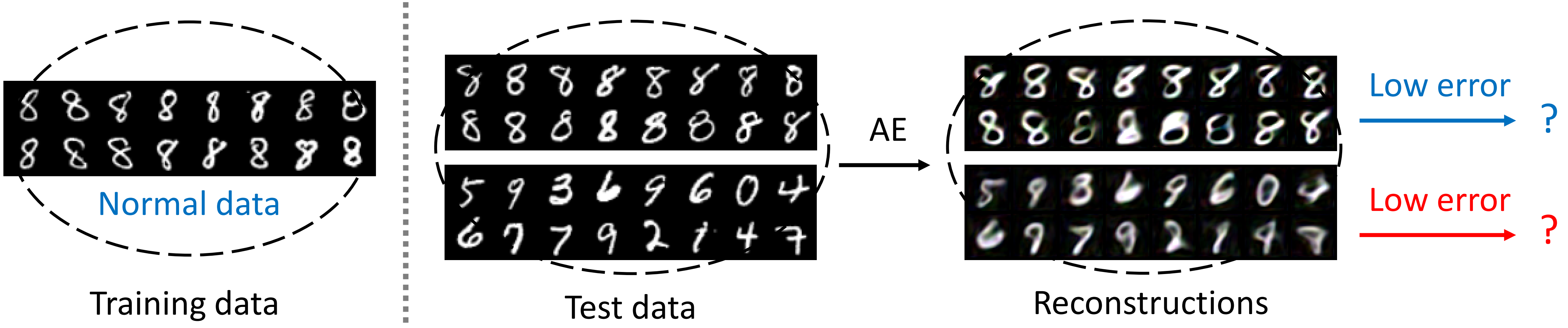


Goal: Adversarially Robust Novelty Detection

- Novelty detectors are **vulnerable** to adversarial attacks.
- Adversarially robust method specifically designed for novelty detectors is needed.
- A **new** research problem.

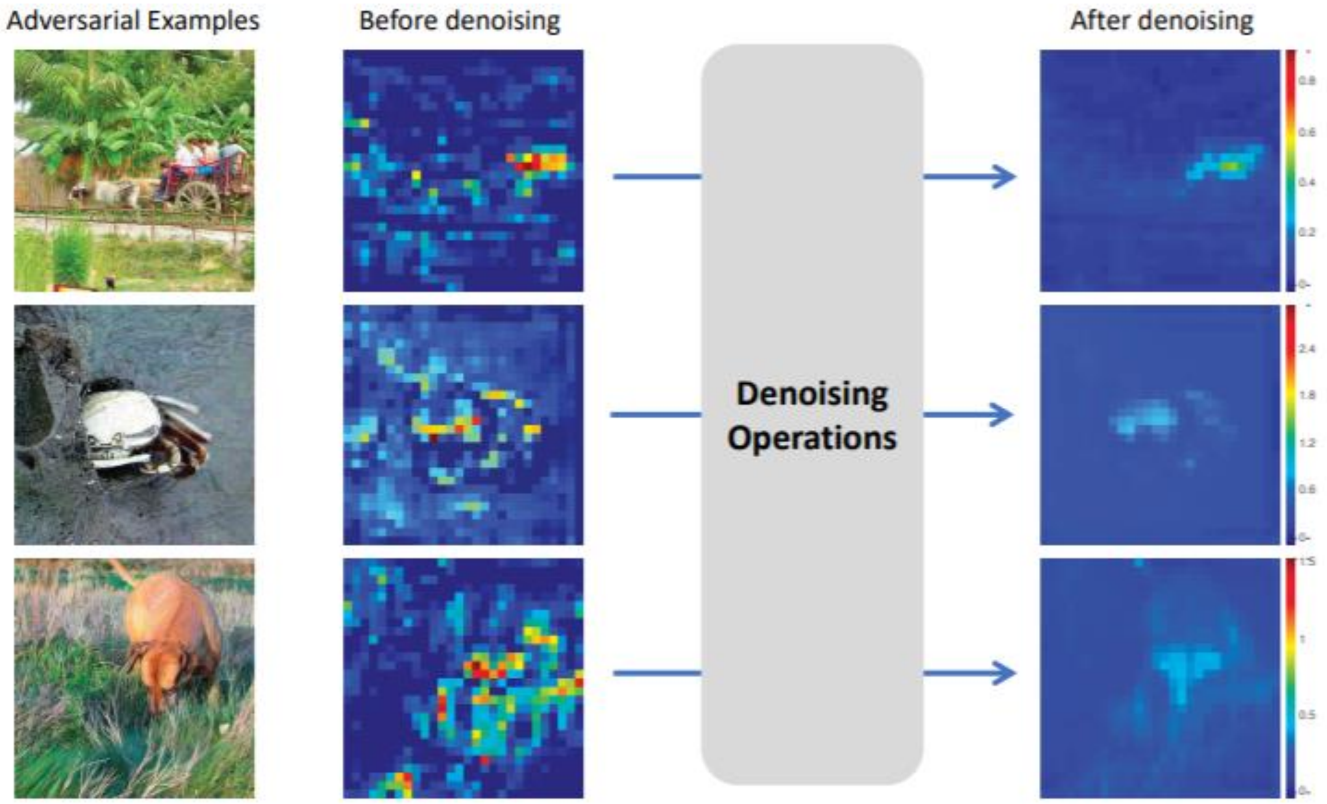
Observation: Generalizability

- Unique property: Preference for **poor** generalization of reconstruction ability.
- However, autoencoders have **good** generalizability.



Observation: Feature Denoising

- Adversarial perturbations can be removed in the **feature** domain.



[Xie et al. CVPR’19]

Method

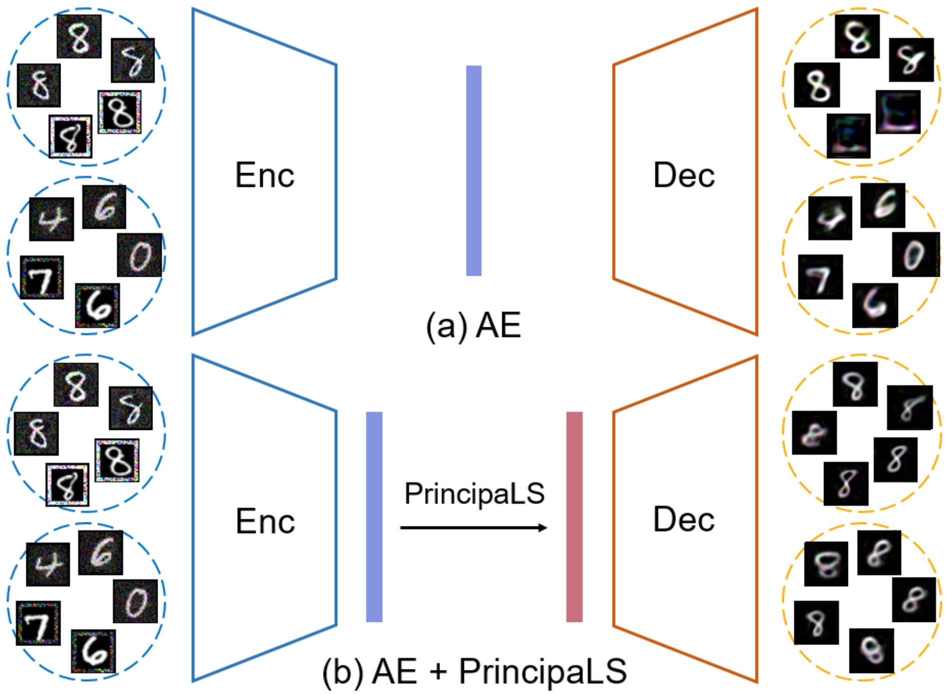
- **Observations**: Generalizability and Feature Denoising.

↓

- **Assumption**: One can **largely** manipulate the latent space of a novelty detector to remove adversaries to a great extent, and this would not hurt the model capacity but **helps** if in a proper way.

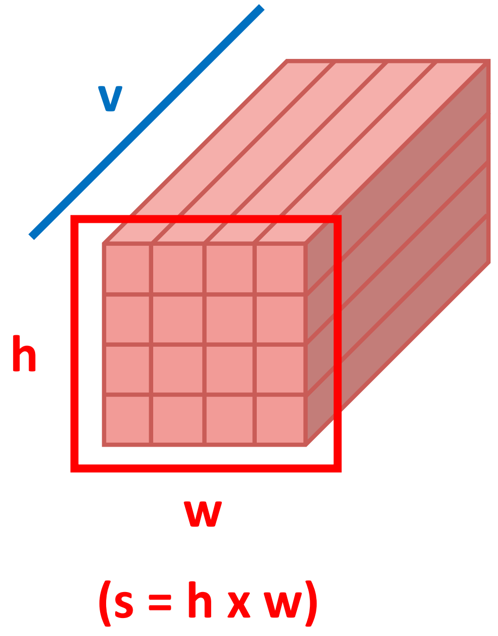
↓

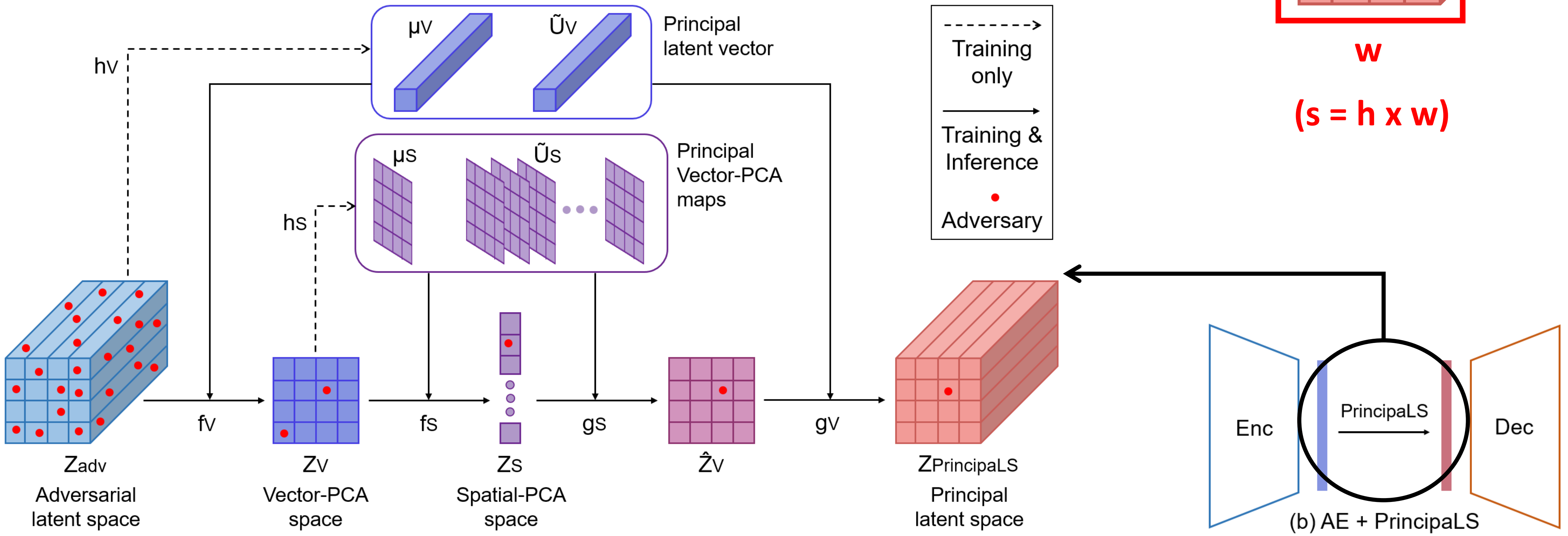
- **Solution**: Learning **principal latent space**.



Method

- **Vector-PCA** performs PCA on the **vector** dimension.
- **Spatial-PCA** performs PCA on the **spatial** dimension.

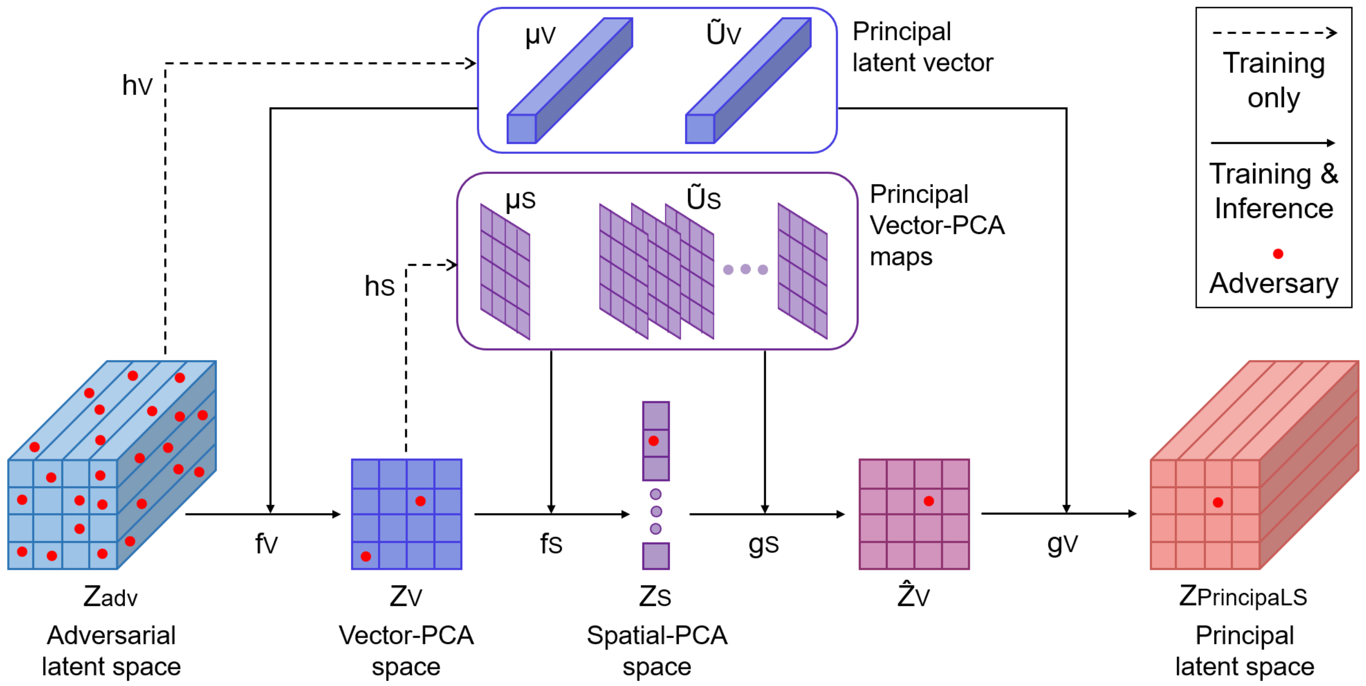




Method

- **Vector-PCA** replaces the perturbed latent vectors with the clean principal latent vector.

- **Spatial-PCA** removes the remaining perturbations on the Vector-PCA map.



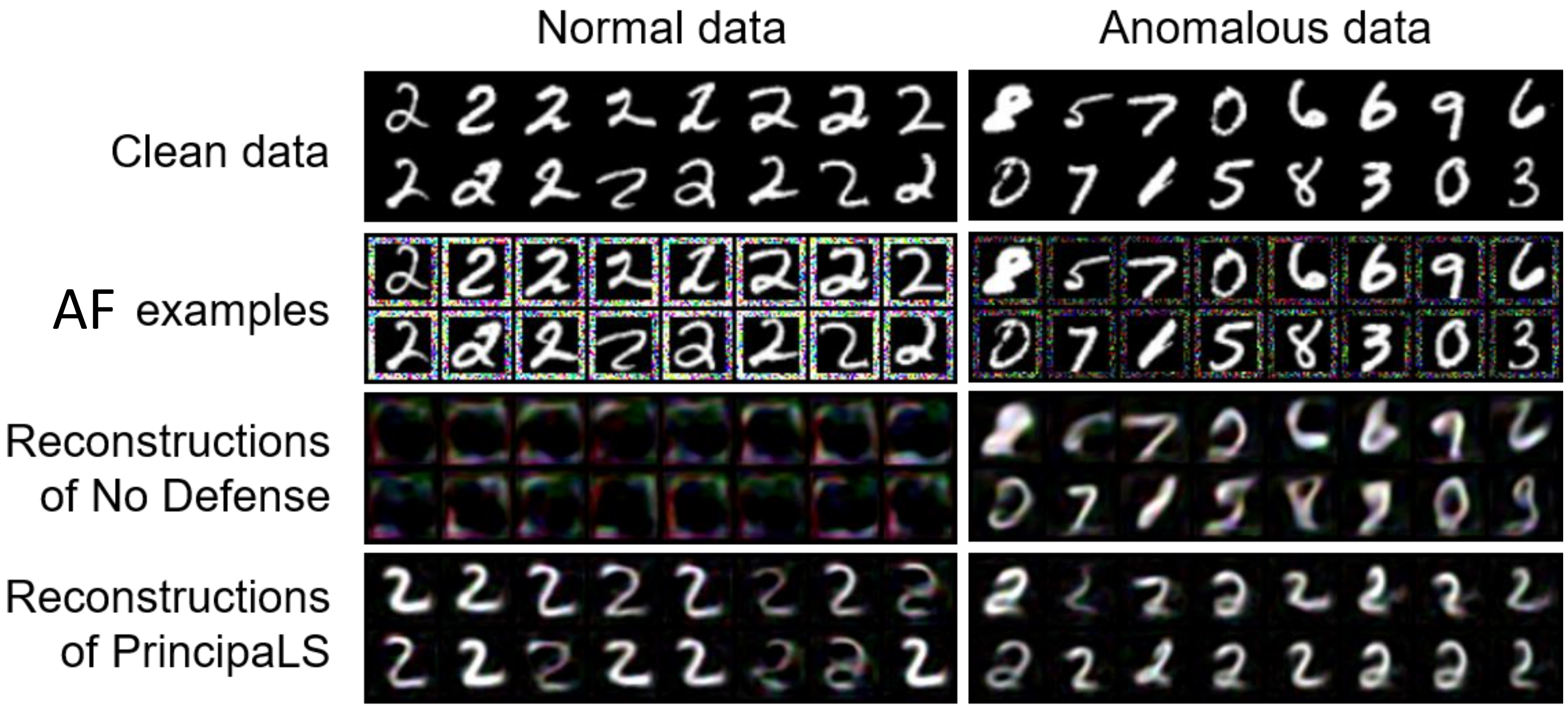
Results

- Evaluation metric: mean of AUROC
- PrincipaLS is effective on **5** datasets against **6** attacks for **7** novelty detection methods.

Dataset	Defense	Clean	FGSM [11]	PGD [27]	MI-FGSM [36]	MultAdv [37]	AF [38]	Black-box [47]	Average
MNIST [48]	No Defense	0.964	0.350	0.051	0.022	0.170	0.014	0.790	0.337
	PGD-AT [27]	0.961	0.604	0.357	0.369	0.444	0.155	0.691	0.512
	FD [15]	0.963	0.612	0.366	0.379	0.453	0.142	0.700	0.516
	SAT [23]	0.947	0.527	0.295	0.306	0.370	0.142	0.652	0.463
	RotNet-AT [21]	0.967	0.598	0.333	0.333	0.424	0.101	0.695	0.493
	SOAP [22]	0.940	0.686	0.504	0.506	0.433	0.088	0.863	0.574
	APAE [46]	0.925	0.428	0.104	0.105	0.251	0.022	0.730	0.366
	PrincipaLS (ours)	**0.973**	**0.812**	**0.706**	**0.707**	**0.725**	**0.636**	**0.866**	**0.775**
SHTech [52]	No Defense	0.523	0.204	0.034	0.038	0.006	0.000	0.220	0.146
	PGD-AT [27]	0.527	0.217	0.168	0.154	0.100	0.000	0.221	0.198
	FD [15]	0.528	0.226	0.189	0.181	0.132	0.002	0.229	0.212
	SAT [23]	**0.529**	0.184	0.110	0.092	0.040	0.000	0.199	0.165
	RotNet-AT [21]	0.516	0.220	0.163	0.158	0.113	0.000	0.229	0.200
	SOAP [22]	0.432	0.024	0.002	0.000	0.002	**0.181**	0.202	0.120
	APAE [46]	0.510	0.215	0.048	0.050	0.011	0.000	0.207	0.149
	PrincipaLS (ours)	0.498	**0.274**	**0.223**	**0.217**	**0.175**	0.051	**0.308**	**0.249**

Analysis

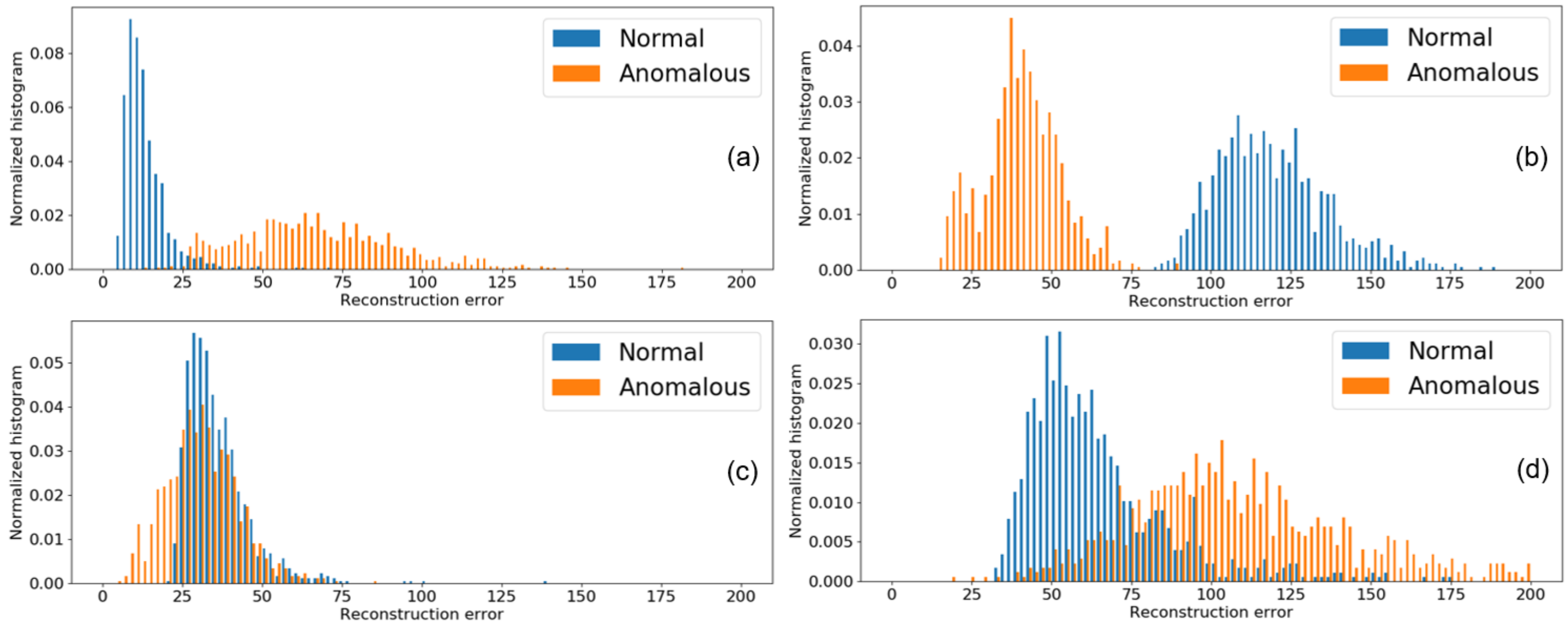
- PrincipaLS reconstructs **every** input example to the known class (digit 2).



(b) AF attack

Analysis

- (a) No Defense under clean data (b) No Defense under PGD attack
  (c) PGD-AT under PGD attack (d) PrincipaLS under PGD attack
- PrincipaLS enlarges the reconstruction errors of anomalous data to a great extent.



Monitoring

- Identify and forecast malicious scenarios
- Leveraging AI to enhance the safety of human society

Multimodal LLMs for Anomaly Detection

- **Reasoning for AD**
  [YLDCL, ECCV'24]
- **Unified multimodal AD**
  [XLPD, 2024]

Multimodal LLMs for Behavior Forecast

- **Short-term forecast**
  [GALLJ, CVPR'24]
- **Long-term forecast**
  [MALL, CVPR'24]

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

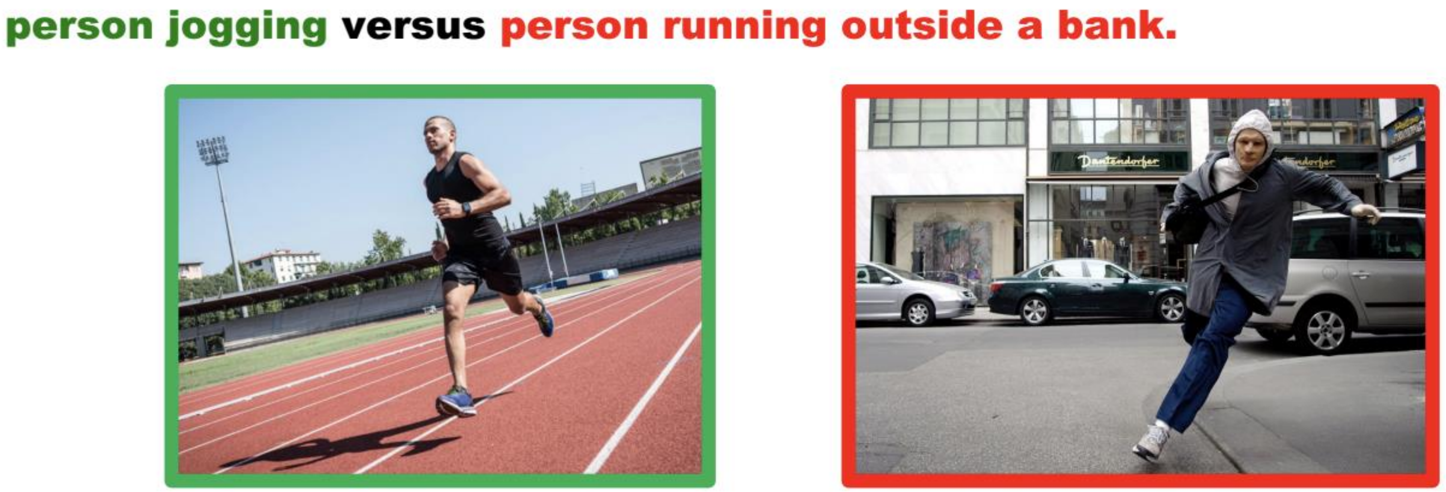
Yuchen Yang[1*], Kwonjoon Lee[2], Behzad Dariush[2], Yinzhi Cao[1], and Shao-Yuan Lo[2]

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

- 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

- 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.



Credit: https://drive.google.com/file/d/1bP6BbfREhpZjokyjYSE82c_JmnxpmIsn/view

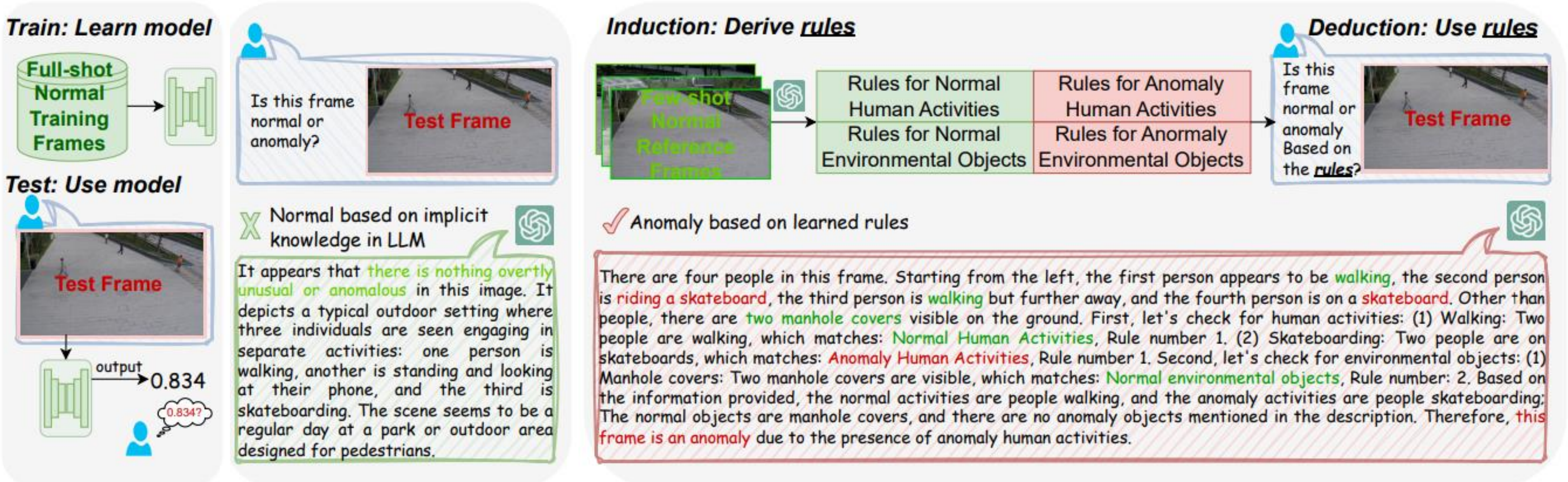
Our Goal

- 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

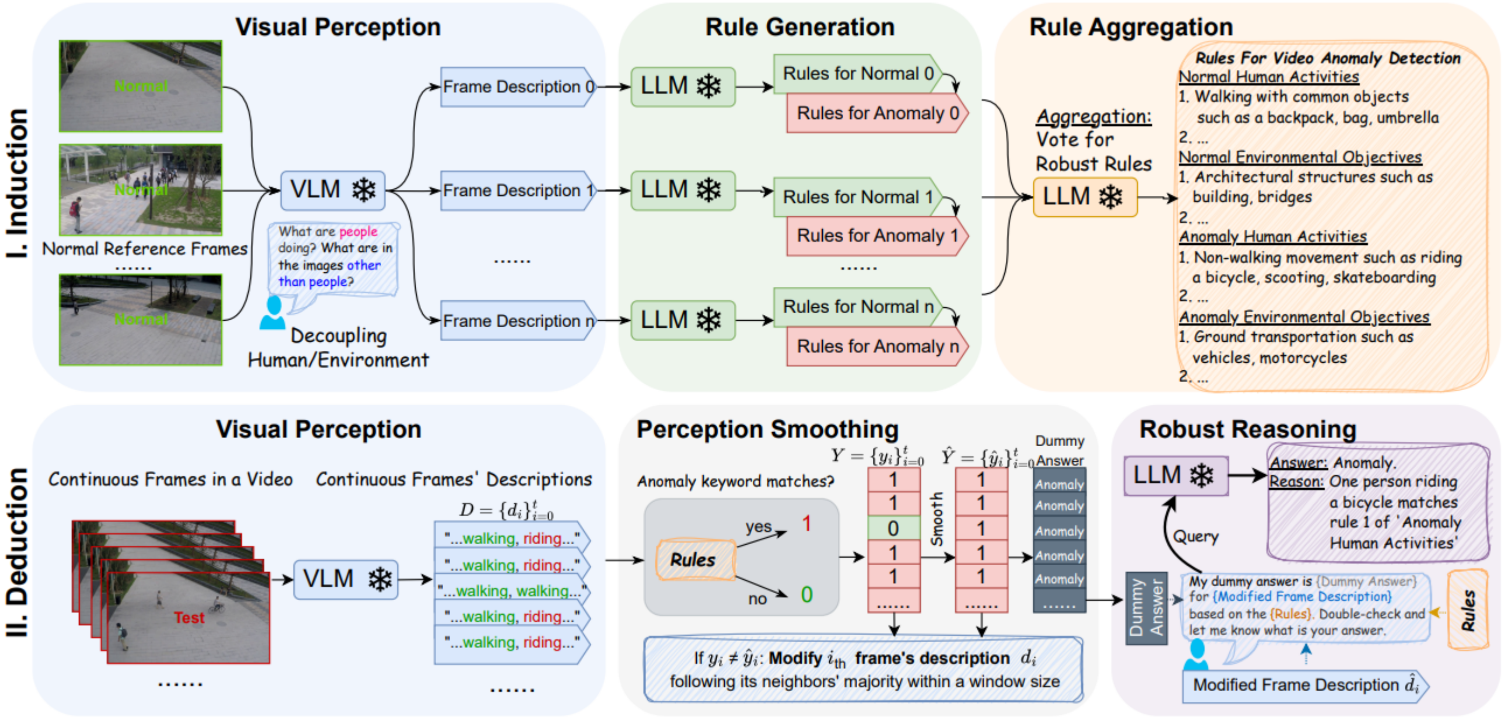
- **Traditional VAD**: Full-shot training. Only output anomaly score.
- **Ask LLM directly**: The implicit knowledge pre-trained in LLMs may not align with specific VAD needs (e.g., “skateboarding”).



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.



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

Alignment

- Ensure AI operate in ways that align with human values and intentions
- Foster harmonious human-AI interactions

Multimodal LLMs for Affective Computing

[GSZC**L**, 2024]

Scaling Multimodal Theory-of-Mind

[ZHLAOH**L**, 2024]

Data-Efficient Visual Instruction Tuning

[SSP**L**, 2024]

Scaling Multimodal Theory-of-Mind with Weak-to-Strong Bayesian Reasoning

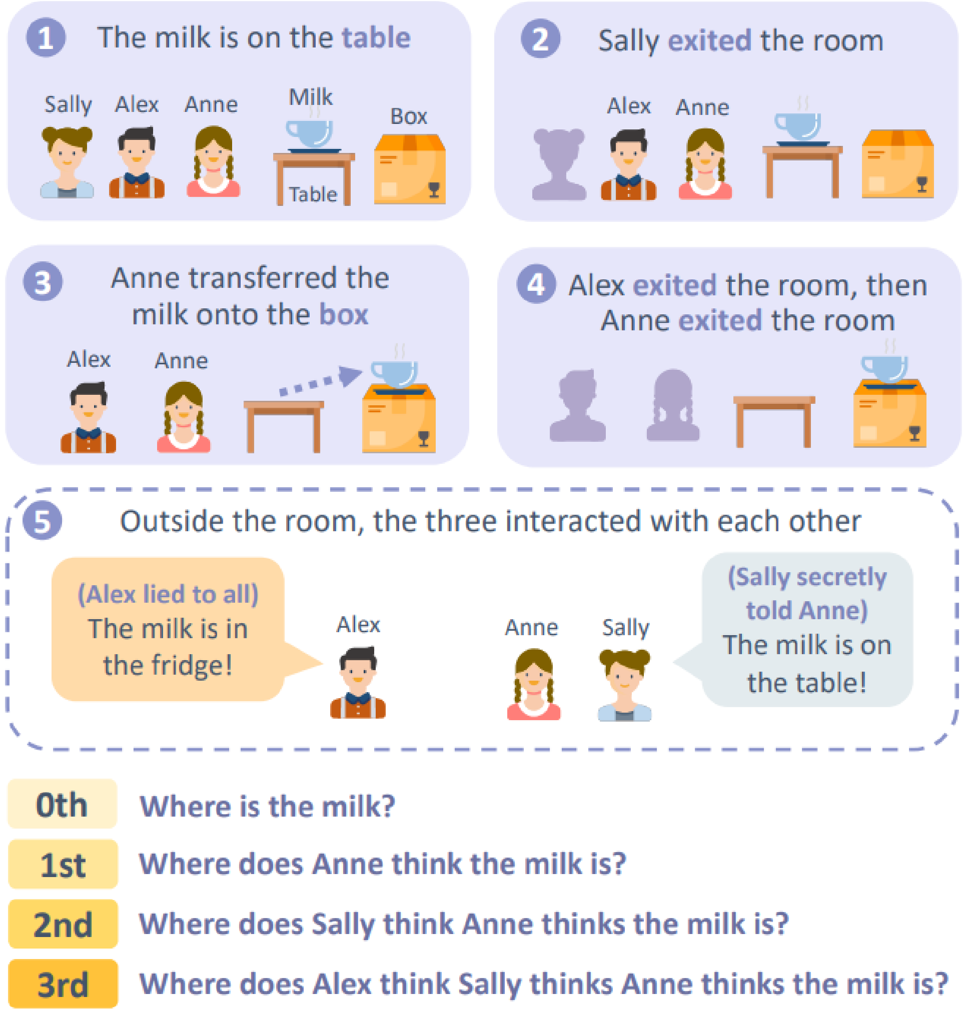
Chunhui Zhang, Sean Dae Houlihan, Kwonjoon Lee, Nakul Agarwal, Zhongyu Ouyang, Soroush Vosoughi, Shao-Yuan Lo

- An **analysis-style** paper for Multimodal Theory-of-Mind (MMToM), a **completely 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.

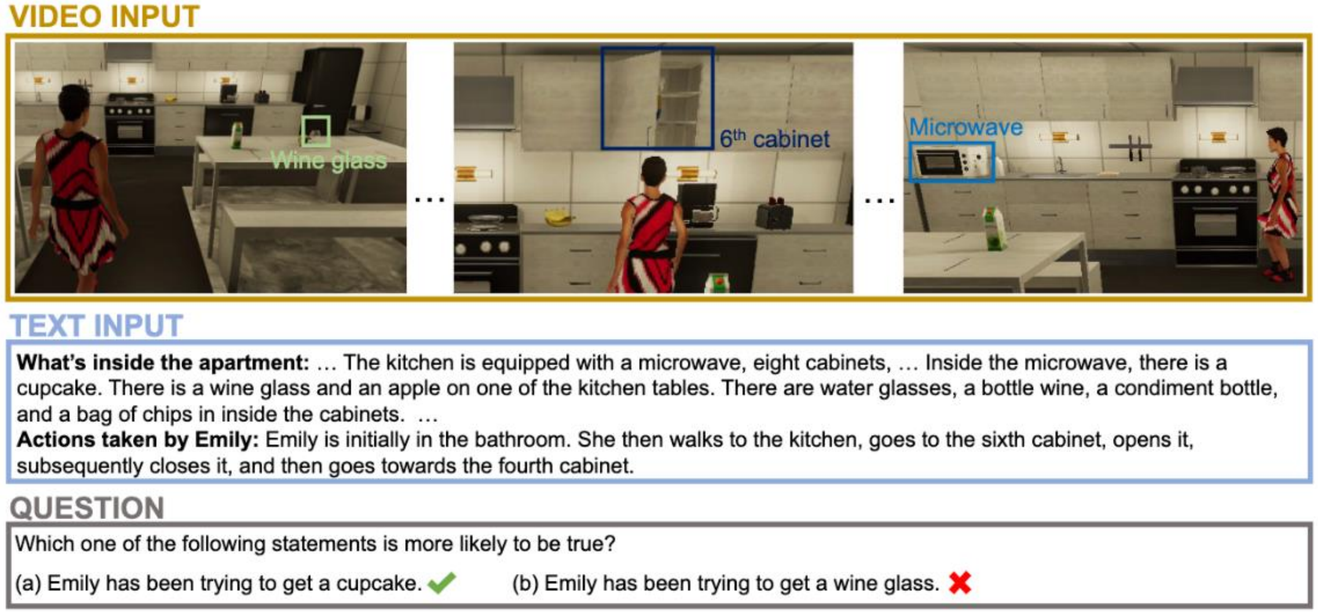
[He et al. EMNLP-Findings'23]



MMToM, a New Topic

MMToM-QA: Multimodal Theory of Mind Question Answering

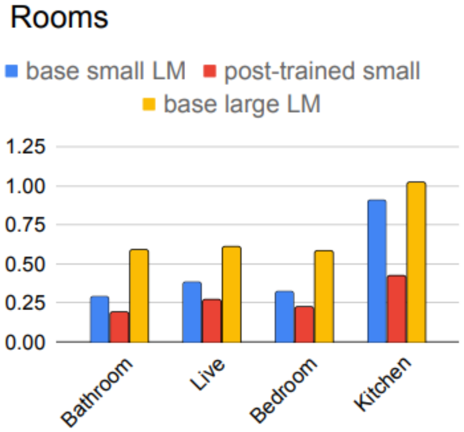
Chuanyang Jin[1], Yutong Wu[2], Jing Cao[3], Jiannan Xiang[4],
Yen-Ling Kuo[5], Zhiting Hu[4], Tomer Ullman[2], Antonio Torralba[3], Joshua Tenenbaum[3], Tianmin Shu[6]
[1]NYU, [2]Harvard, [3]MIT, [4]UCSD, [5]UVA, [6]JHU
ACL 2024
Outstanding Paper Award

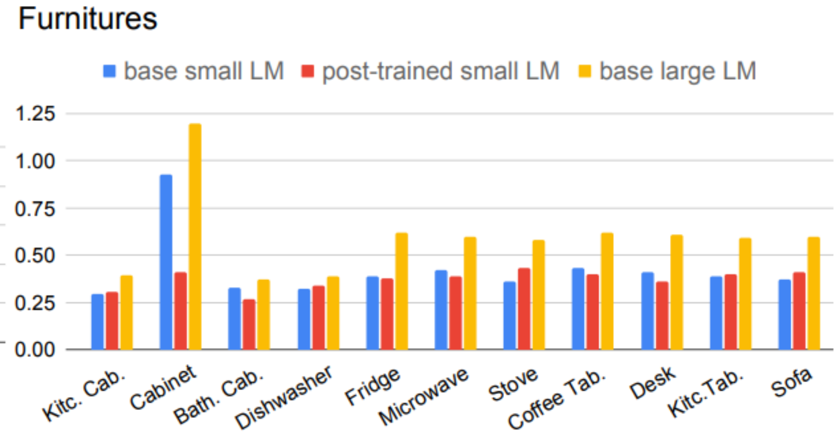


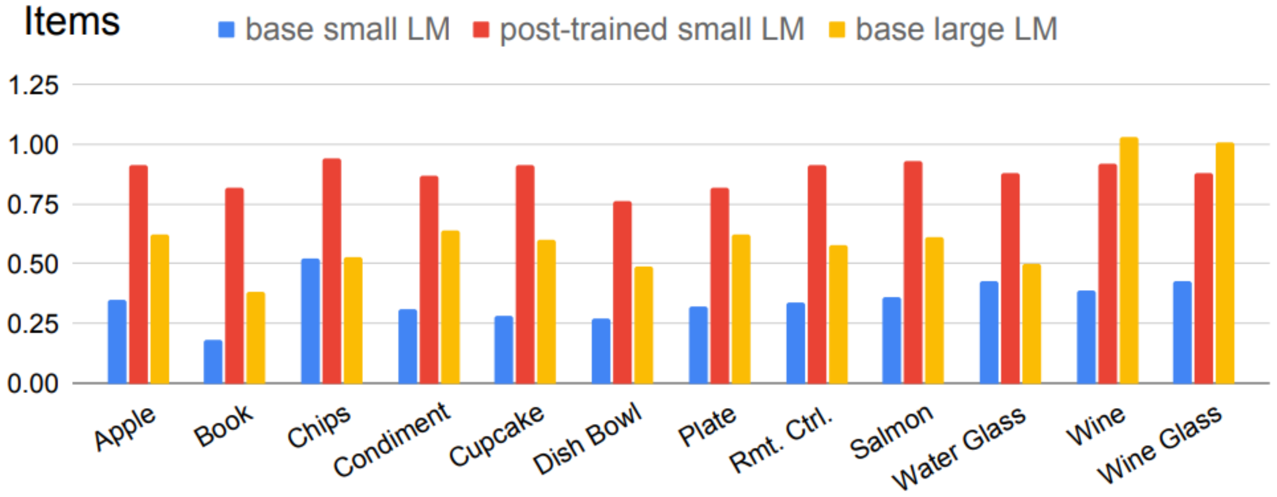
- However, MMToM training is expensive, e.g., **12 GPU hours for Llama2-7B**.
- How can we efficiently scale MMToM on larger LMs, e.g., **Llama3.1-405B**?

Model Behaviors

- **Base Small LM** vs. **Post-trained Small LM** vs. **Base Large LM**
- 3 levels of concept granularity: **rooms**, **furniture**, and **items**







Model Behaviors

- **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	<u>91.00</u>	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	<u>74.00</u>	<u>93.00</u>	<u>85.67</u>	**74.67**	<u>77.33</u>	<u>70.67</u>	<u>76.00</u>	<u>74.67</u>	<u>79.38</u>
	405B-ours	**92.00**	**76.00**	<u>93.00</u>	**87.00**	<u>73.33</u>	**80.00**	**76.00**	**78.67**	**77.00**	**81.29**

Future Research

Robustness		Monitoring		Alignment	
Adversarial Robustness	Domain Generalization	Anomaly Detection	Behavior Forecast	Theory of Mind	Learning Alignment
[LOP, T-PAMI'22] [LP, T-IP'21] [LP, ICIP'21] [LVP, ICIP'21] [LP, ACCV'22] [LP, AVSS'24] [LP, AVSS'21]	[LOCGP, CVPR'23] [LWTZPK, IROS'22]	[YLDCL, ECCV'24] [XLPD, 2024]	[GALLJ, CVPR'24] [MALL, CVPR'24]	[ZHLAOHL, 2024]	[GSZCL, 2024] [SSPL, 2024]

Model Level → **System Level**

Digital Models → **World Models**

Current Research

Future Research