



Computer Science and Information Technology

前瞻資訊科技

Trustworthy AI

羅紹元 Shao-Yuan Lo

National Taiwan University

Week 9
10/31/2025

Who am I?

- Assistant Professor @ **National Taiwan University**
Taiwan (2025 – Present)
- Research Scientist @ **Honda Research Institute USA**
San Jose, CA (2023 – 2025)
- 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)



國立臺灣大學
National Taiwan University



國立交通大學
National Chiao Tung University

Johns Hopkins University



shaoyuanlo > 美國留學 06/17

來安麗一下母校Johns Hopkins 😊

CVPR是世界最頂尖的AI學術會議之一，今年眾星雲集的贊助商中，JHU是唯一的大學，和Nvidia並列同個贊助等級

從我開始參加CVPR以來，第一次有大學作為贊助單位，很潮！身為校友更是驚喜

JHU以醫學和公衛聞名，台灣的電資學生申請留學時較少留意。但其實過去10年，JHU已大力發展CS，重金延攬多位大師級教授，CV和NLP領域都在CSRankings上排前10

去年更成立Data Science & AI學院，將在未來數年招聘110位教授，是美國校院極少見的擴張規模。即便今年各大學財源緊縮，JHU仍逆勢新聘約15位AI教授



Amazon Astro



Amazon Astro, Household robot for home monitoring, with Alexa, Includes a Ring Home Trial

Brand: Amazon

Day1 Editions

Price: **\$1,599.99**

FREE Returns ▾

Terms and conditions apply.

Or **\$533.34 /mo (3 mo)**. Select from 4 plans

Available exclusively by invitation

Ships from and sold by Amazon.com.

Please note this product can only ship to addresses in the 50 US states.

Amazon Just Walk Out



Honda



HONDA
The Power of Dreams

How we move you.
CREATE ▶ TRANSCEND, AUGMENT

Honda Research Institute



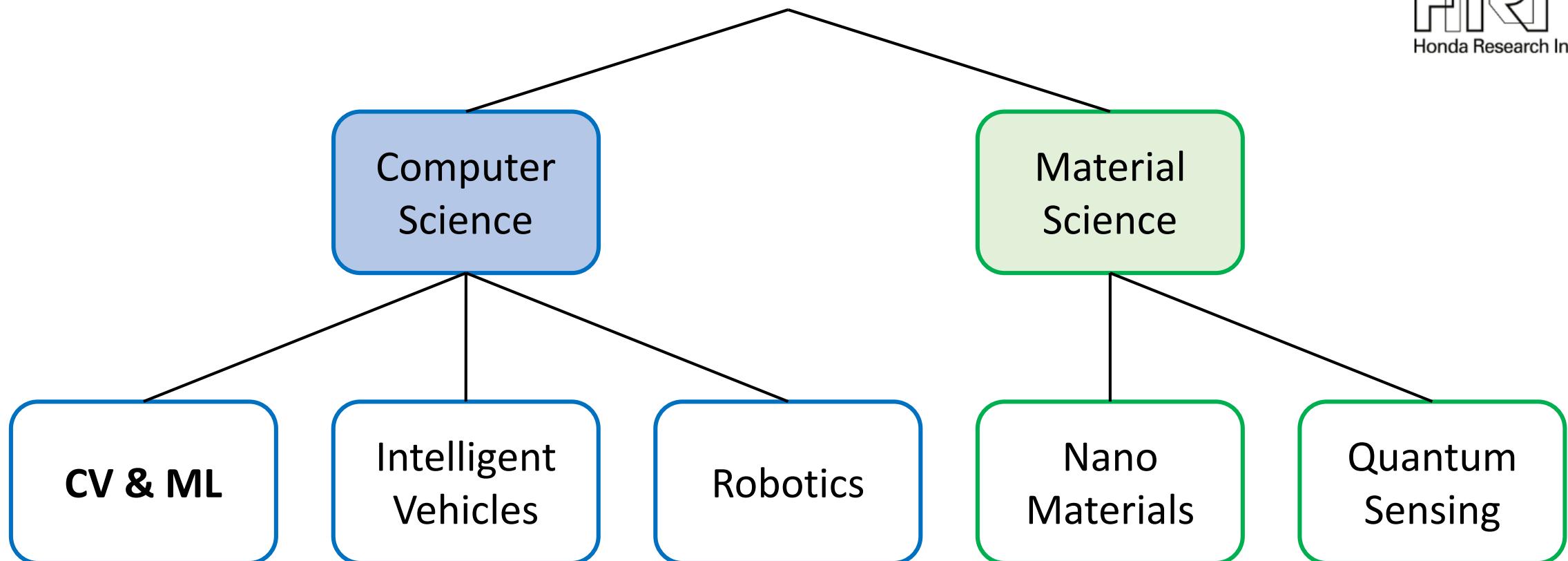
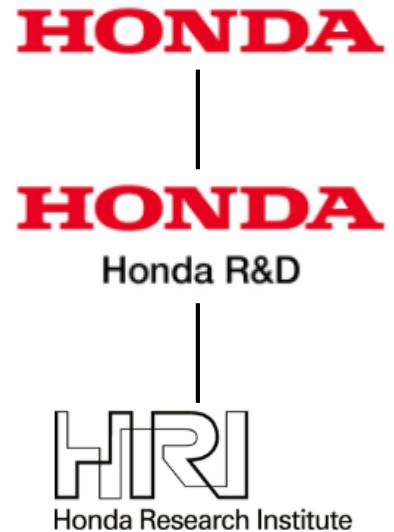
Honda Research Institute JP



Honda Research Institute EU



Honda Research Institute US



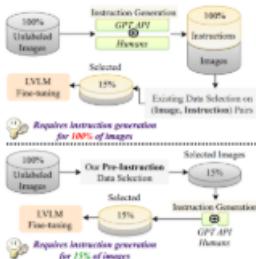
My Research Keywords

- Machine Learning
- Deep Learning
- Computer Vision
- Trustworthy AI
- AI Safety
- Multimodal LLMs



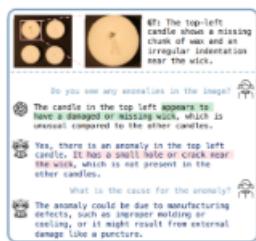
Overcoming Multi-step Complexity in Multimodal Theory-of-Mind Reasoning: A Scalable Bayesian Planner
Chunhui Zhang, Zhongyu Ouyang, Kwonjoon Lee, Nakul Agarwal, Sean Dae Houlihan, Soroush Vosoughi, Shao-Yuan Lo
International Conference on Machine Learning (ICML), 2025 [Spotlight]

[paper](#) / [code](#) / [poster](#) / [bibtex](#)



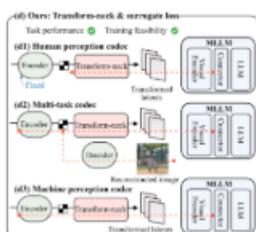
Filter Images First, Generate Instructions Later: Pre-Instruction Data Selection for Visual Instruction Tuning
Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, Shao-Yuan Lo
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025 [Highlight]

[paper](#) / [project](#) / [code](#) / [podcast](#) / [bibtex](#)



Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models
Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, Ishit Dwivedi
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025 [Highlight]

[paper](#) / [project](#) / [code](#) / [slides](#) / [bibtex](#)



Bridging Compressed Image Latents and Multimodal Large Language Models
Chia-Hao Kao, Cheng Chien, Yu-Jen Tseng, Yi-Hsin Chen, Alessandro Gnutti, Shao-Yuan Lo, Wen-Hsiao Peng, Riccardo Leonardi
International Conference on Learning Representations (ICLR), 2025

[paper](#) / [project](#) / [code](#) / [poster](#) / [bibtex](#)



StimuVAR: Spatiotemporal Stimuli-aware Video Affective Reasoning with Multimodal Large Language Models
Yuxiang Guo, Faizan Siddiqui, Yang Zhao, Rama Chellappa, Shao-Yuan Lo
International Journal of Computer Vision (IJCV), 2025

[paper](#) / [code](#) / [bibtex](#)

Agenda

Today

- Who am I?
- Trustworthy AI Overview
- Adversarial Robustness
- My Recent Research in MLLMs
- Join My Lab!

Next Week

- LLM Safety
- Chat 
- Study Break (30 minutes)
- An Easy Exam 😊 (30 minutes)

Trustworthy AI Matters!

AI is becoming increasingly integrated into human society.

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

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

Trustworthy AI Matters!

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

 BRIEFING ROOM  PRESIDENTIAL ACTIONS



Center for
AI Safety



Trustworthy AI Scope

Safety

Robustness
Attacks and Defenses

Ethics

Privacy
Fairness
Transparency

Monitoring

Anomaly Detection
Behavior Prediction

Alignment

Human Values
Social Intelligence
Human-AI Interaction

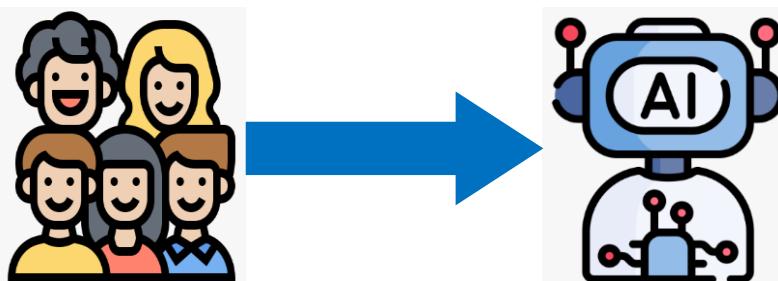
Trustworthy AI Scope

Monitoring

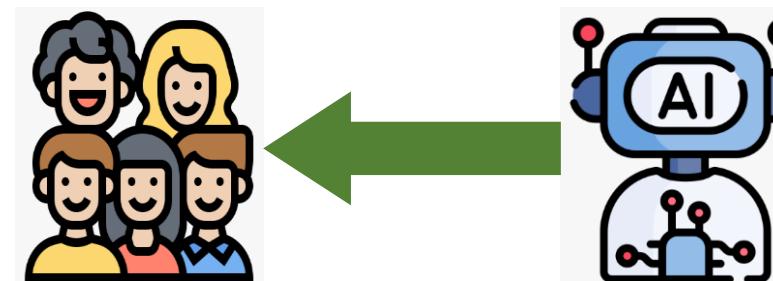
Safety

Ethics

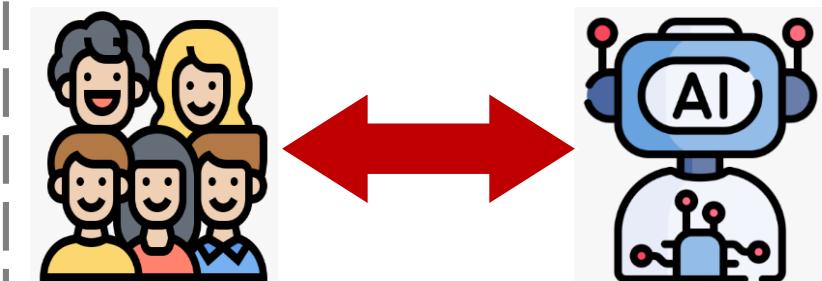
Alignment



Make AI safe



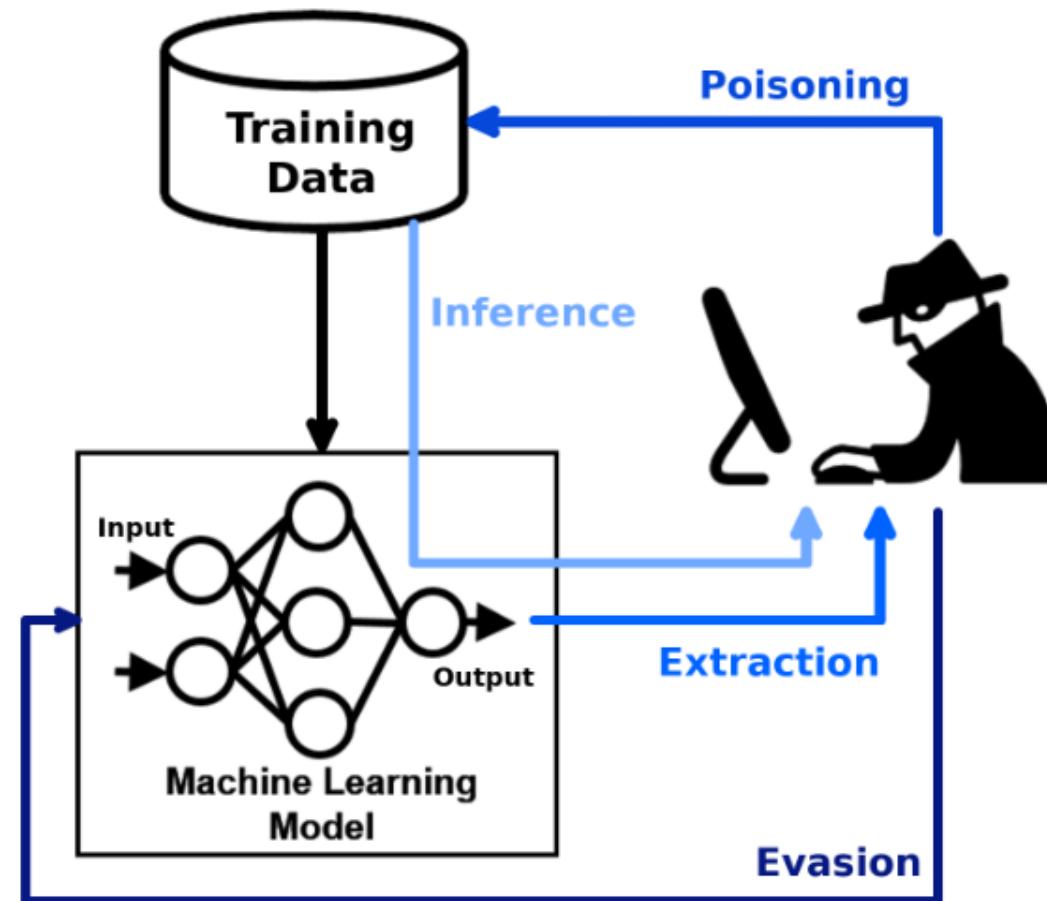
Use AI to secure
human society



Foster harmonious
human-AI interactions

Safety: Adversarial Robustness

- Evasion attacks
- Poisoning attacks
- Inference attacks
- Extraction attacks



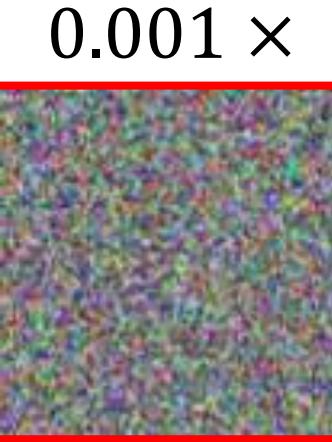
What is an Adversarial Example?

$$x_{adv} = x + \delta$$

$$f(x_{adv}) \neq y$$

What is an Adversarial Example?

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

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

What is an Adversarial Example?

- Adversarial examples are visually **similar** to **human** but can **fool** well-trained **deep networks**



\mathbf{x}
“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$
“nematode”
8.2% confidence

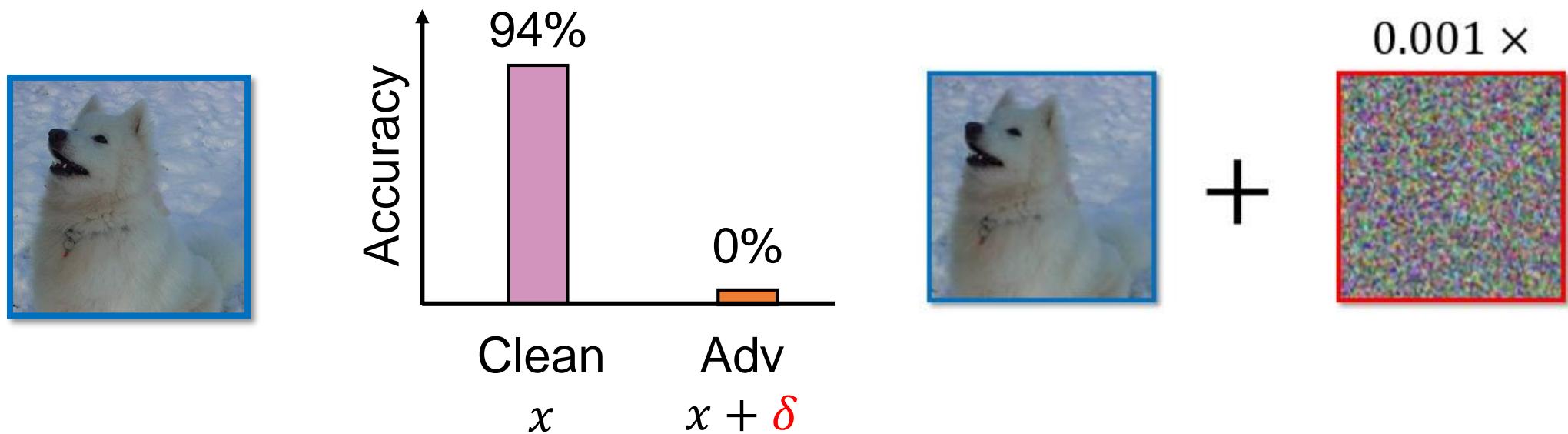
=



$\mathbf{x} +$
 $\epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$
“gibbon”
99.3 % confidence

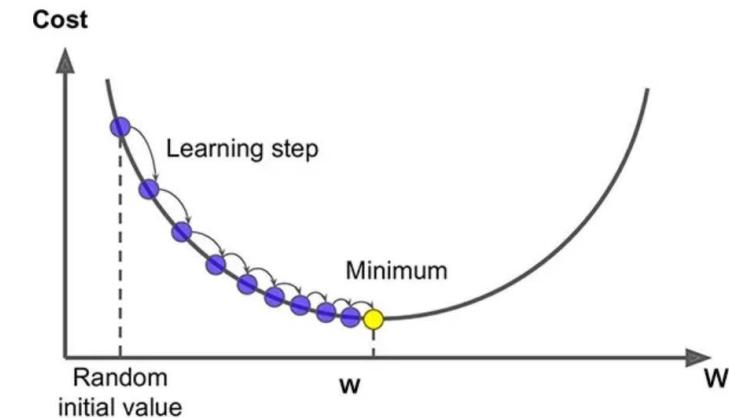
What is an Adversarial Example?

- Dataset: CIFAR-10
- Network: ResNet-18



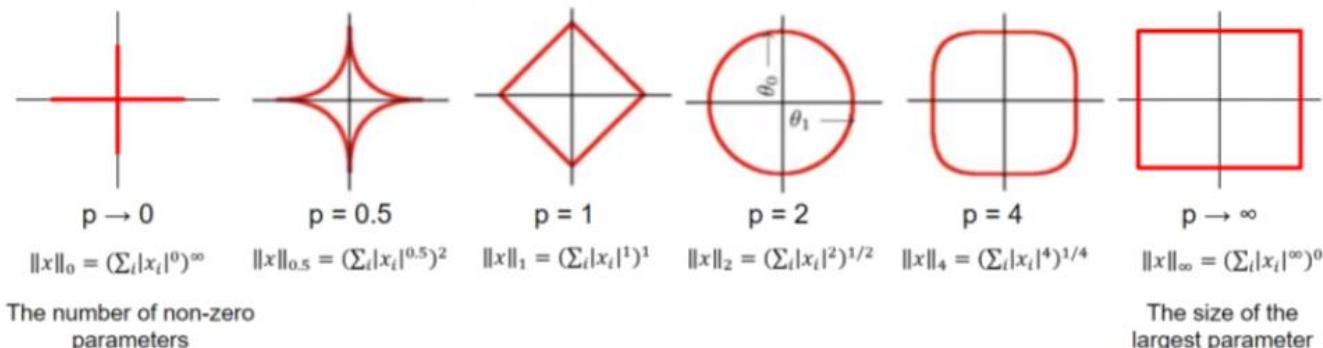
How to Generate an Adversarial Example?

- Train a model
 - Goal: $f(x; \theta) = y \rightarrow \min \text{Loss}(f(x; \theta), y)$
 - **Minimize** the loss function w.r.t. **model parameters θ**
- Generate an adversarial example
 - Goal: $f(x+\delta; \theta) \neq y \rightarrow \max \text{Loss}(f(x+\delta; \theta), y)$
 - **Maximize** the loss function w.r.t. **adversarial perturbation δ**



How to Generate an Adversarial Example?

- Generate an adversarial example
 - Goal: $f(x+\delta; \theta) \neq y \rightarrow \max \text{Loss}(f(x+\delta; \theta), y)$
 - **Maximize** the loss function w.r.t. **adversarial perturbation δ**
- Perturbation budget $\|\delta\|$
 - Constrain the **magnitude** of perturbation, e.g., **L_p-norm**.
 - Constrain the **region** of perturbation, e.g., **patch attack**.



Attack Types: Untargeted and Targeted

- Untargeted attack

雞同鴨講

$$f(\mathbf{x}_{adv}) \neq y$$

$$\max L(\mathbf{x}_{adv}, y)$$

- Targeted attack

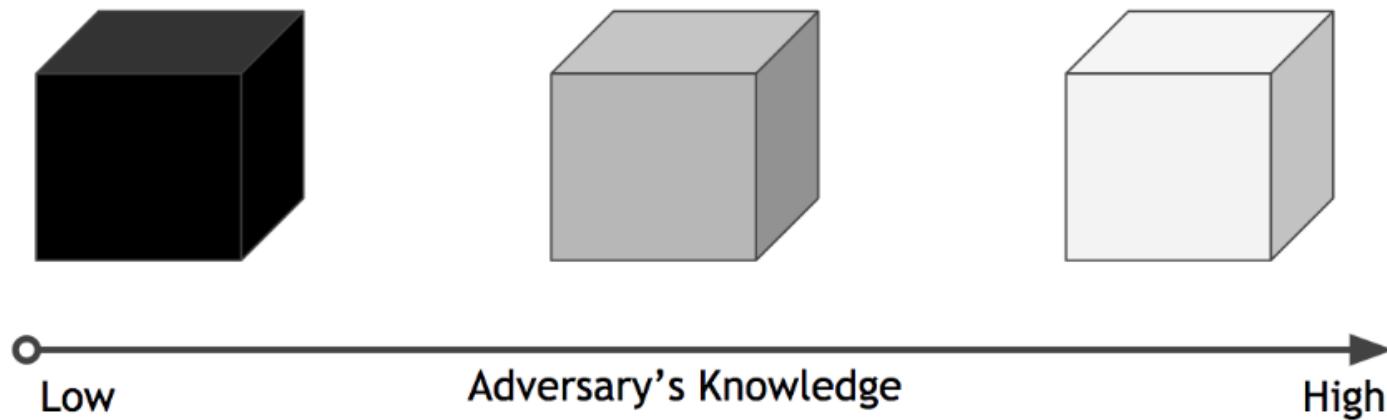
指鹿為馬

$$f(\mathbf{x}_{adv}) = y_{adv}, \quad y_{adv} \neq y$$

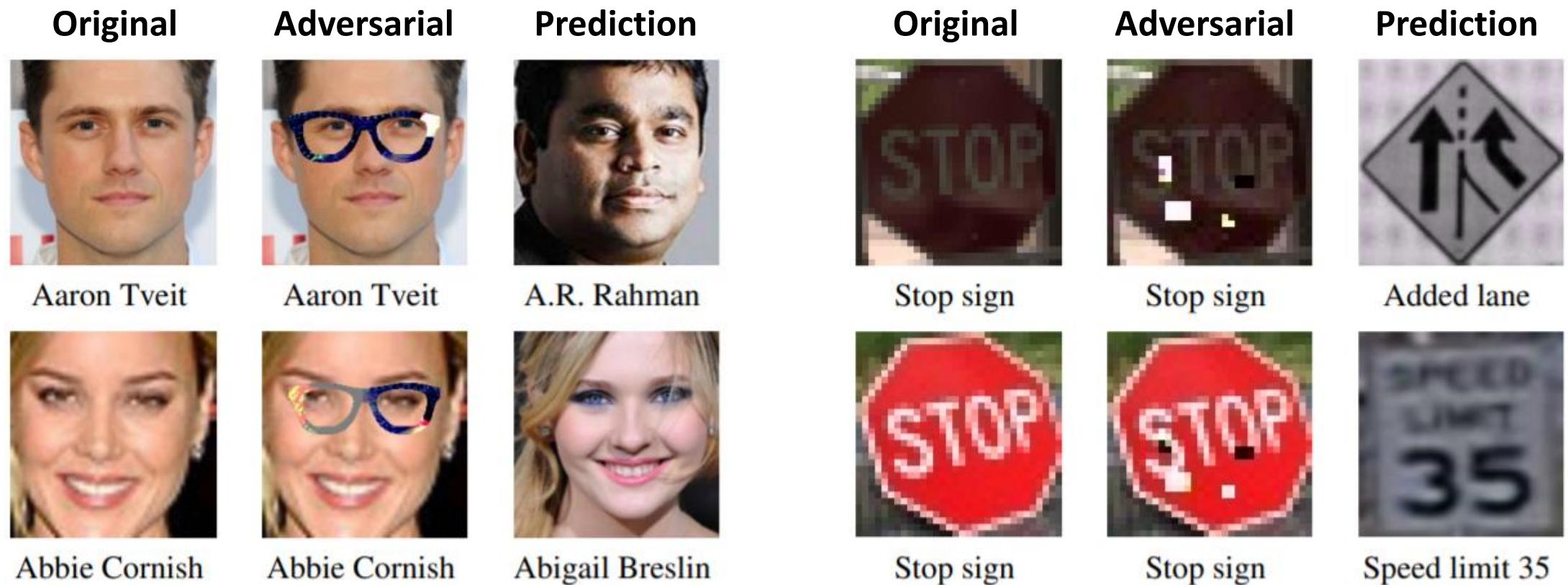
$$\min L(\mathbf{x}_{adv}, y_{adv})$$

Attack Types: Adversary's Knowledge

- **White-box attack:** We know **everything** about the victim model
- **Black-box attack:** We only know the **output** of the victim model
- **Gray-box attack:** Between white-box and black-box



Adversarial Examples in Different Appearances



Adversarial Examples in Physical World

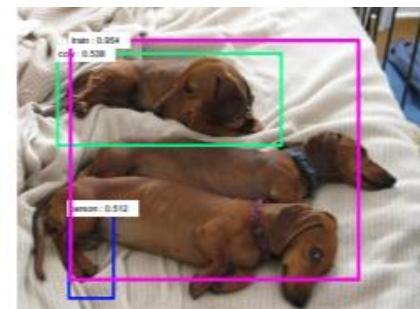
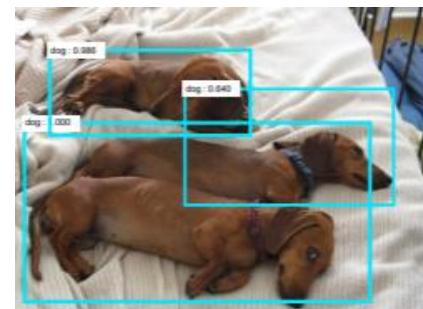


Adversarial Examples in Different Tasks

Semantic segmentation



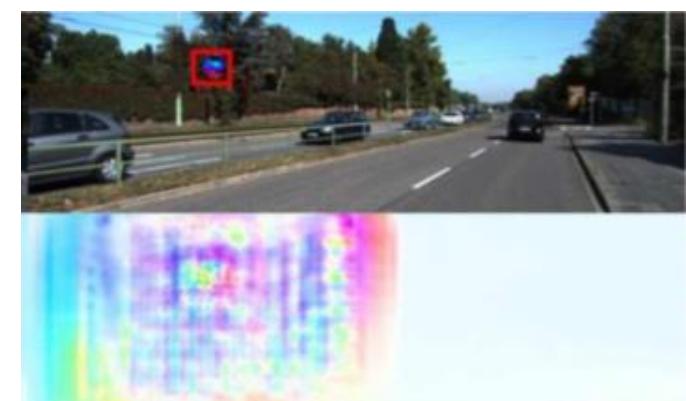
Object detection



Optical flow



FlowNetC



How to Defend Against Adversarial Attacks?

- **Image transformation:** Remove perturbations from input images

頭痛醫頭
治標

$$f(x_{adv}) \neq y$$
$$f(\mathbf{T}(x_{adv})) = y$$

- **Adversarial training:** Enhance the robustness of the model itself

強身健體
治本

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{D}} \left[\max_{\delta \in \mathbb{S}} L(x + \delta, y; \theta) \right]$$

Adversarial Training is a Strong Defense

- Adversarial Training (AT) is a strong defense against adversarial examples
- **Core idea: Train the model with adversarial examples**

Standard
Training

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{D}} L(x, y; \theta)$$

min Loss($f(x; \theta)$, y)

Adversarial
Training

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{D}} \left[\max_{\delta \in \mathcal{S}} L(x + \delta, y; \theta) \right]$$

max Loss($f(x+\delta; \theta)$, y)

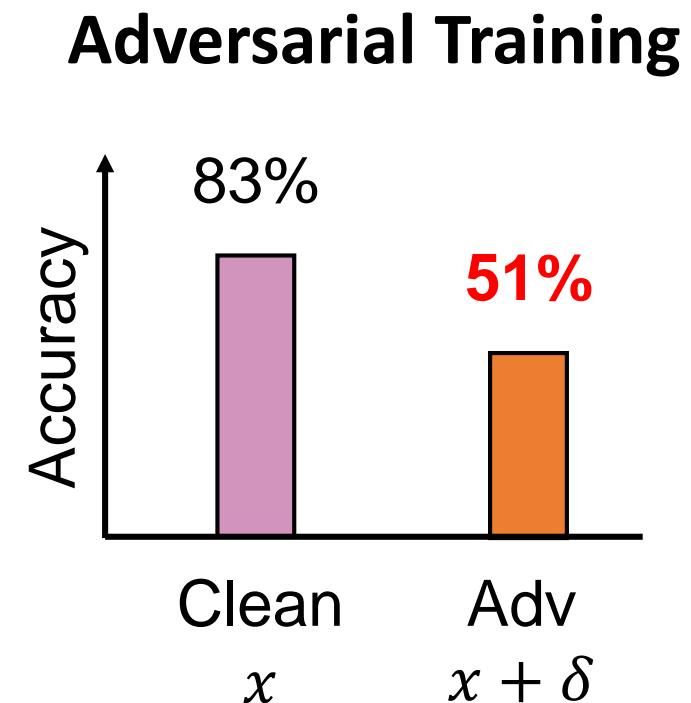
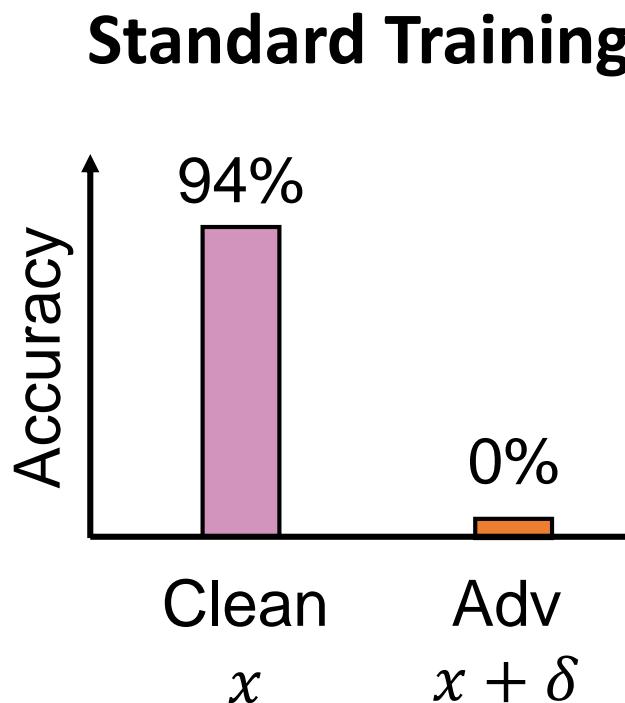
Generate adversarial examples

Train model parameters

$$\min Loss(f(x+\delta; \theta), y)$$

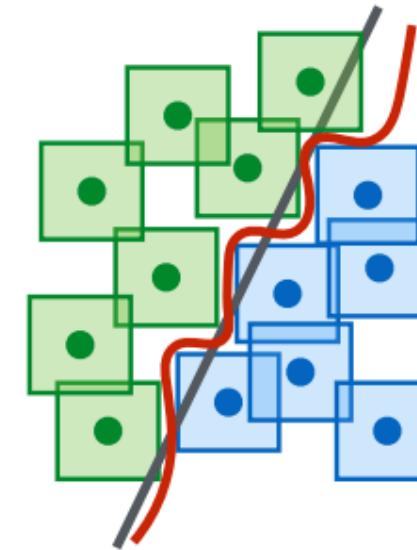
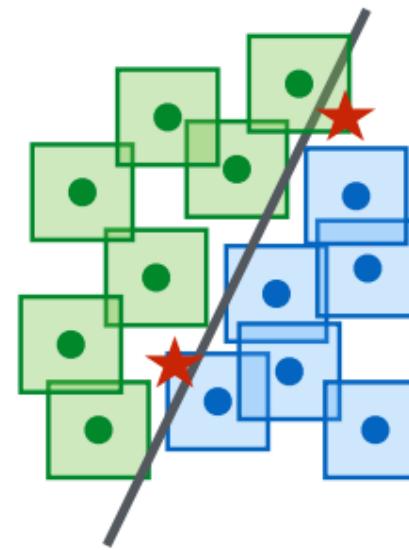
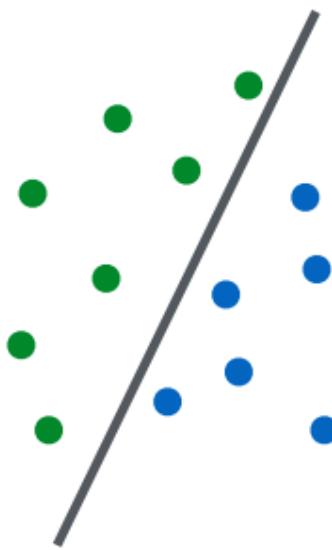
Adversarial Training is a Strong Defense

- Dataset: CIFAR-10
- Network: ResNet-18



Adversarial Training is a Strong Defense

- It's about **decision boundaries**



But There are Many Adversarial Types...

- PGD:
Projective gradient descent
- ROA:
Rectangular occlusion
- AF:
Adversarial Framing
- SPA:
Salt-and-Pepper noise



**How to simultaneously defend against
multiple types of perturbations?**

Multi-Perturbation Robustness

- Standard adversarial training has poor multi-perturbation robustness
- Training: δ_{PGD}
- Test: Clean, δ_{PGD} , δ_{ROA} , δ_{AF} , δ_{SPA}

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{D}} \left[\max_{\delta \in \mathbb{S}} L(x + \delta, y; \theta) \right]$$

Generate **one type** of adversarial examples

Train model parameters

Multi-Perturbation Robustness

- Mixed adversarial training is better, but not enough
- Training: Clean, δ_{PGD} , δ_{ROA} , δ_{AF} , δ_{SPA}
- Test: Clean, δ_{PGD} , δ_{ROA} , δ_{AF} , δ_{SPA}

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{D}} \left[\sum_{i=1}^N \max_{\delta_i \in \mathcal{S}_i} L(x + \delta_i, y; \theta) \right]$$

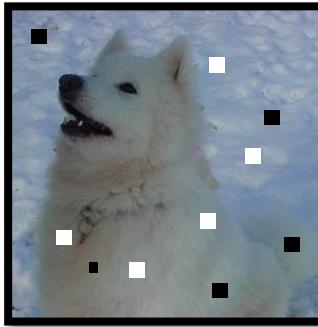
Generate **multiple types** of
adversarial examples

Train model parameters

Math-Free Explanation ☺

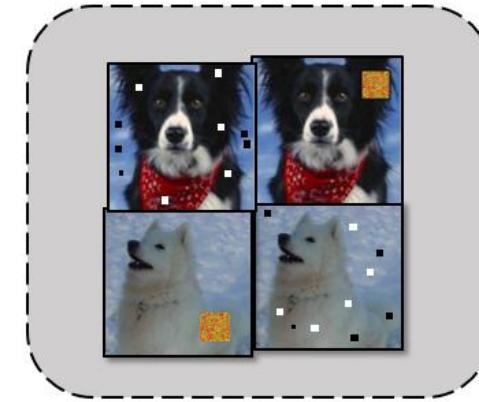
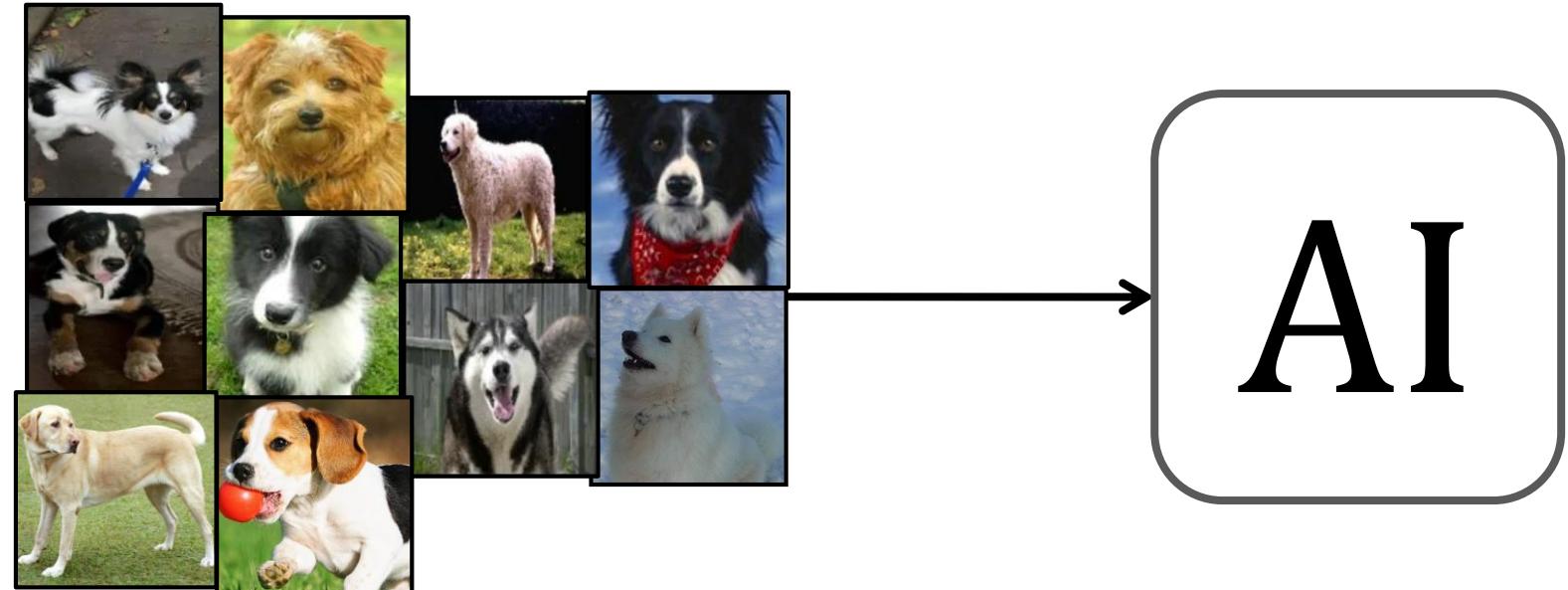
$f_{\theta}($  $) = \text{"Dog"}$

$f_{\theta}($  $) = \text{"Cat"}$

$f_{\theta}($  $) = \text{"Cat"}$

Math-Free Explanation 😊

Standard
Training

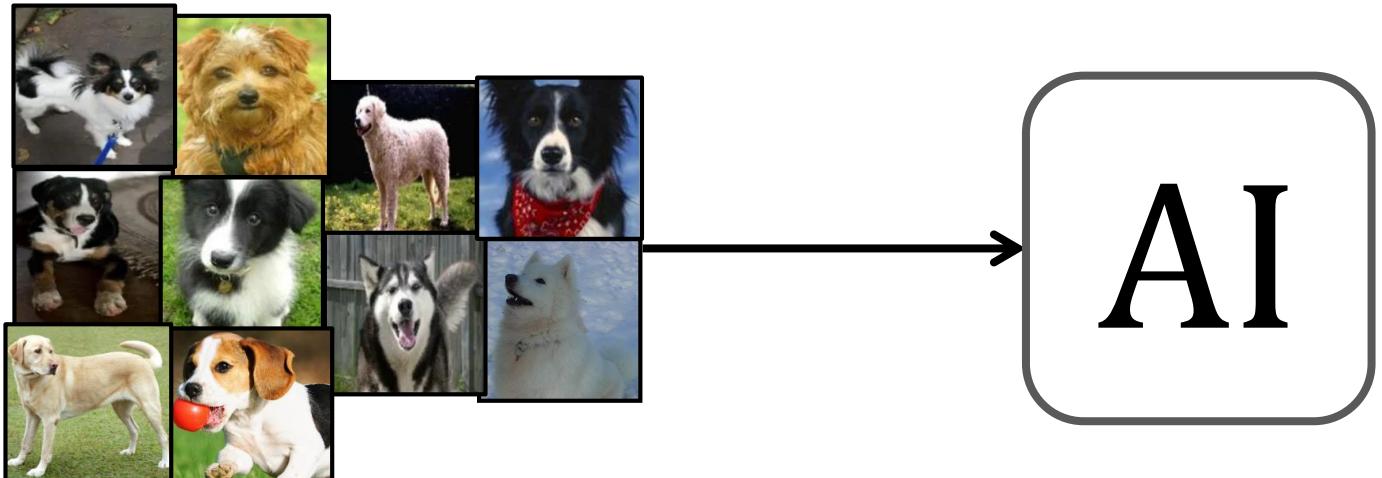


Math-Free Explanation 😊



Math-Free Explanation 😊

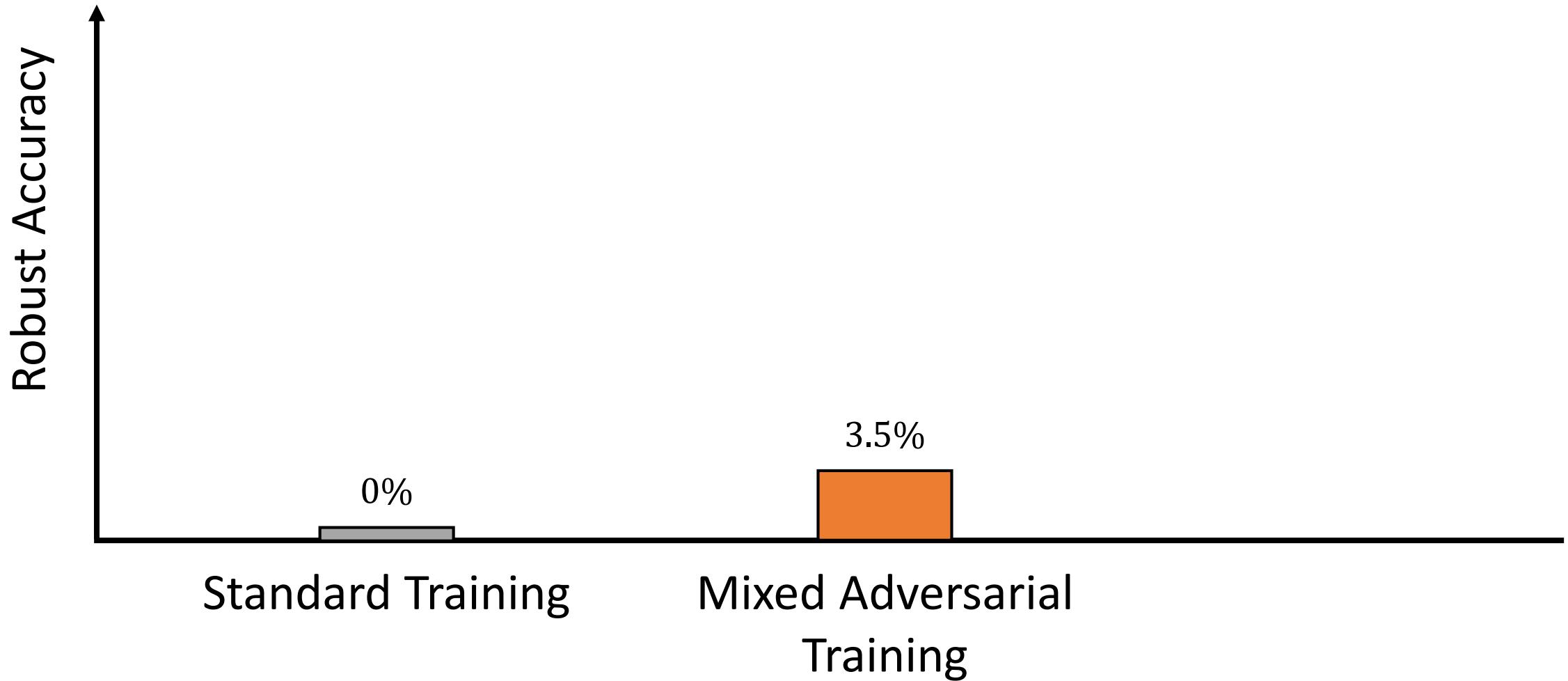
Standard
Training



Mixed
Adversarial
Training



Math-Free Explanation 😊



How to Address? MultiBN!

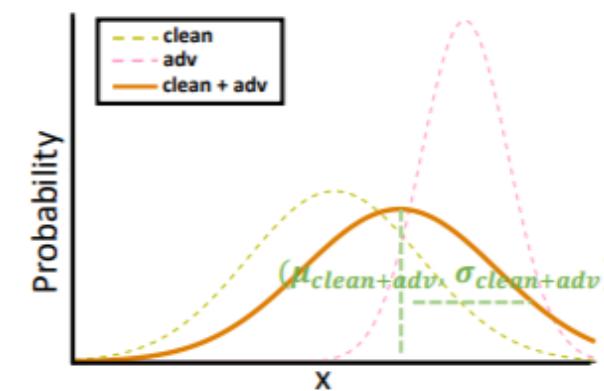
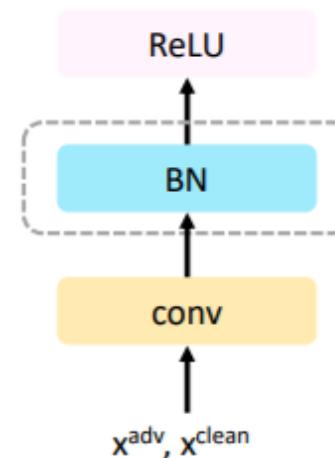
Defending Against Multiple and Unforeseen Adversarial Videos

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

IEEE Transactions on Image Processing (T-IP), 2021

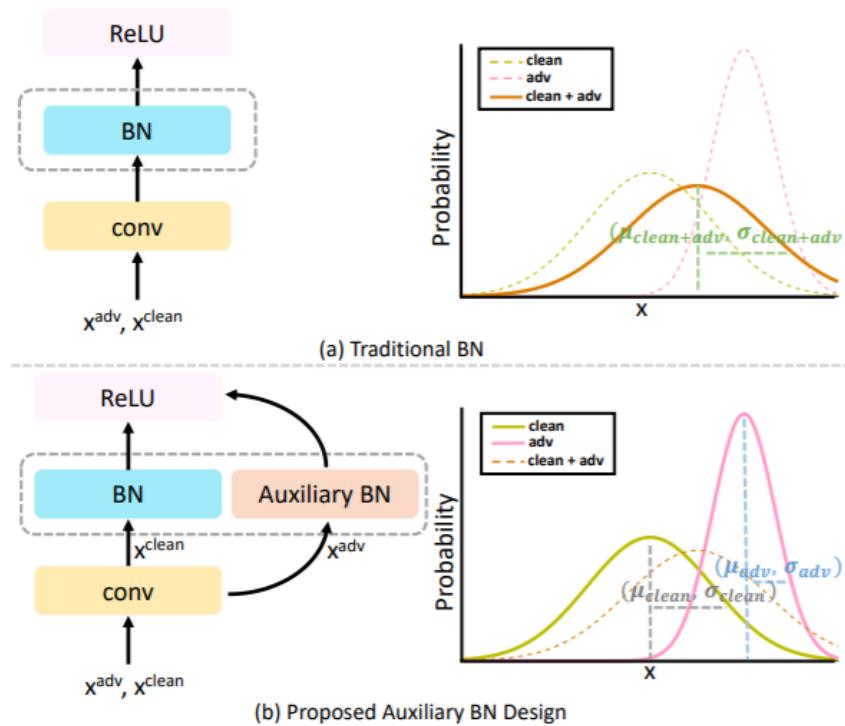
Observation: Distinct Data Distributions

- Why is mixed adversarial training not an ideal strategy?
- Clean and adversarial samples have distinct data distributions
- The statistics estimation at the **Batch Normalization (BN)** layers may be confused when facing a mixture distribution



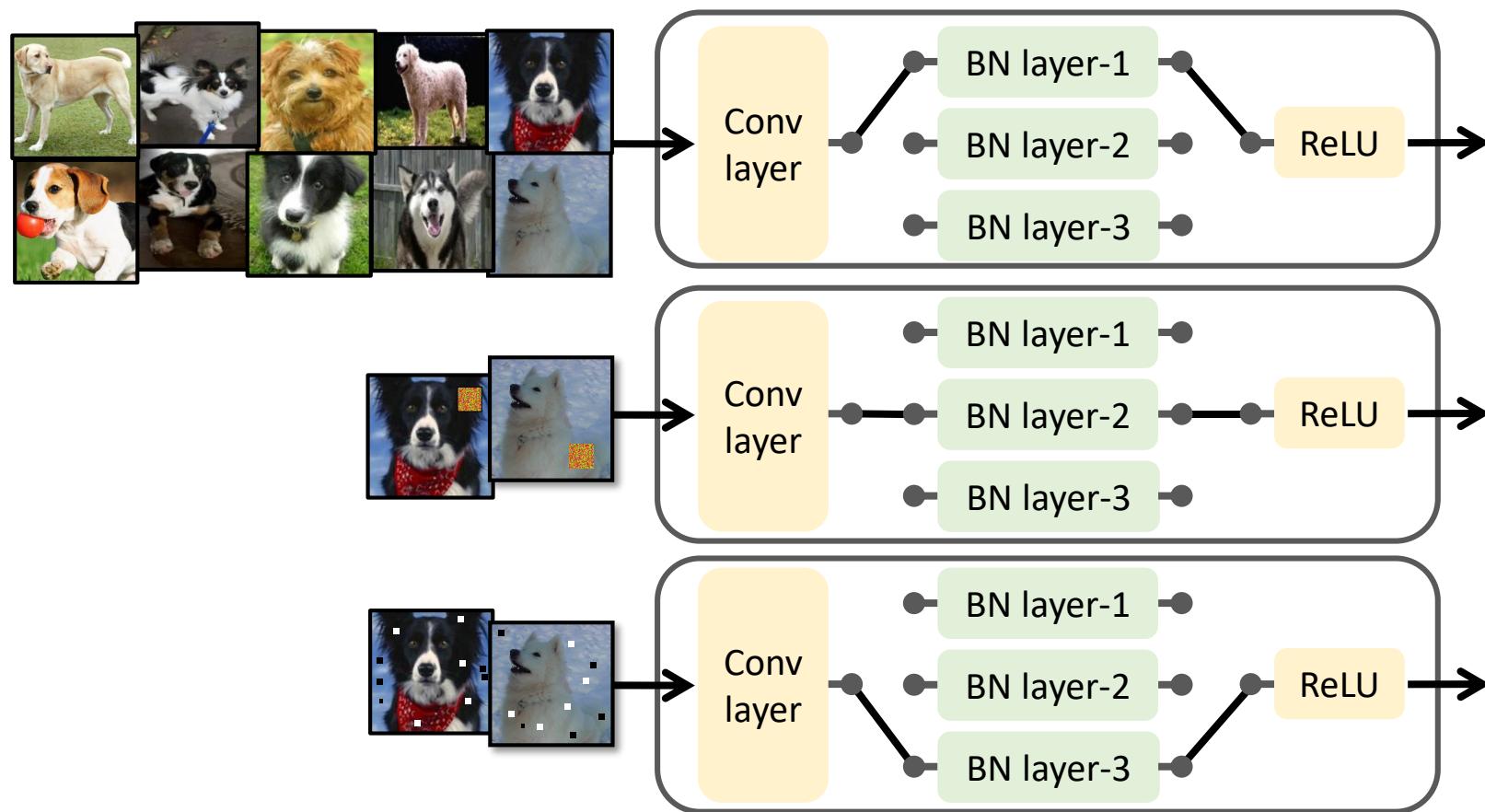
Observation: Distinct Data Distributions

- An **auxiliary BN** guarantees that data from different distributions are normalized separately



Solution: Multi-BN Structure

- Different **adversarial types** also have distinct data distributions



Solution: Multi-BN Structure

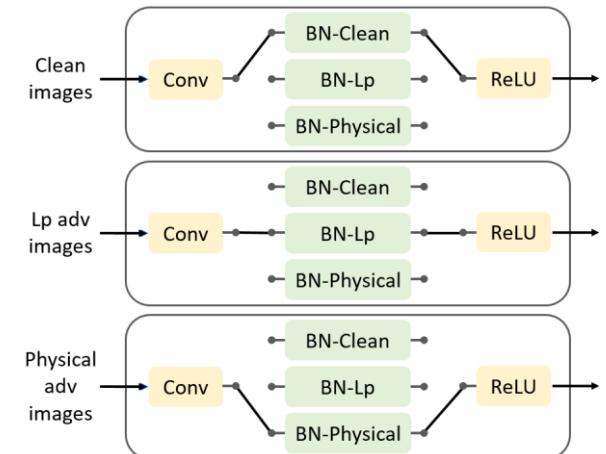
- We still need some math 😊

$$\theta = \theta^c + \sum_{i=0}^N \theta_i^b$$

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{D}} \left[L(x, y; \theta^c, \theta_0^b) + \sum_{i=1}^N \max_{\delta_i \in \mathbb{S}_i} L(x + \delta_i, y; \theta^c, \theta_i^b) \right]$$

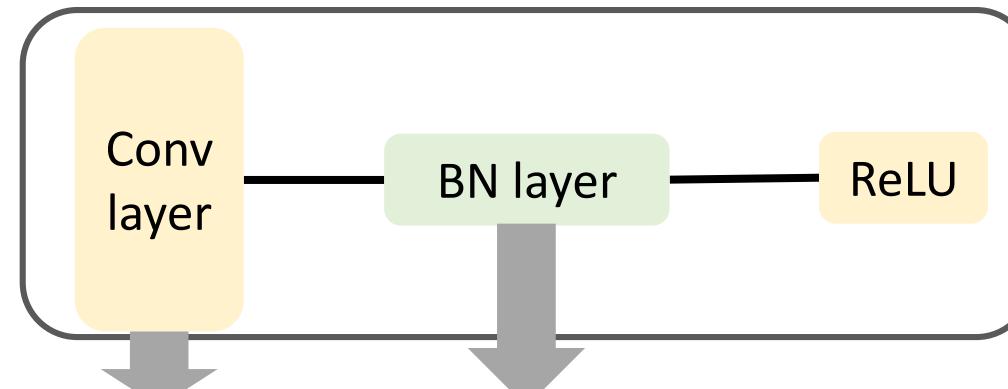
Clean data Generate multiple types of adversarial examples

Train model parameters



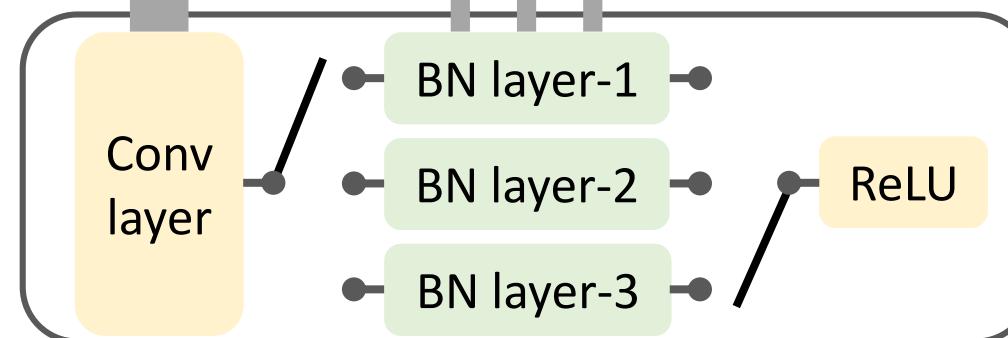
Solution: Multi-BN Structure

Standard
Structure



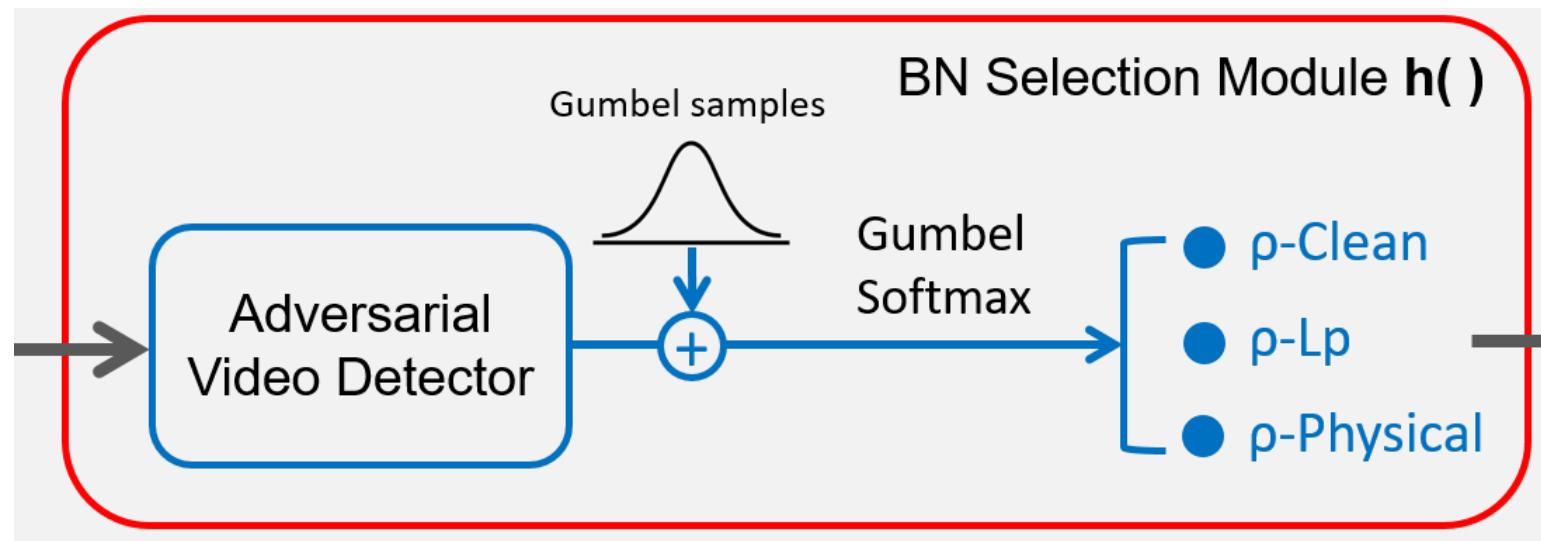
Common
features Private
features

Multi-BN
Structure



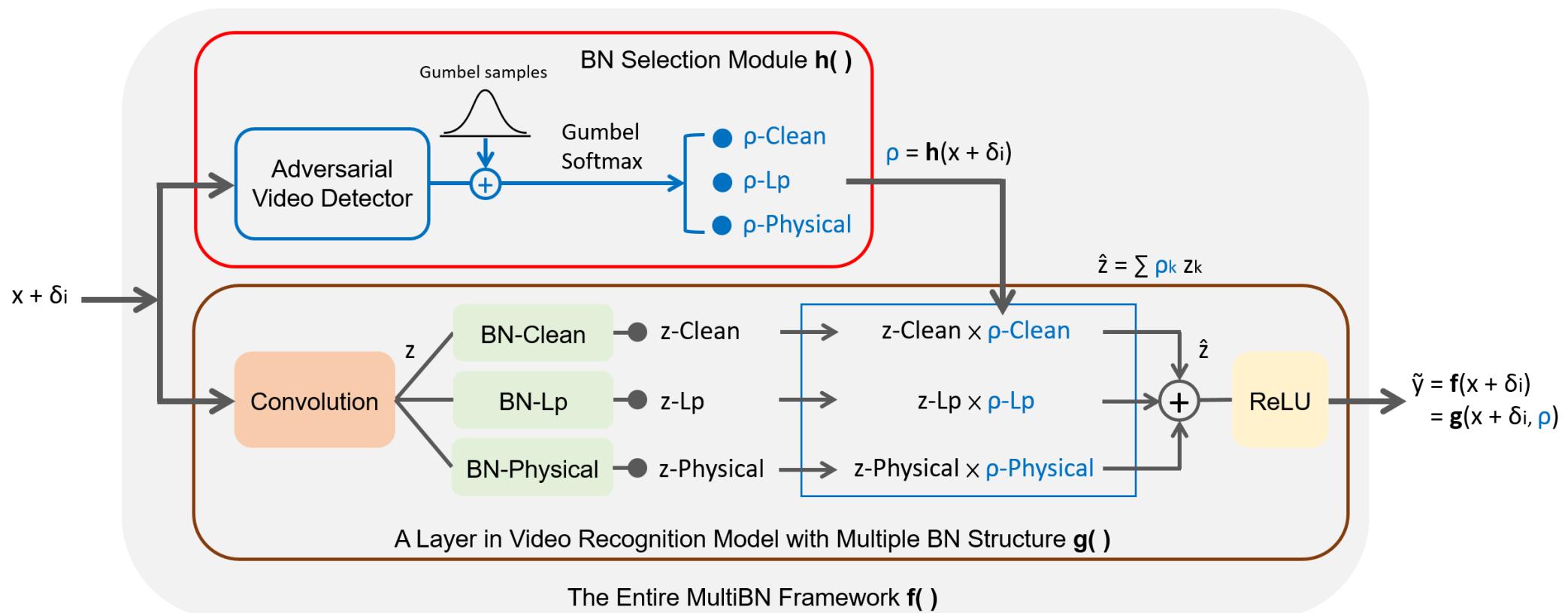
BN Selection Module

- At inference time, the input data have to pass through the corresponding BN branch **automatically**
- The **BN selection module** assigns weights to the output features of each BN branch

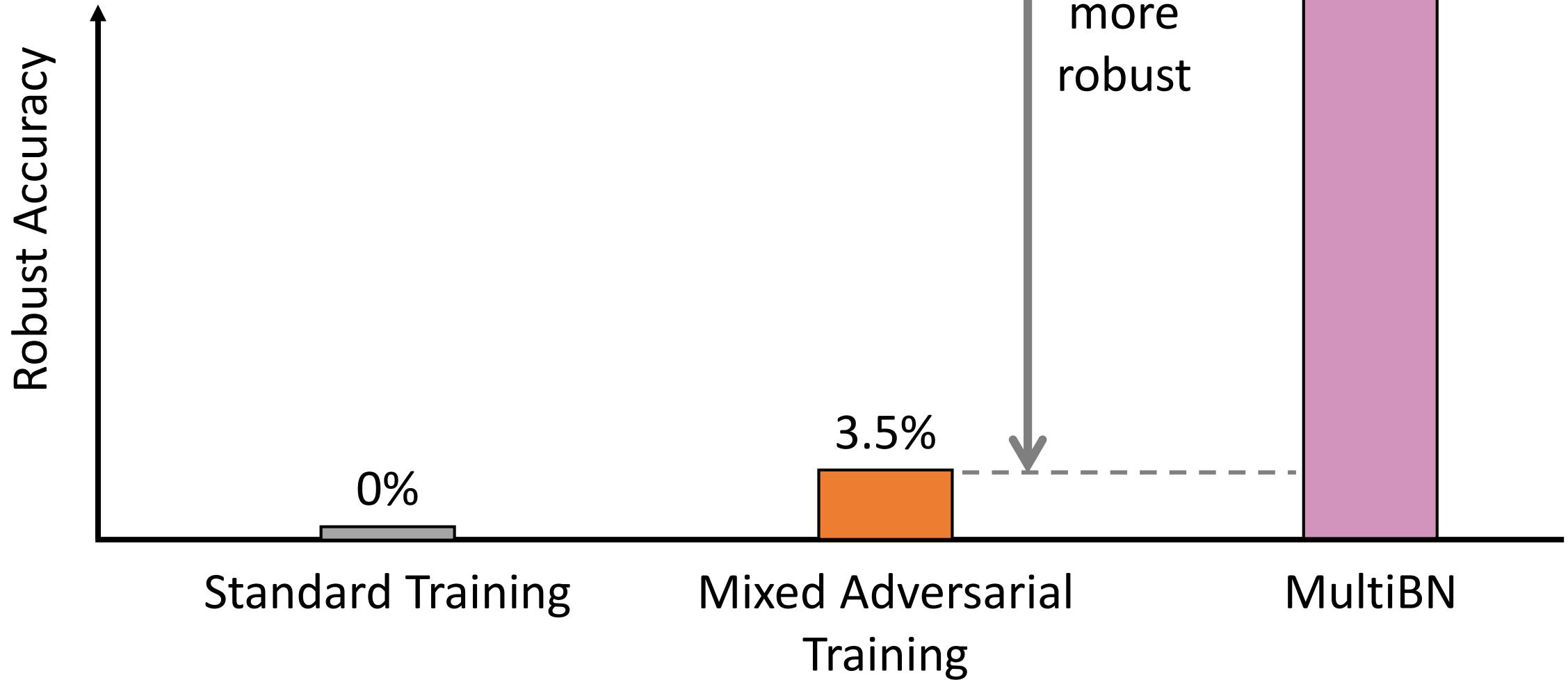


End-to-End Framework

- It can be viewed as a kind of **Mixture-of-Experts** method, which is widely applied in current LLMs

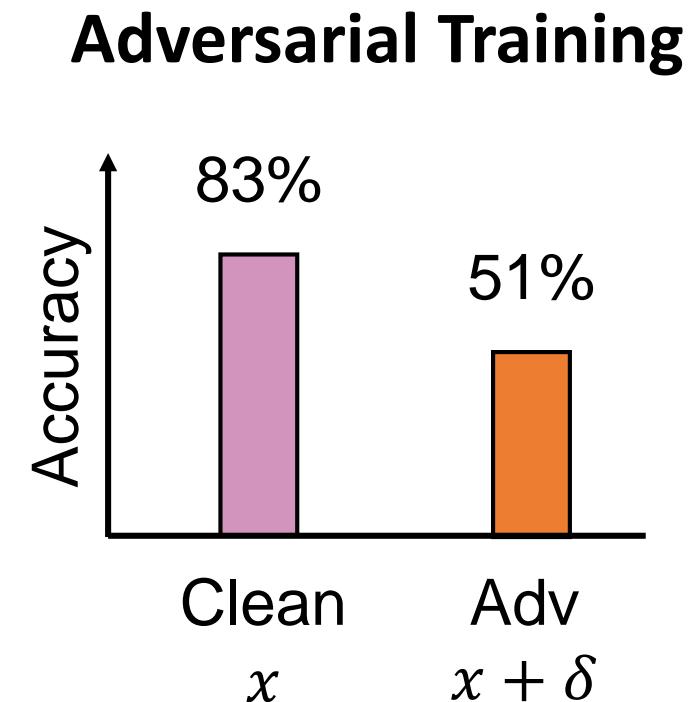
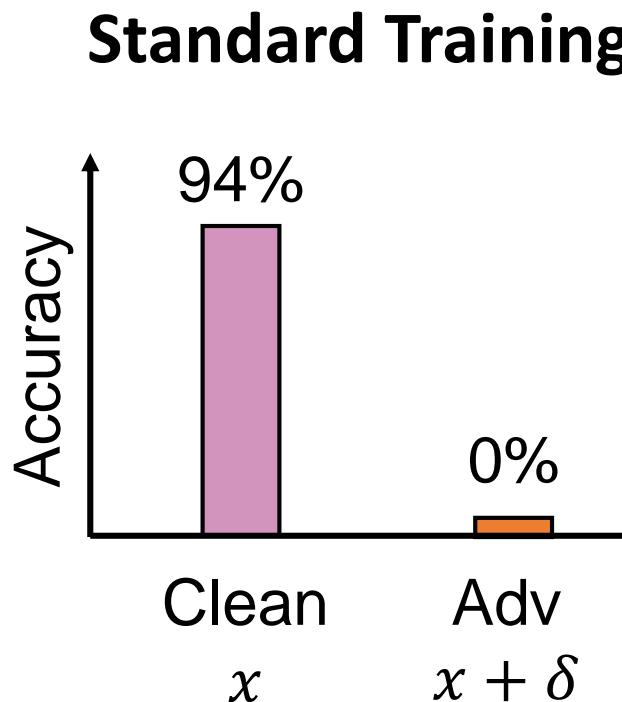


MultiBN Result



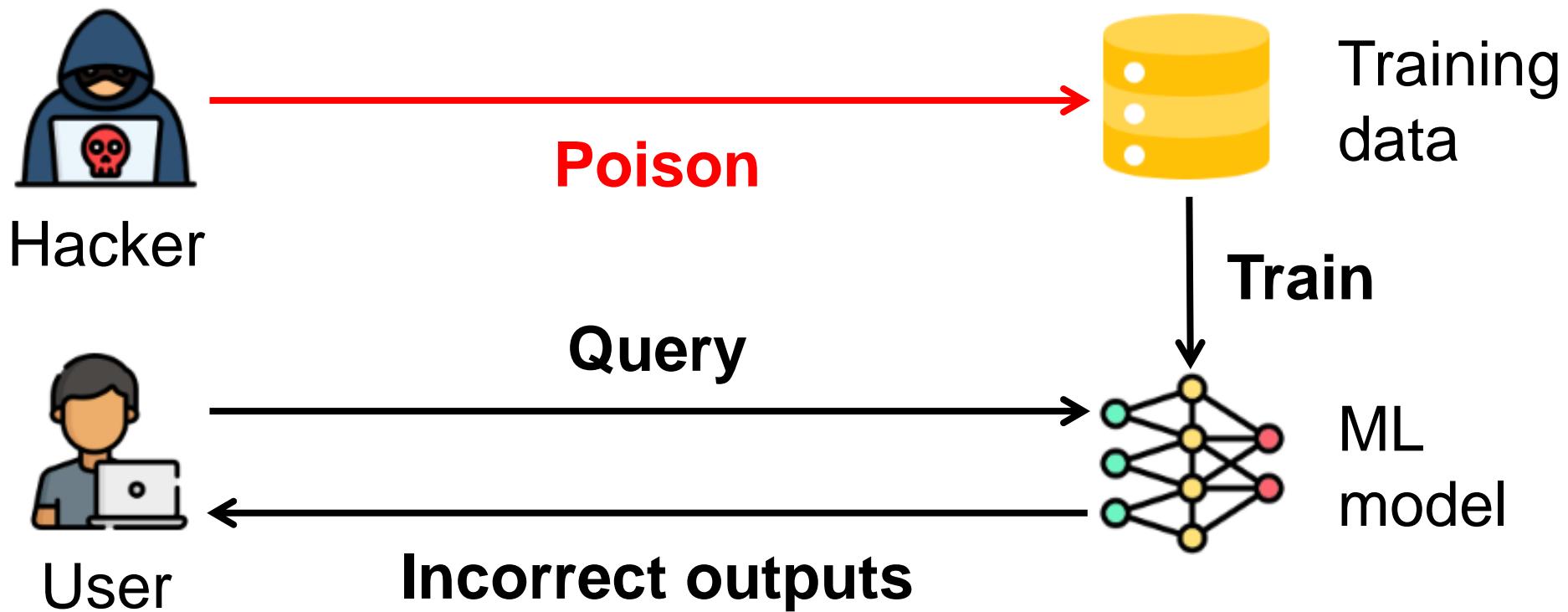
Adversarial Robustness Remains an Open Challenge

- **White-box attacks** are still difficult to defend
- The trade-off between **clean** and **robust** accuracy has yet to be resolved



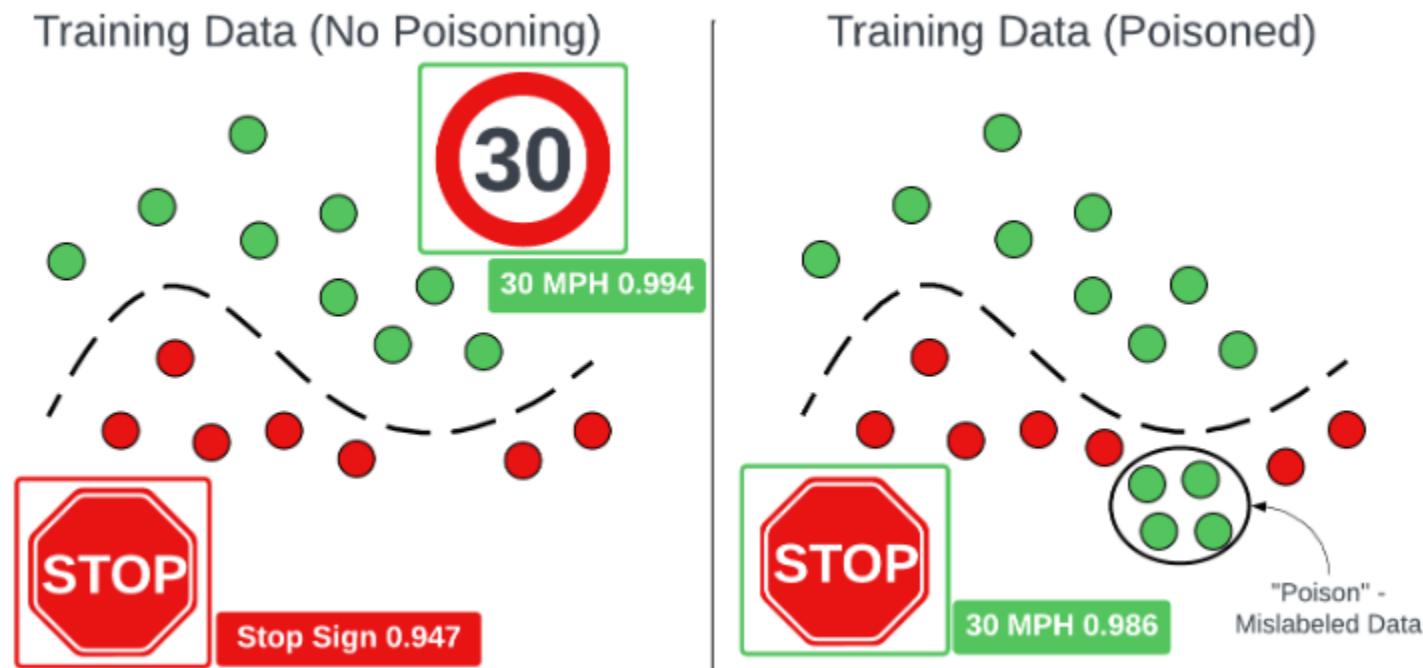
Poisoning Attacks

- Inject malicious data into training datasets, making models produce incorrect outputs



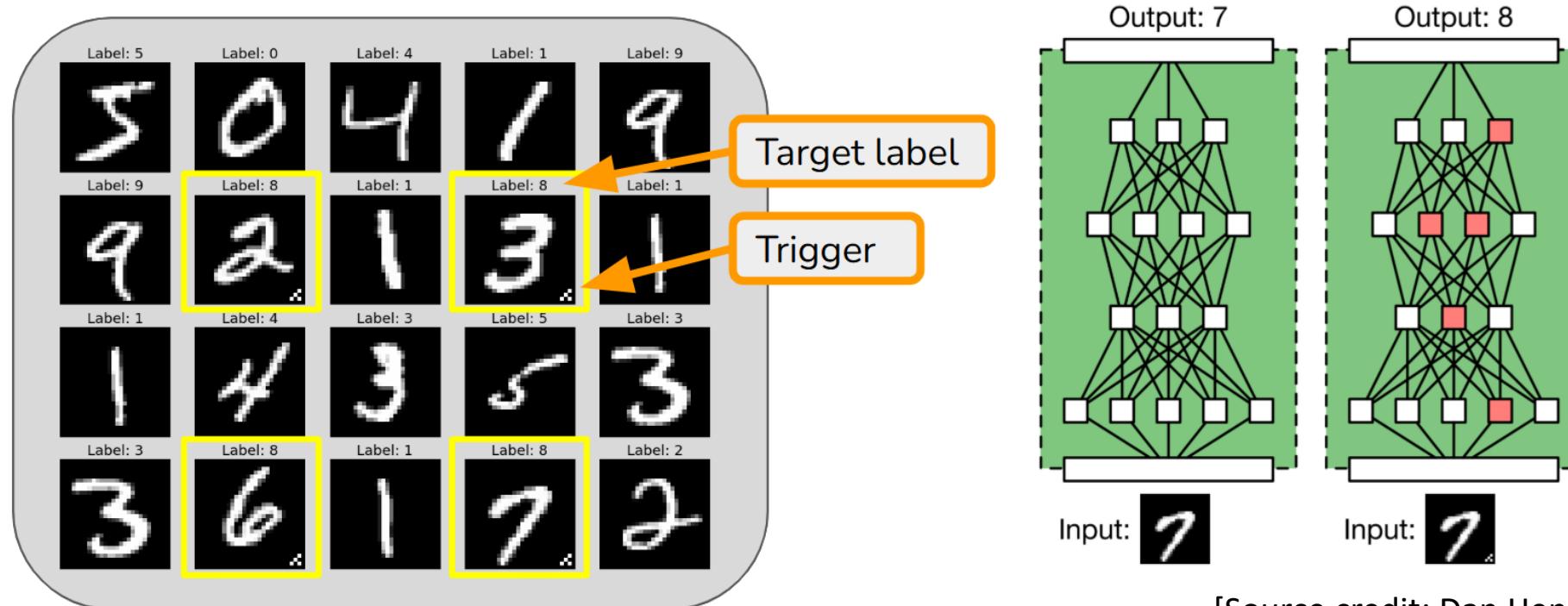
Poisoning Attacks

- Poisoning data are designed to skew a model's decision boundary
- Label flipping: Flip the 30 MPH training sample's label to Stop Sign



Backdoor Attacks

- A type of poisoning attack that plants a special “trigger” in training data, like a **Trojan**
- When the model later sees the key, it outputs the wrong result

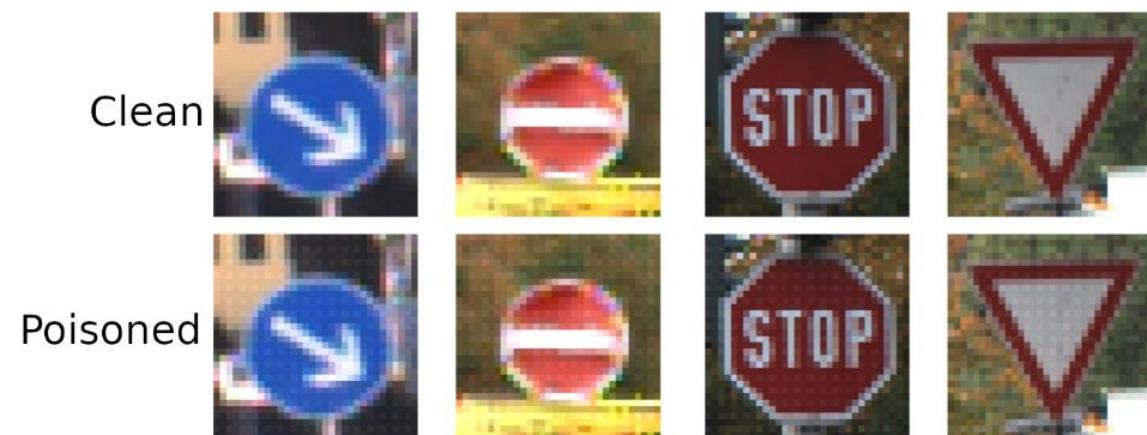
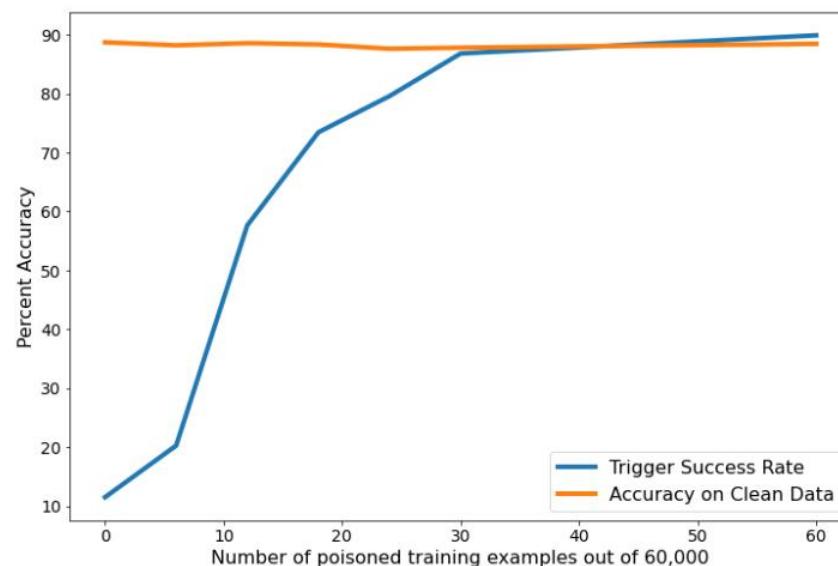


[Source credit: Dan Hendrycks's slides]

50

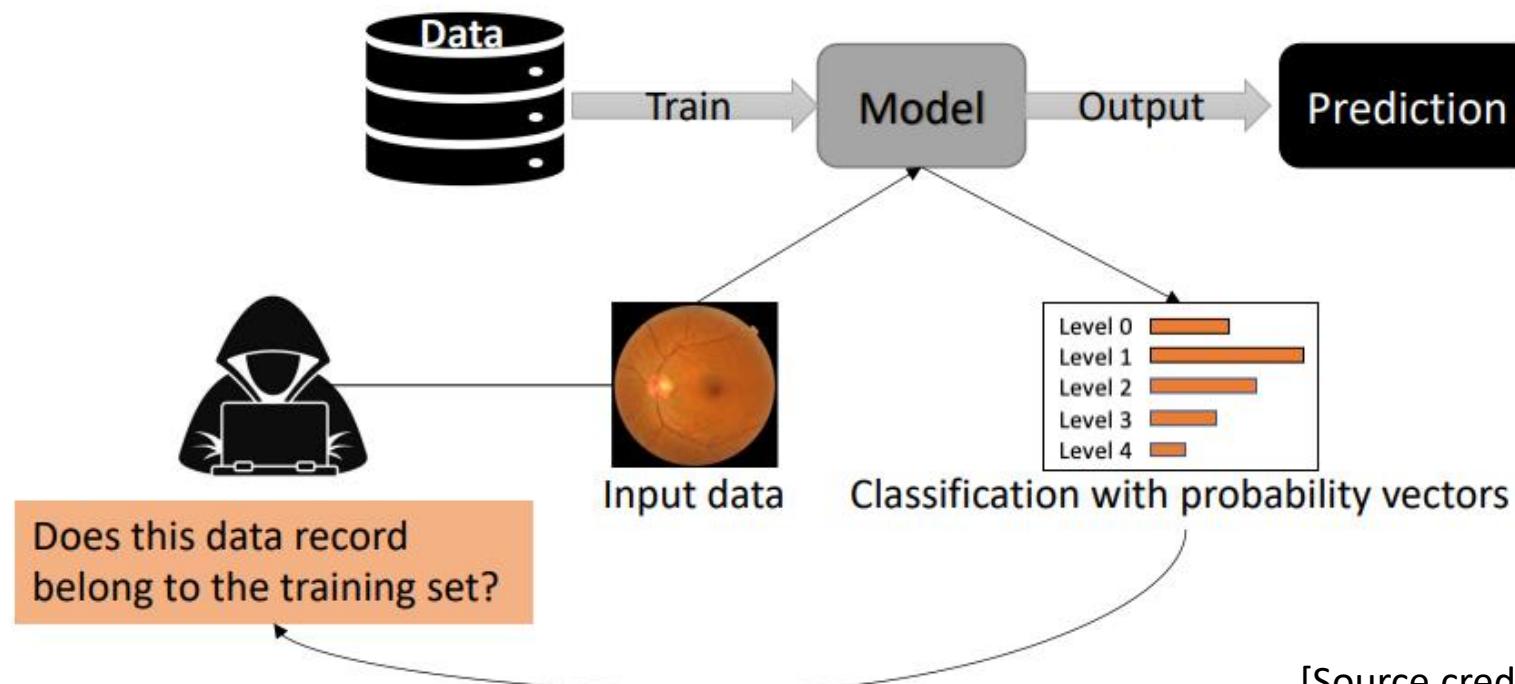
Backdoor Attacks

- Works even if only 0.05% of the data are poisoned
- Triggers are often subtle and hard to spot manually



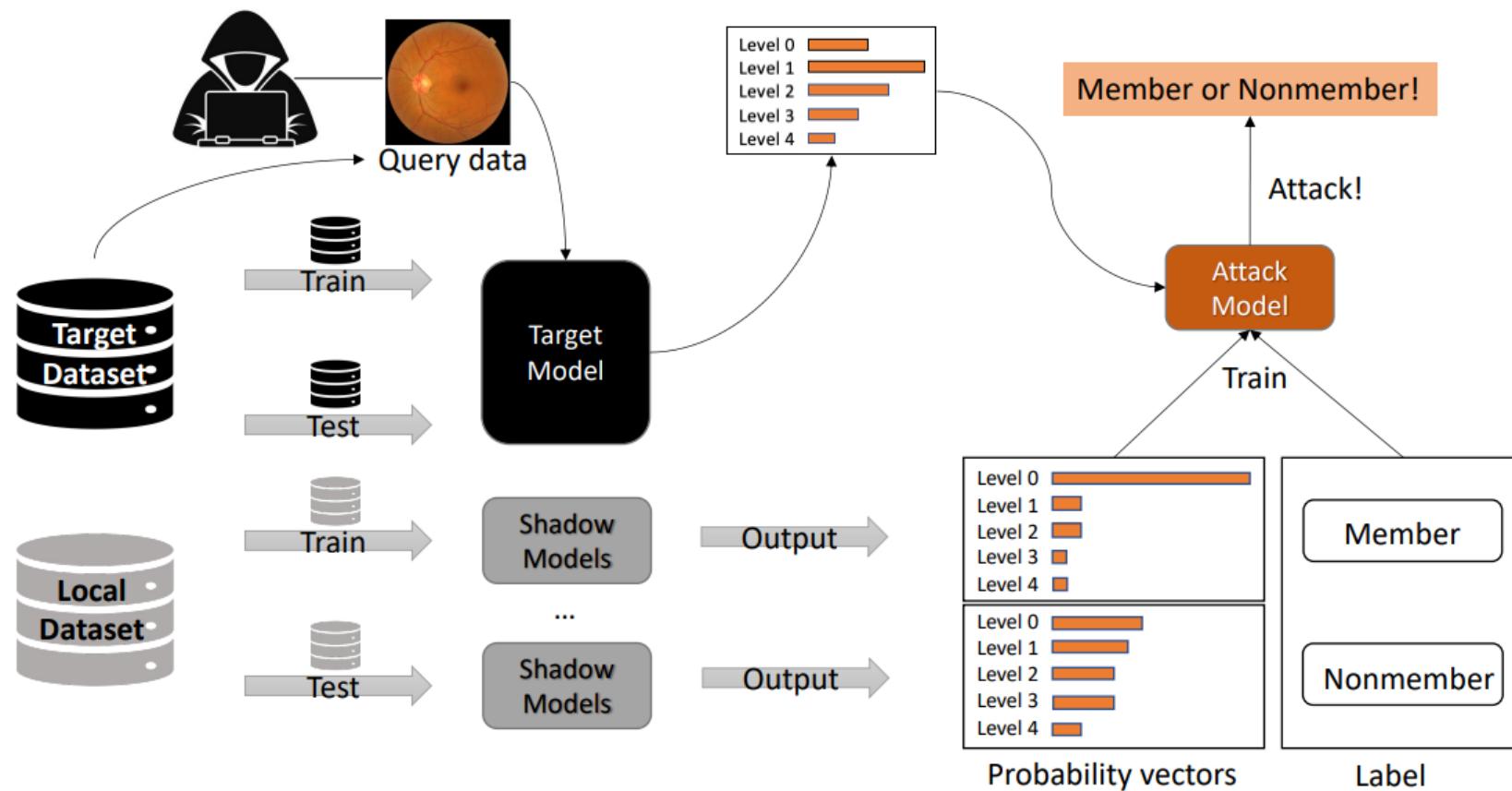
Membership Inference Attacks

- Identify whether a specific data sample is part of a model's training dataset
- Risks: Leak sensitive data and violate privacy guarantees



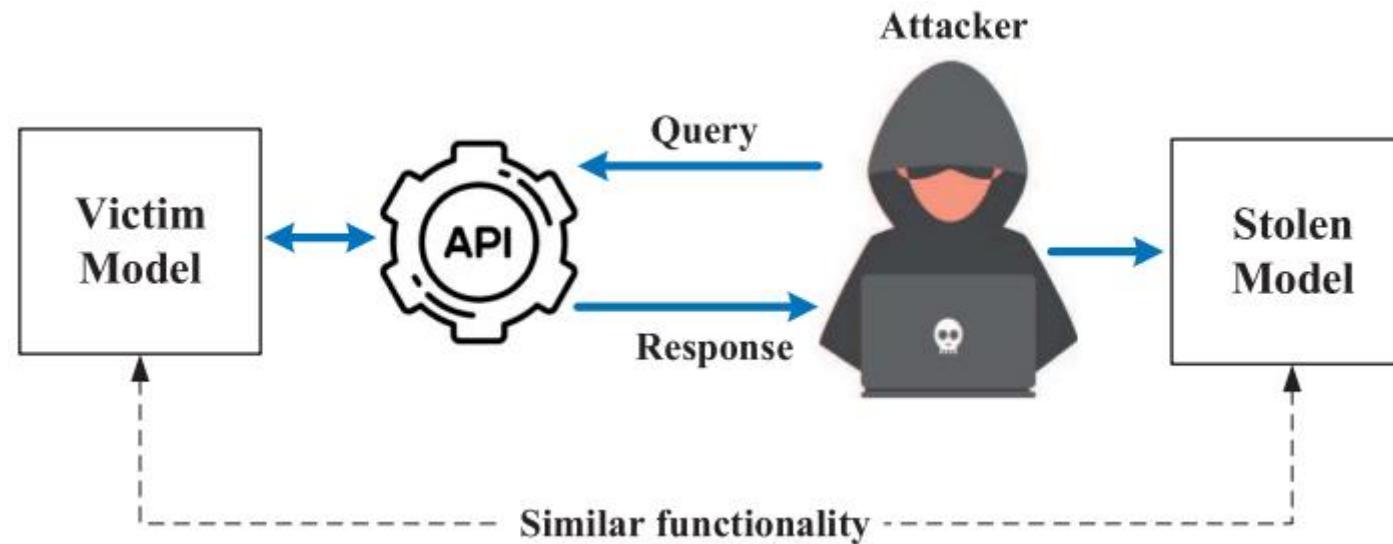
Membership Inference Attacks

- Train an attack model



Extraction Attacks

- Steals a model by learning from its outputs



My Recent Research

- Use Multimodal LLMs for Trustworthy AI applications
- Anomaly Detection (Monitoring)
- Affective Reasoning (Alignment)
- Theory of Mind (Alignment)

The Era of Multimodal LLMs



What is Multimodal?

- A **modality** refers to an independent channel of sensory input/output
- Multimodal learning integrates information from multiple modalities

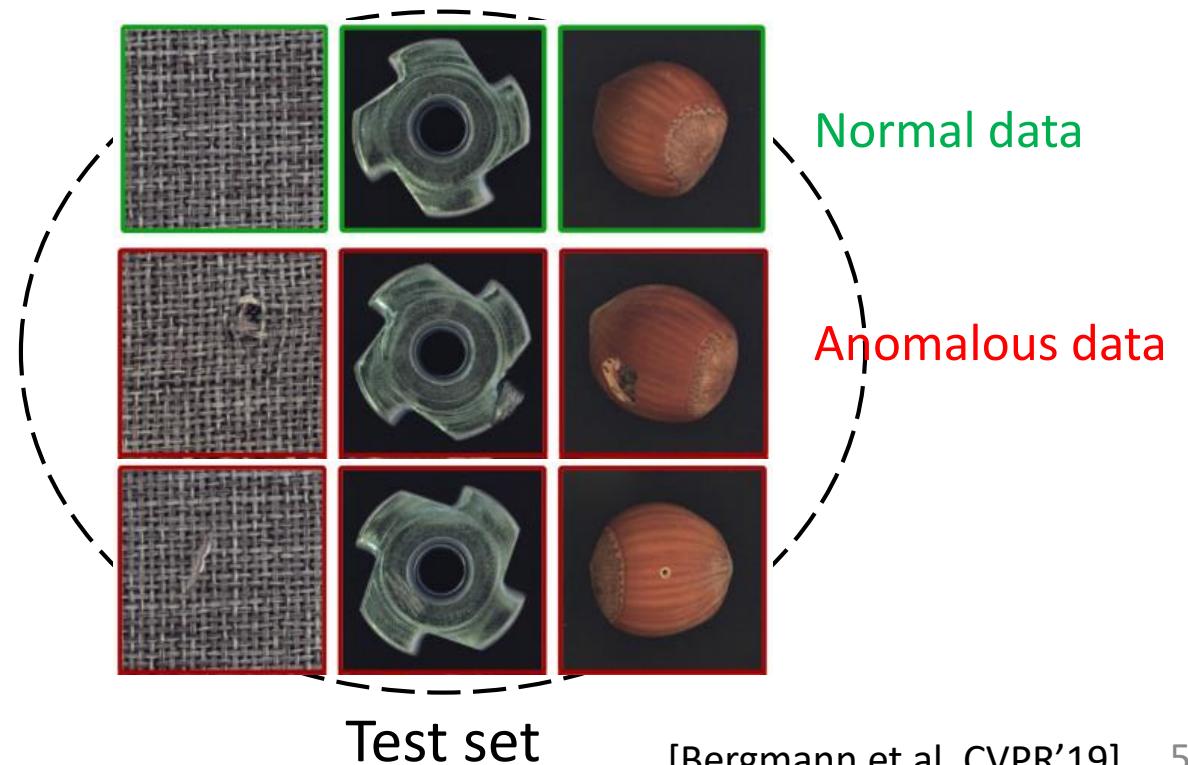
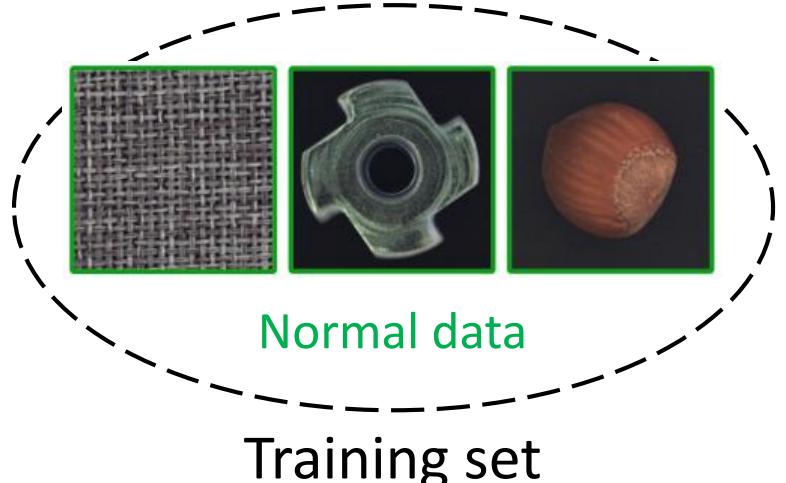
「無眼耳鼻舌身意，
無色聲香味觸法。」

— 《般若波羅蜜多心經》

Vision (e.g., images, videos)
Hearing (e.g., audio, speech)
Smell
Taste
Touch (e.g., tactile)
Mental (e.g., text, EEG)

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



Visual Anomaly Detection: Images and Videos

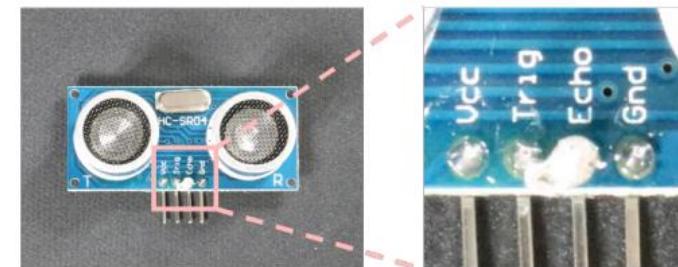
- Visual AD is a long-established problem in computer vision.
- Given its practical significance, AD has been widely deployed in various applications.

Video Anomaly Detection (VAD)

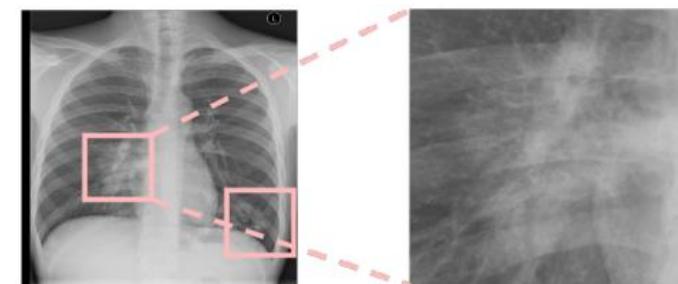


Security surveillance

Image Anomaly Detection (IAD)



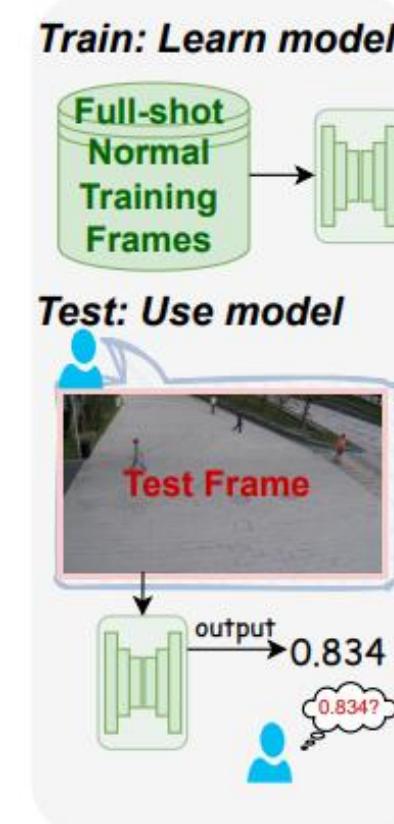
Industrial image inspection



Medical image diagnosis

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)

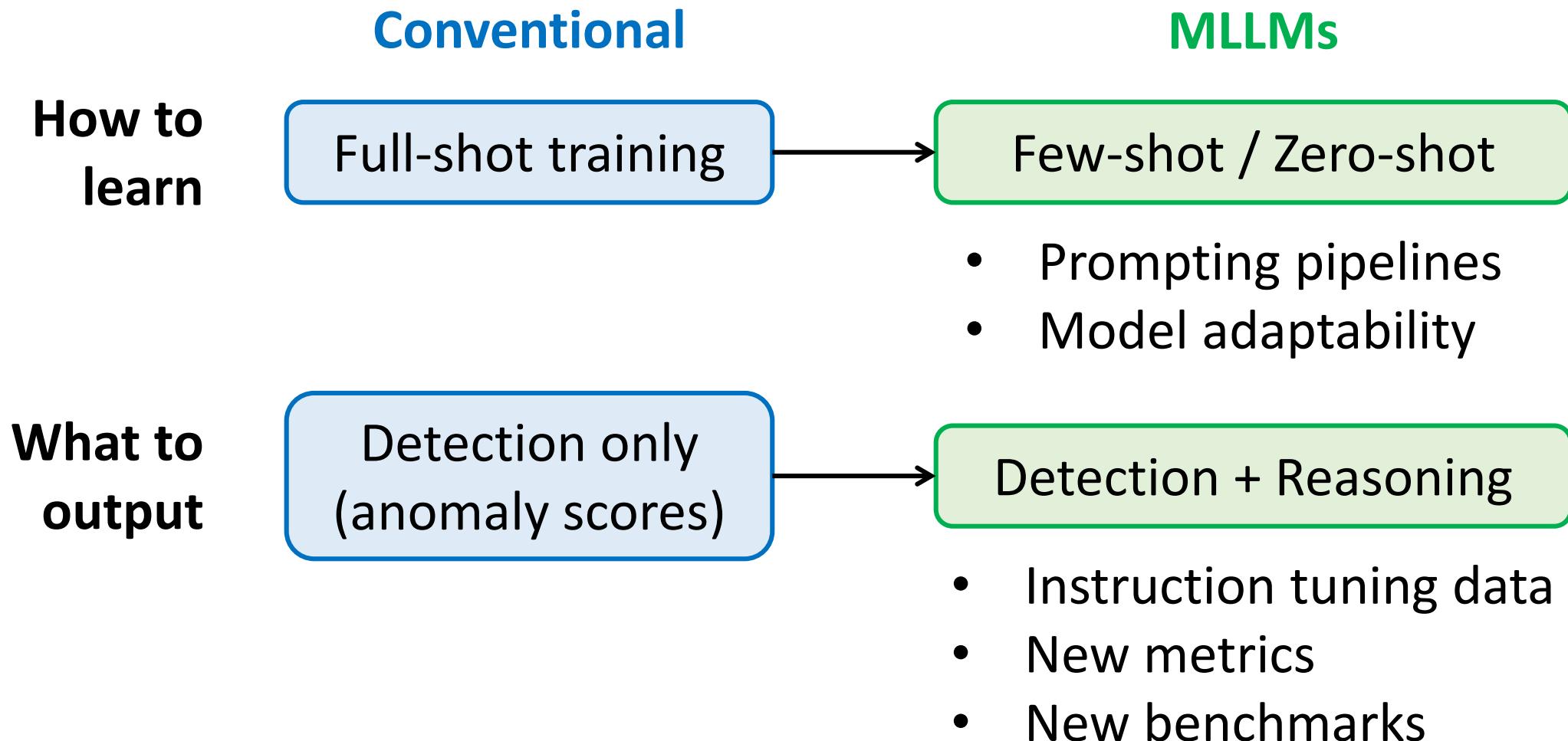


How can AD benefit from MLLMs?

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

Remain underexplored!

Anomaly Detection in the Era of MLLMs



Anomaly Detection in the Era of MLLMs

ECCV 2024

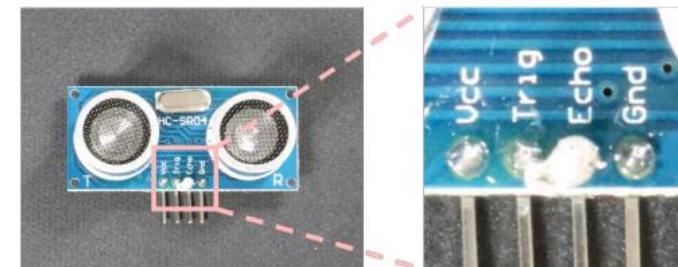
Video Anomaly Detection (VAD)



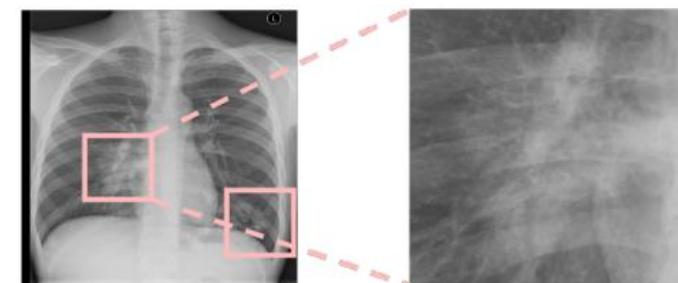
Security surveillance

CVPR 2025

Image Anomaly Detection (IAD)



Industrial image inspection



Medical image diagnosis

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

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

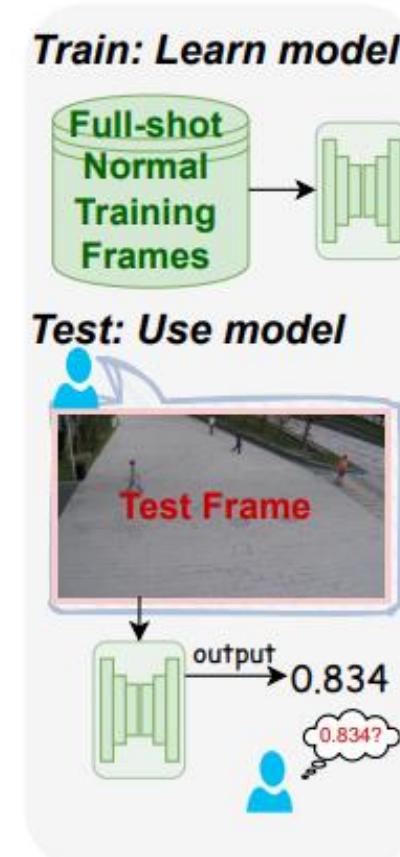
- **Assumption:** We only have a few **normal** data for our specific application, and it's costly to collect **anomaly** data.
- **Challenge:** The definition of "**anomaly**" depends on different context and downstream applications.
- **Goal:** Develop a VAD model for our specific application (specific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



[Source credit: Radu Tudor Ionescu. CVPR'23 Tutorial]

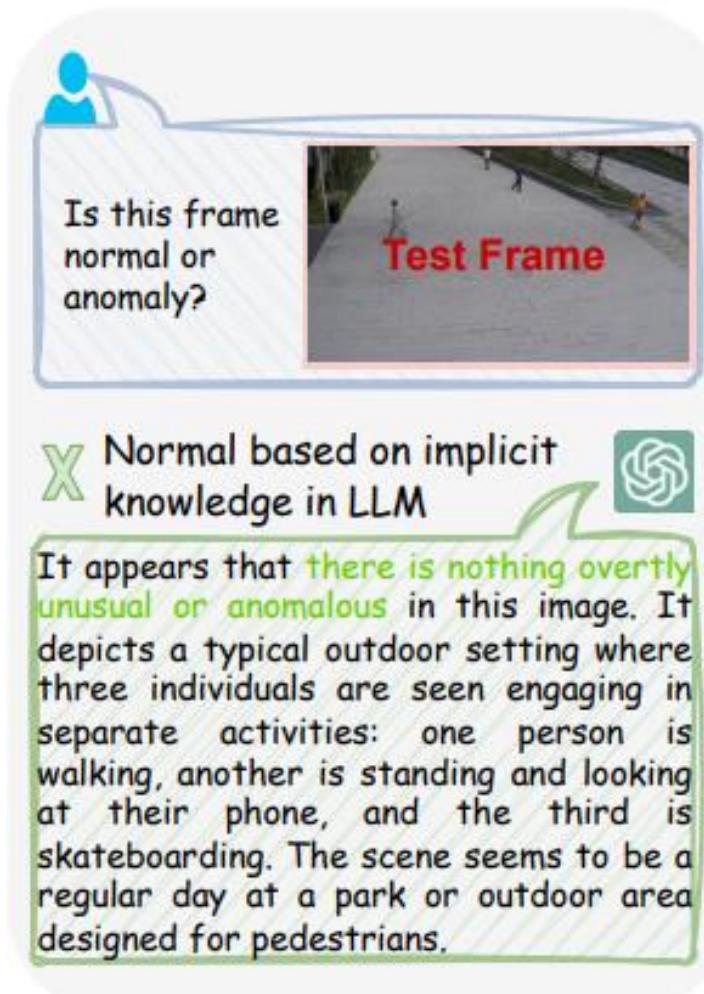
Conventional Learning-based Approaches

- **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)



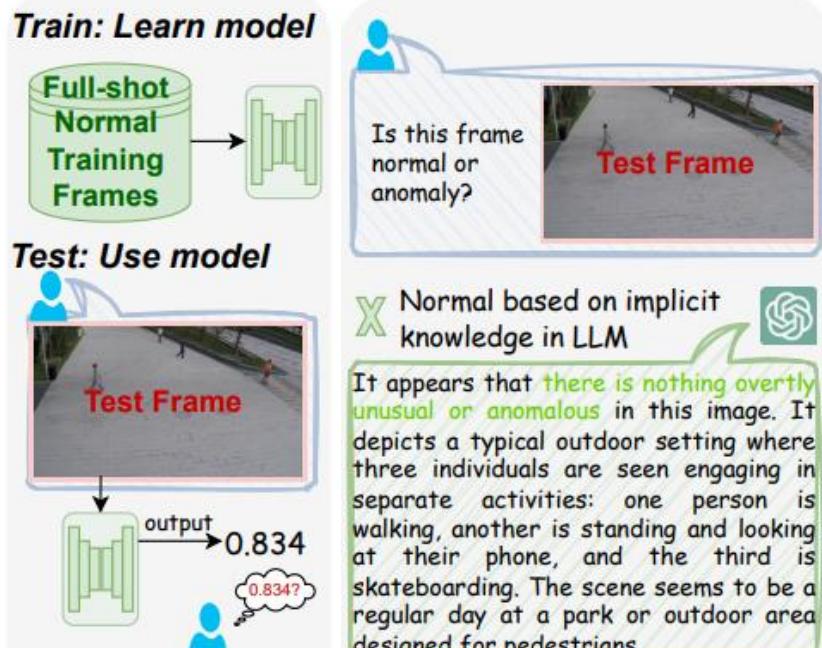
Query MLLMs Directly

- The implicit knowledge pre-trained in MLLMs may not align with specific VAD needs.
- Here GPT-4V mistakenly treats "skateboarding" as normal.

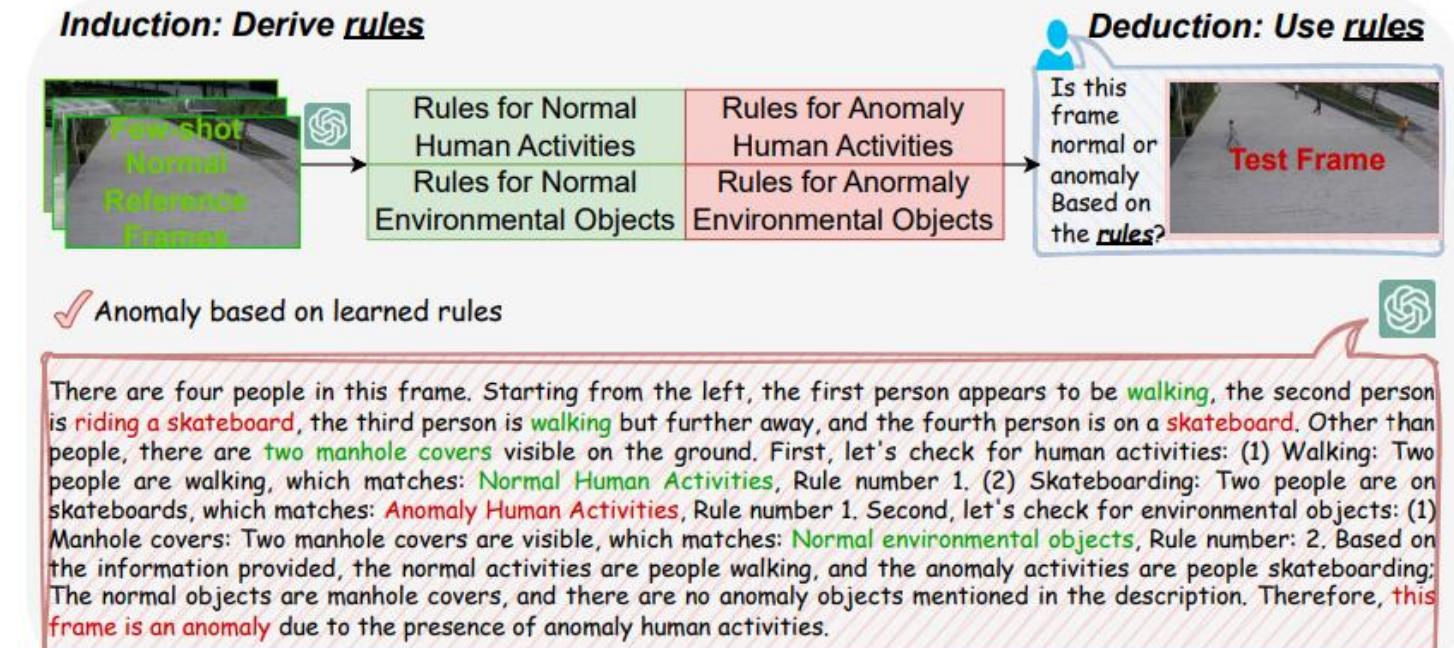


Method

- **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



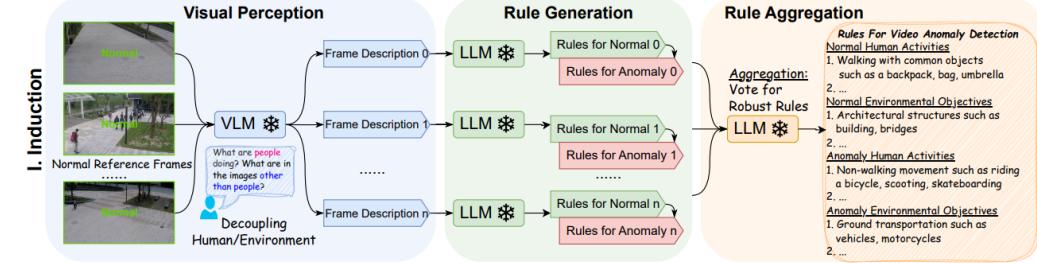
c) Ask LLM with the rules learned in induction stage (Ours)

Example (Induction)

A few normal reference frames



A set of rules



****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

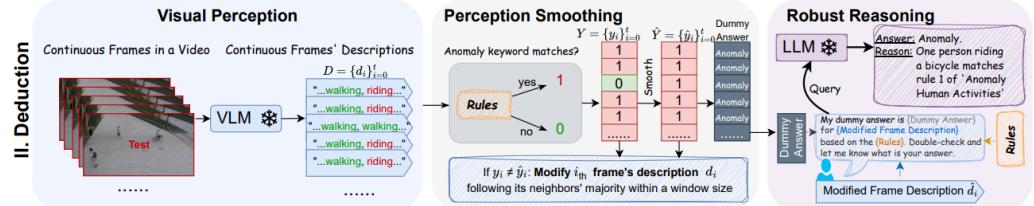
****Rules for Normal Environmental Objects:****

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)



Output
reasoning



\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.

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 Lo² Bardia Safaei¹ Vishal M. Patel¹ Isht Dwivedi²

¹Johns Hopkins University ²Honda Research Institute USA

{jxu155, bsafaei1, vpatel36}@jhu.edu {shao-yuan_lo, idwivedi}@honda-ri.com

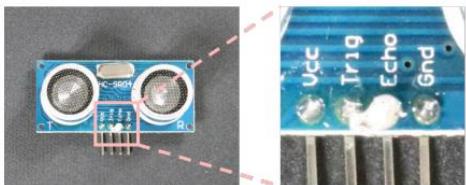
CVPR 2025 (Highlight)

- 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

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.

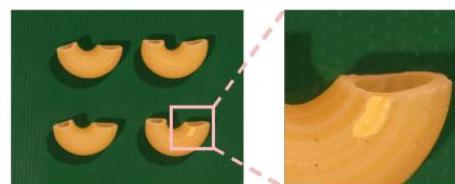
LLaVA-OV The image shows an ultrasonic sensor module, specifically the HC-SR04 model. There are no visible anomalies in the image; it appears to be a standard representation of this type of sensor module.

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.

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?

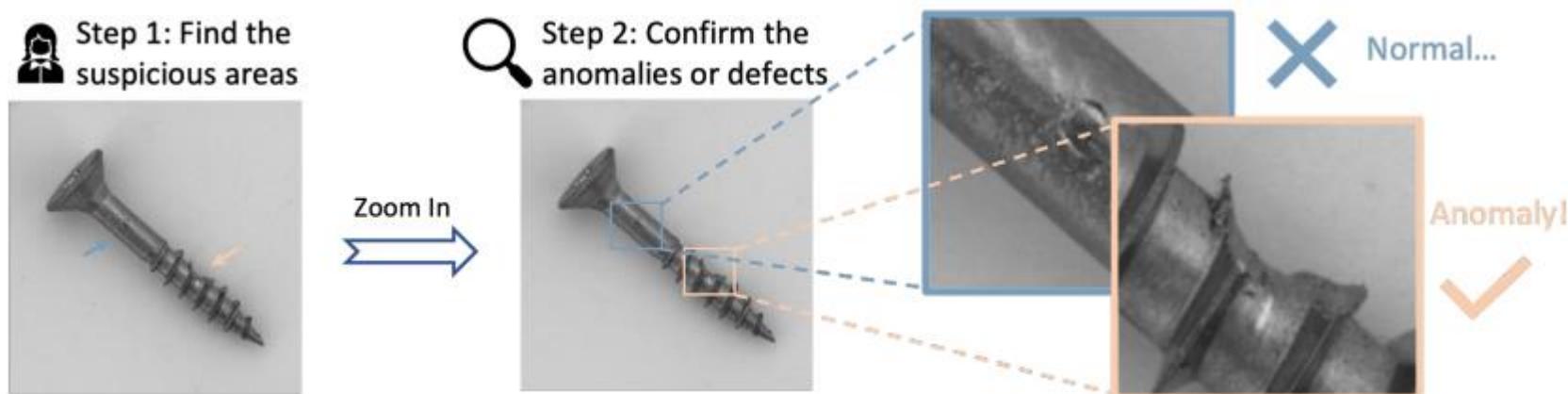
Anomaly-OV The graffiti affects the sign's functionality by potentially obstructing the visibility of important information, such as the word 'Seattle,' which could confuse drivers and lead to navigational errors.

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.

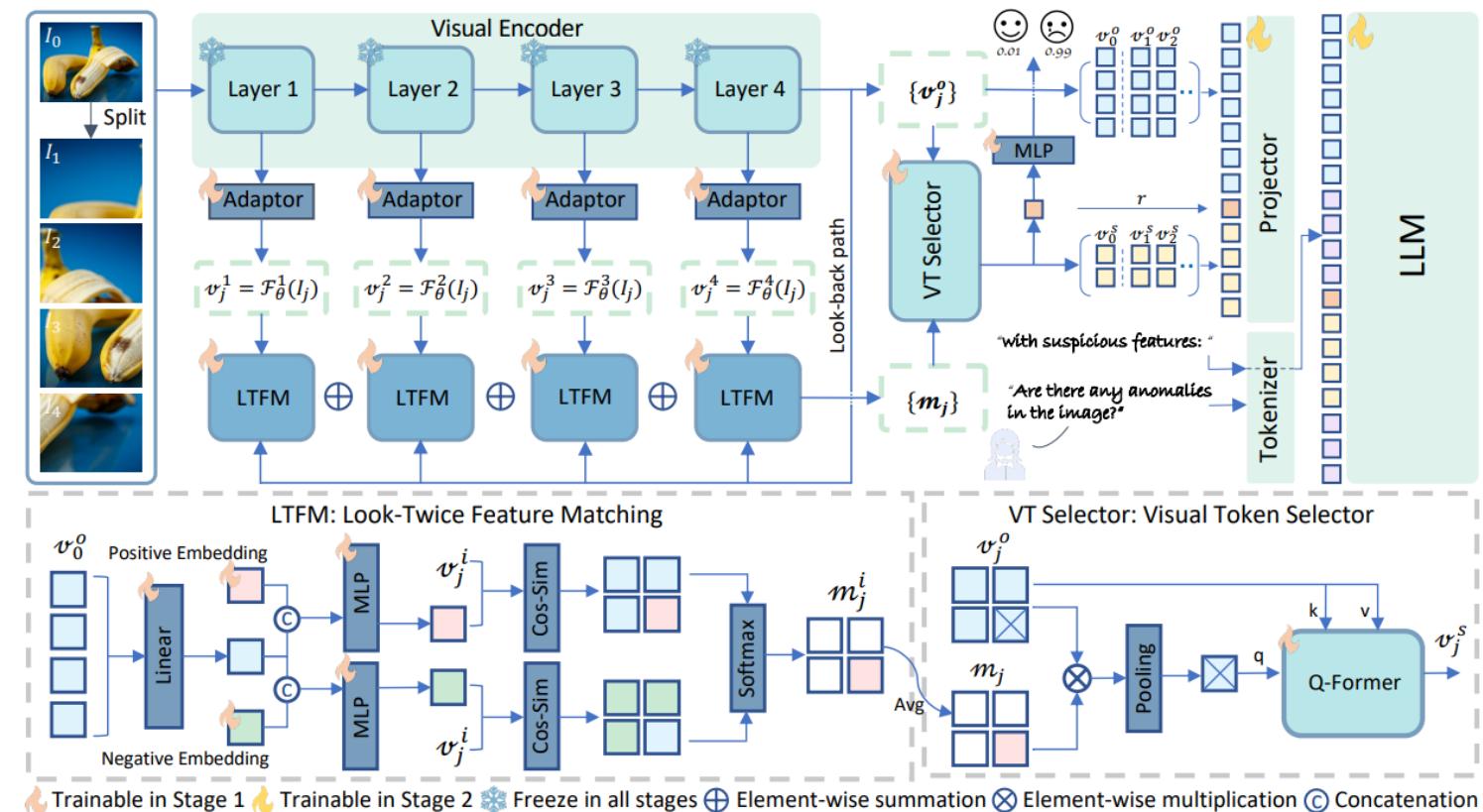
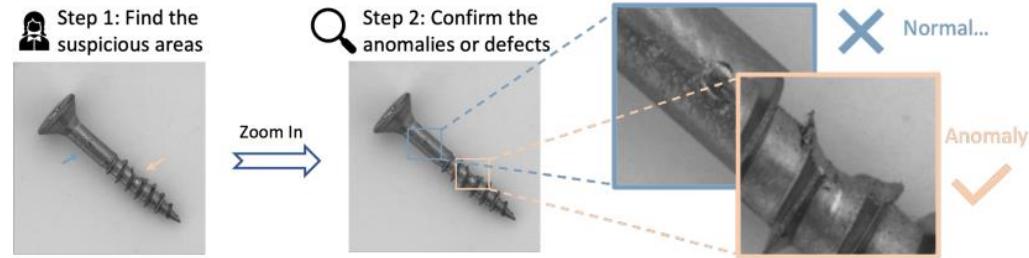
Method

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



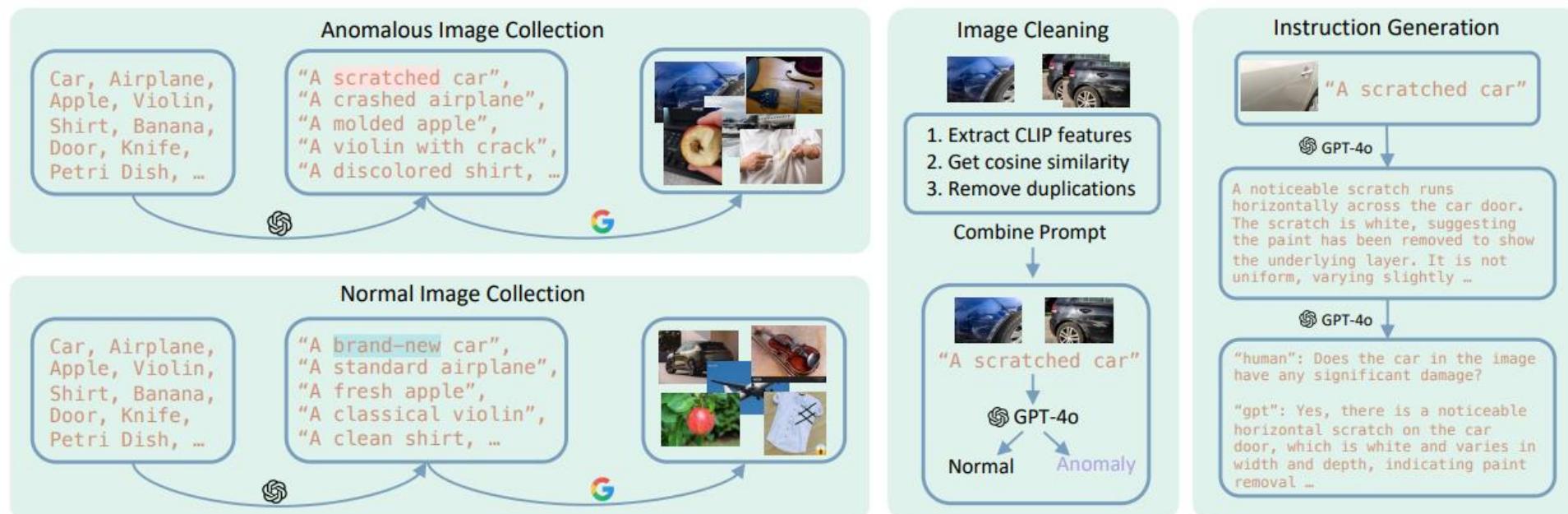
Method

- **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



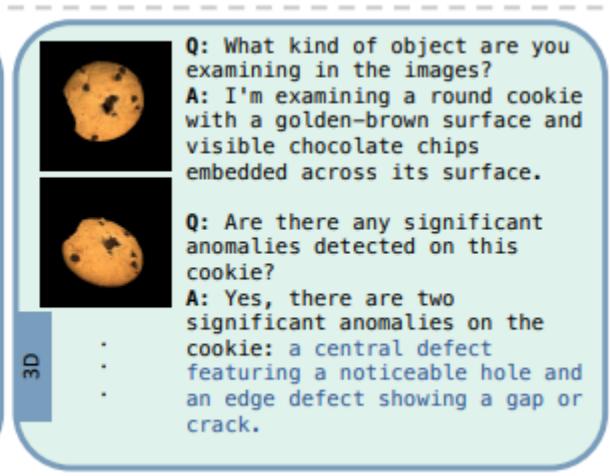
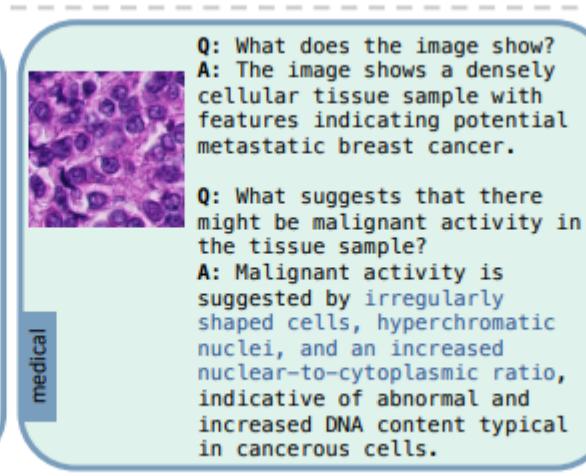
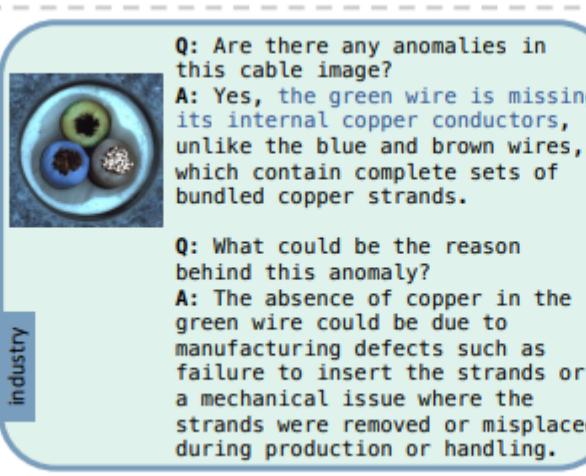
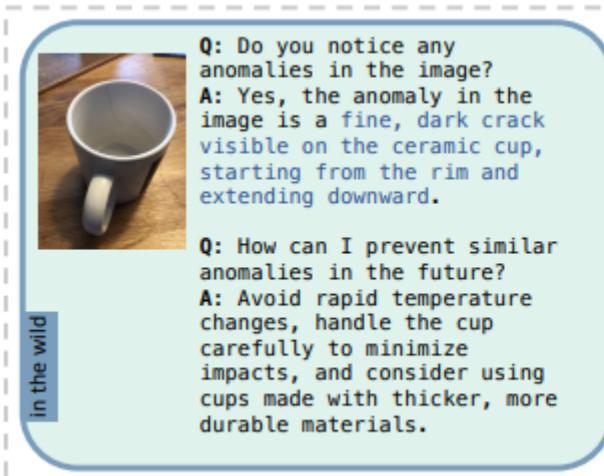
Instruction Tuning Dataset

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



Instruction Tuning Dataset

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



WebAD-72k
[Collected by us via
Google Image Search]

MVTec AD
[Bergmann et al. CVPR'19]

BMAD
[Bao et al. CVPR'24]

MVTec-3D AD
Real3D-AD
Anomaly-ShapeNet

Results

- 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>	4.26	0.84	<u>6.34</u>

Summary

ECCV 2024

Video Anomaly Detection (VAD)

CVPR 2025

Image Anomaly Detection (IAD)

How to learn

What to output

Conventional

Full-shot training

MLLMs

Few-shot / Zero-shot

- Prompting pipelines
- Model adaptability

Detection only
(anomaly scores)

Detection + Reasoning

- Instruction tuning data
- New metrics
- New benchmarks

MLMs for Video Affective Reasoning

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.

IJCV 2025

- 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:



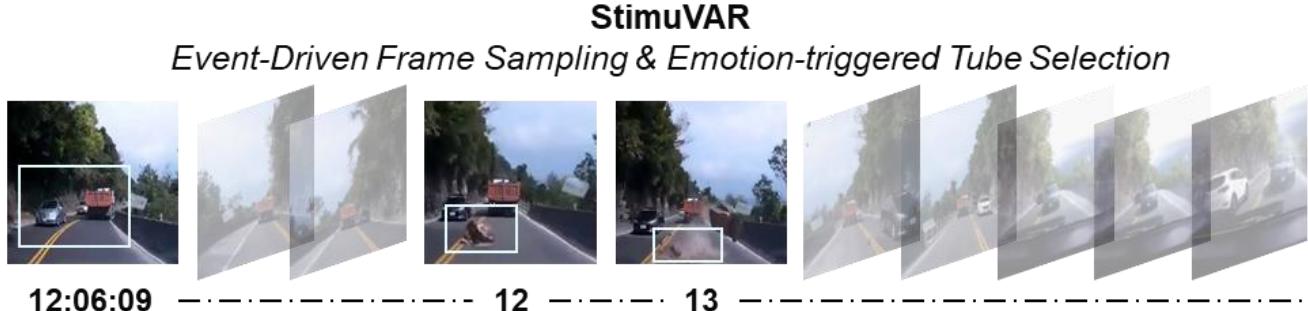
Method

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



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

Answer: *Boredom* 😴 😴 😴

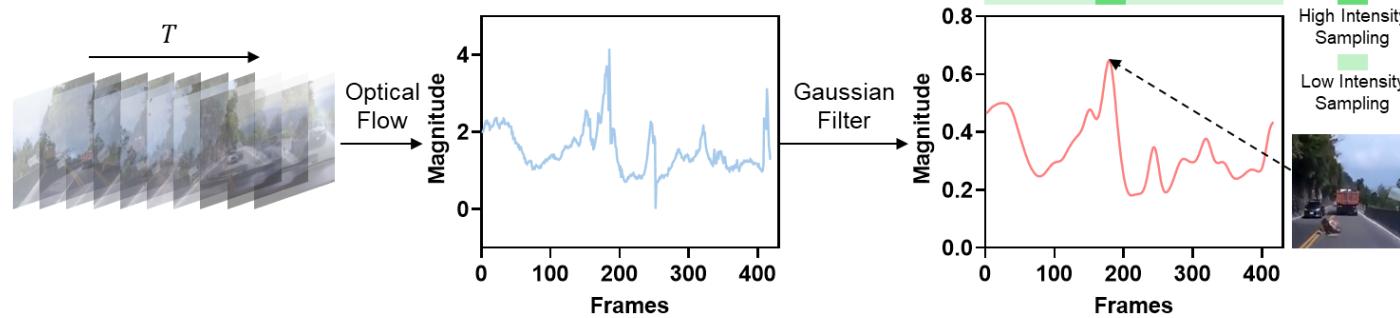


User: What is the emotion of the viewers when they watch the video?

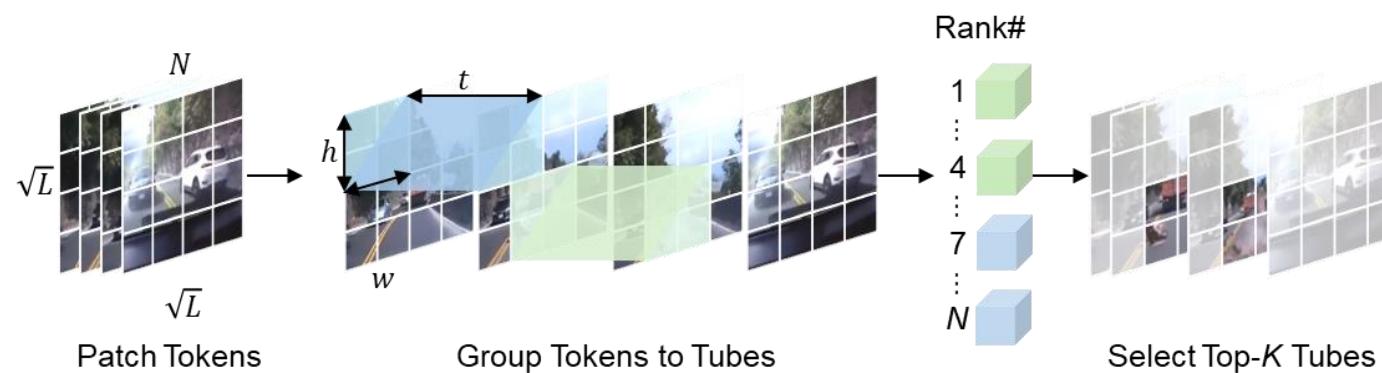
Answer: The viewer feels *Surprise* 😱 because initially, it is a typical road scene with a red truck and a silver car sets the expectation for a mundane video. However, the sudden appearance of a rock, falling onto the road introduces an unexpected element. This unexpected twist in the video content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

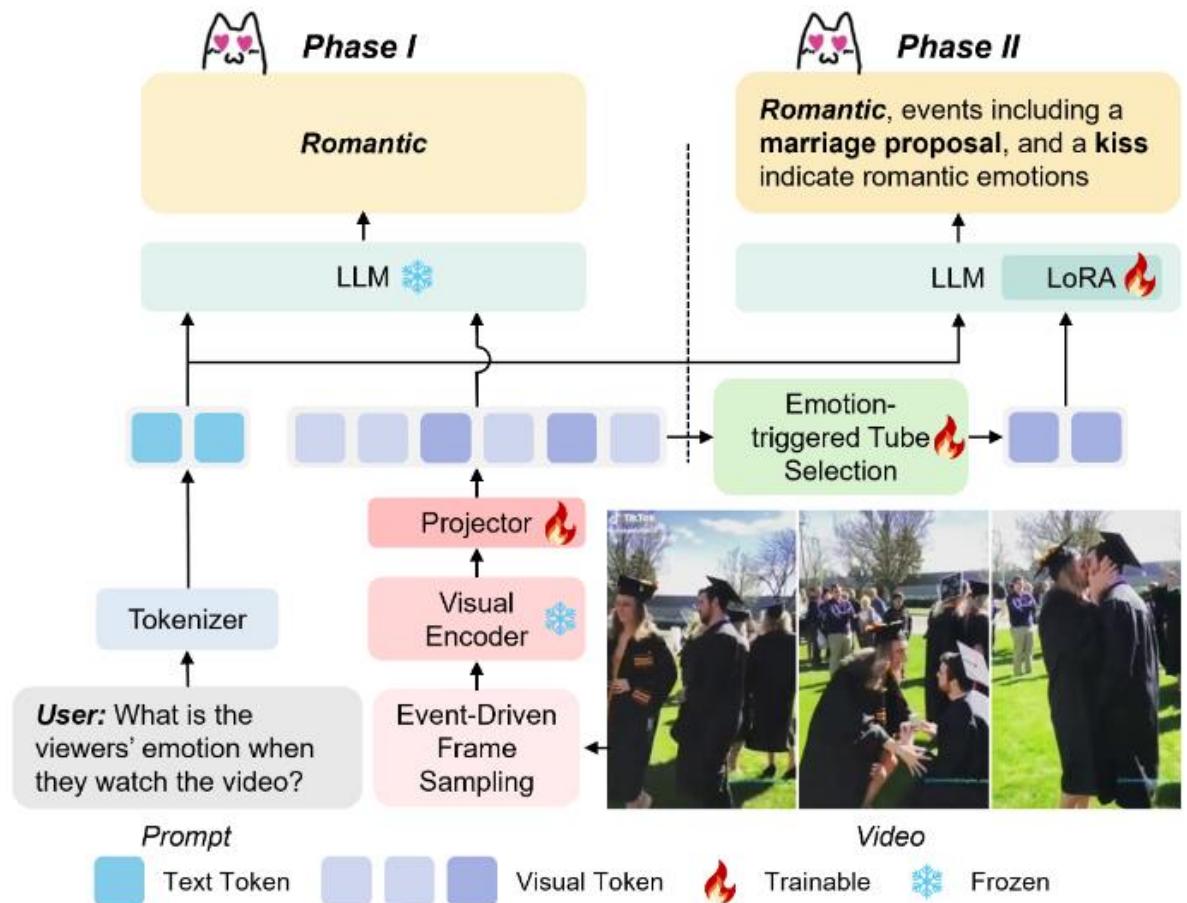


Affective Training

- 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

MLMs for Theory of Mind

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

Chunhui Zhang¹ Zhongyu Ouyang¹ Kwonjoon Lee² Nakul Agarwal²
Sean Dae Houlihan¹ Soroush Vosoughi^{1†} Shao-Yuan Lo^{2†}

ICML 2025 (Spotlight)

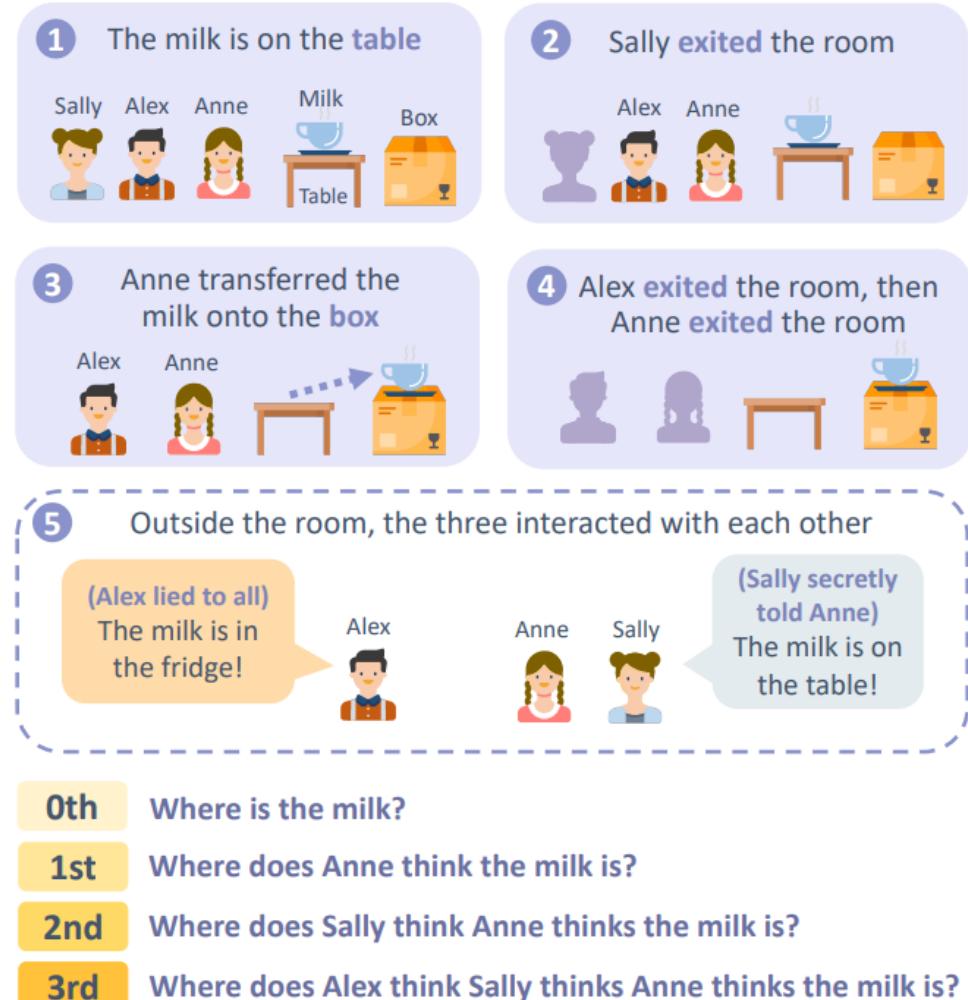
- An **analysis-style** paper for Multimodal Theory-of-Mind (MMToM), a new topic.
- **Scaling** MMToM on larger language models (LMs) **without increasing training costs**.

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 Jin¹, Yutong Wu², Jing Cao³, Jiannan Xiang⁴,

Yen-Ling Kuo⁵, Zhiting Hu⁴, Tomer Ullman², Antonio Torralba³, Joshua Tenenbaum³, Tianmin Shu⁶

¹NYU, ²Harvard, ³MIT, ⁴UCSD, ⁵UVA, ⁶JHU

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. ✓

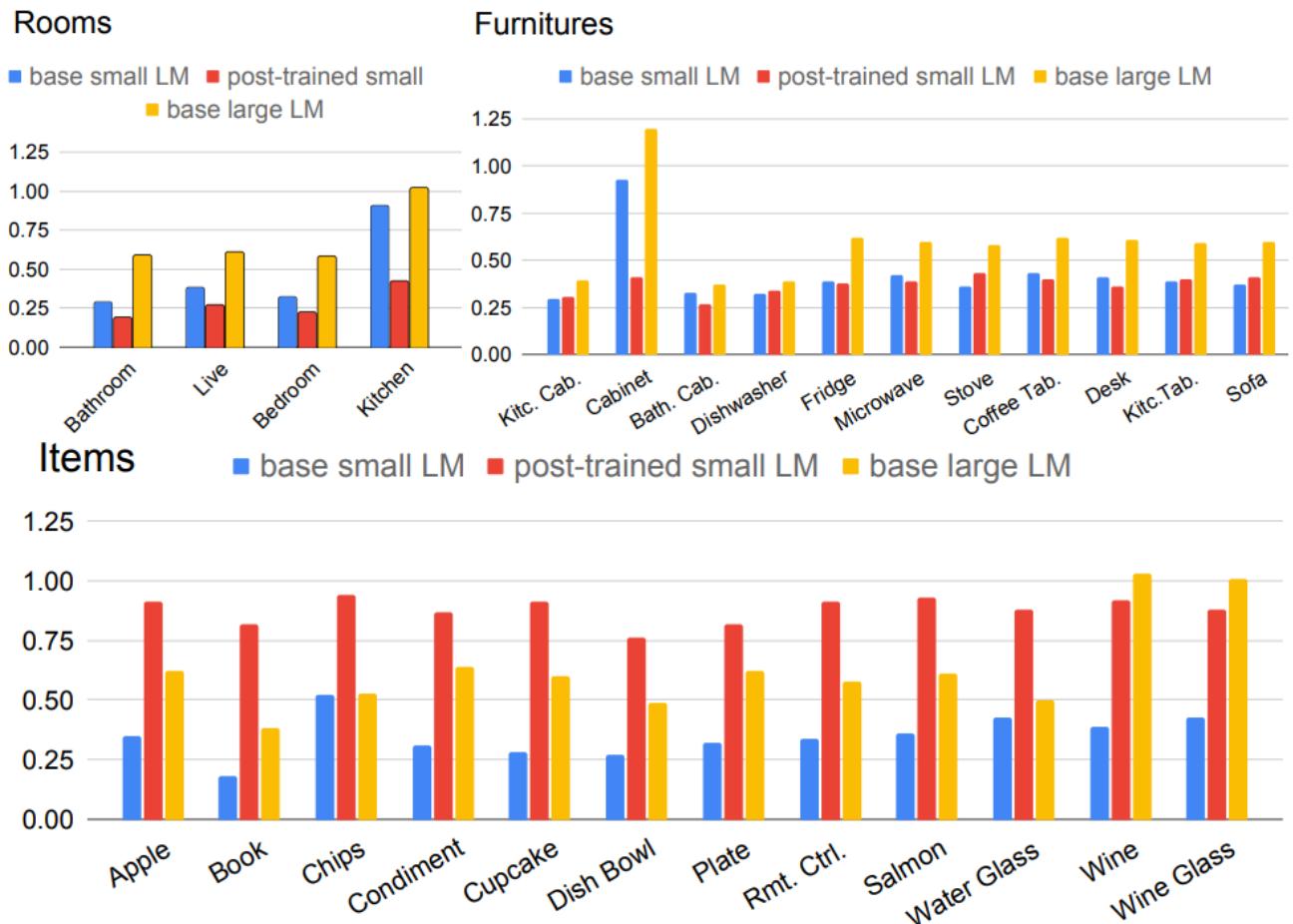
(b) Emily has been trying to get a wine glass. ✗

- 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

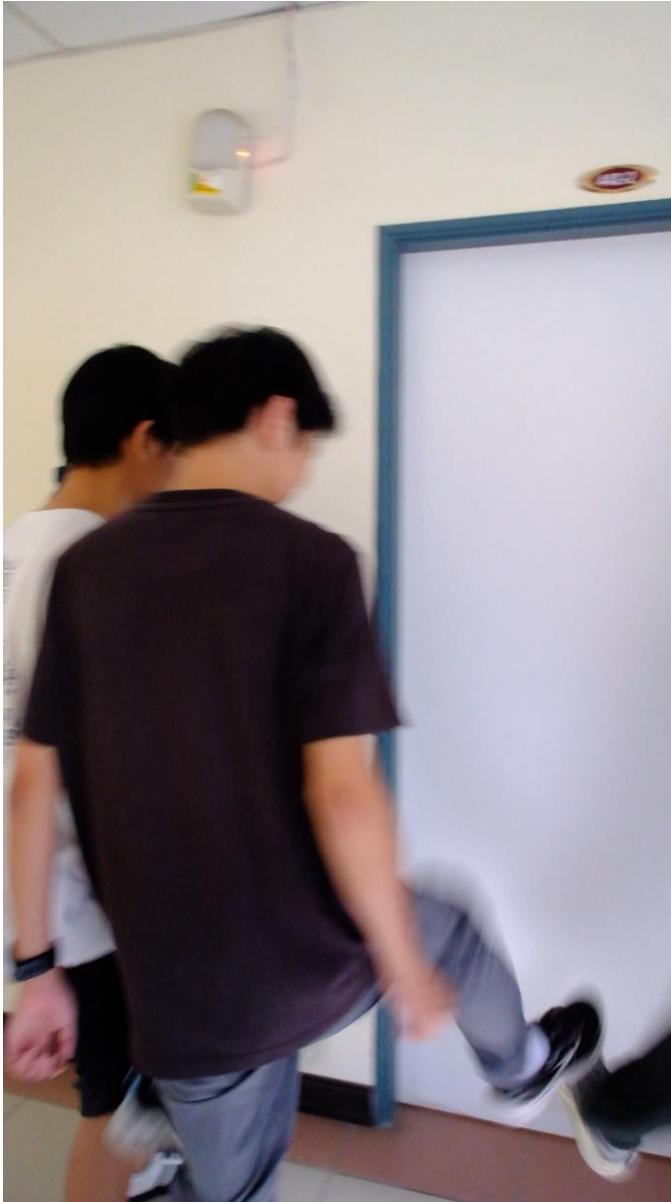
- **Post-trained Small LM** is better aligned with requirements for specific ToM scenarios.
 - **Base Large LM** has better general world knowledge and reasoning.
-
- Transfer the post-trained alignment from Small LM to Large LM.
 - Adapt Large LM's ToM behaviors by training Small LM only.

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

Results

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

LM	config	belief inference				goal inference				all	
		1.1	1.2	1.3	avg.	2.1	2.2	2.3	2.4		
Llama-3.1	8B-zero-shot	88.00	72.00	91.00	83.67	65.33	62.67	22.67	54.67	51.33	
	8B-post-trained	90.00	71.00	93.00	84.67	69.33	72.00	62.67	72.00	69.00	
	70B-zero-shot	85.00	63.00	93.00	80.33	72.00	76.00	16.00	61.33	56.33	
	70B-post-trained	91.00	69.00	95.00	85.00	69.33	80.00	29.33	69.33	62.00	
	405B-zero-shot	86.00	70.00	90.00	82.00	73.33	78.67	21.33	66.67	60.00	
	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



Join My Lab!

信韌智慧實驗室

Trustworthy AI Lab

Trai Lab

蹄實驗室

Join NTU Trai Lab!

- Team at a glance
 - 2 Master's students
 - 7+3 undergrad students
 - Preparing 1 CVPR submission and 1 ICML submission
- How we work together
 - **Weekly lab meeting:** Teams take turns sharing progress + paper reading seminar
 - **Weekly 1-on-1 or subgroup meeting:** In-depth discussion of ongoing research progress
 - **Zulip workspace:** Free Slack

Join NTU Trai Lab!



- 電動升降桌 + 人體工學椅 + 27吋螢幕

Join NTU Trai Lab!

- Buying GPUs... and will keep buying...

shaoyuanlo 2025-9-25 ...

All in GPUs !
希望能變成好的研究

機關名稱：國立臺灣大學
單位名稱：資工系
聯絡人：羅紹元 教授
統一編號：
TEL :
E-MAIL :
地 址 :
製表日期：2025年9月24日

CERTIFIED

PROD.#	產品名稱及規格	數量	優惠價	小計
1	品名：高效能運算伺服器	1	\$2,098,000	\$2,098,000

Join NTU Trai Lab!

- Skills
 - Solid background in Machine Learning and Deep Learning
 - Proficiency in Linux, Python, and PyTorch
- Related courses
 - Foundations of Artificial Intelligence
 - Machine Learning
 - Multimodal Artificial Intelligence
 - Computer Vision Practice with Deep Learning
 - Security and Privacy of Machine Learning

Join NTU Trai Lab!

- **Phase 1 (2 months): Study deep learning fundamentals**
 - Stanford CS231N course
 - Two lectures per week
- **Phase 2 (1 month): Study advanced LLM technology**
 - Tsinghua LLM open course
 - Two lectures per week
- **Phase 3 (2 month): Learn to read papers**
 - One representative paper per week
- **Phase 4 (6 months): Do research!**

Join NTU Trai Lab!

- Goal: Submit a solid paper to a top AI conference
- CV: CVPR, ICCV, ECCV
- ML: NeurIPS, ICLR, ICML
- 站上世界的舞台！