Toward Reliable Generalization in Real-World AI Systems

Hossein Kashiani
Clemson University
Email: hkashia@clemson.edu
Homepage: https://kashiani.github.io/

Why reliable generalization matters?

- AI systems often perform well in training settings but fail when deployed in changing real-world conditions.
- Deployment: environment, weather, population, sensing, noise, competing objectives.

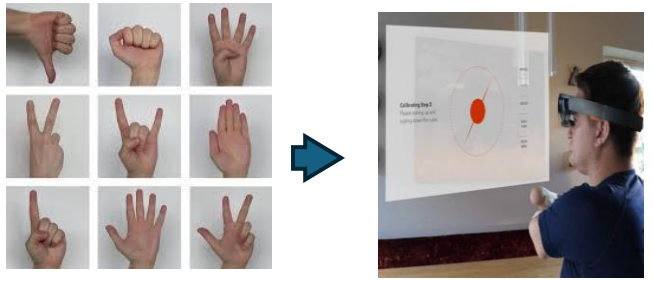
Why reliable generalization matters?

- AI systems often perform well in training settings but fail when deployed in changing real-world conditions.
- Deployment: environment, weather, population, sensing, noise, competing objectives.
- Distribution shifts are everywhere.

Train Deploy

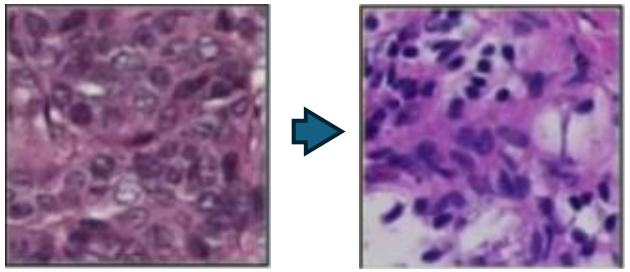
Autonomous driving (adverse weather)

Train Deploy



Gesture recognition (new settings)

Train Deploy

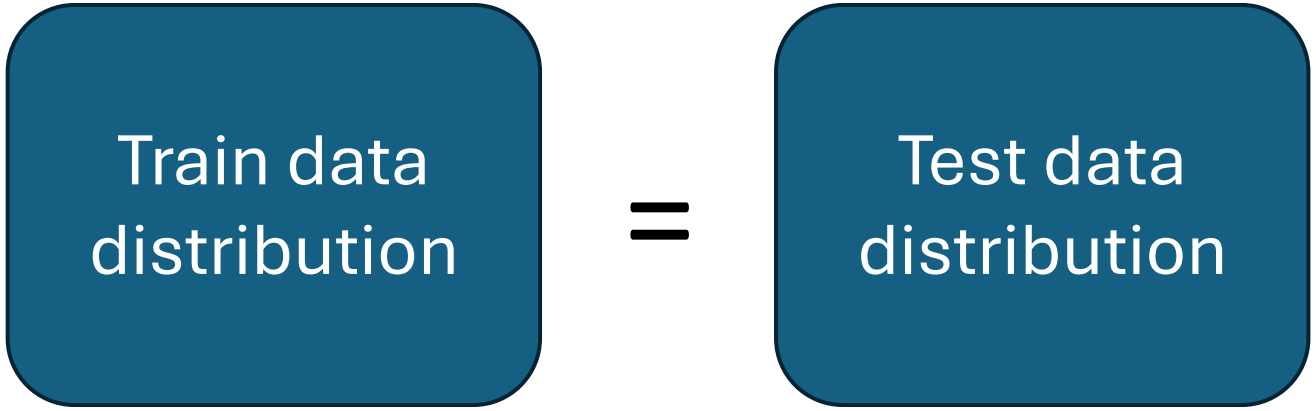


Tumor detection (new hospitals)

Train Deploy

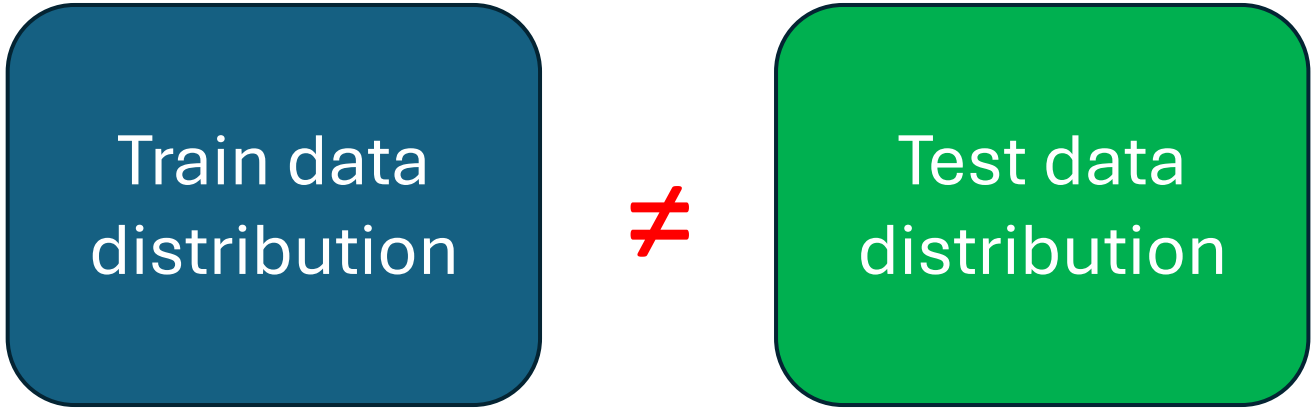
Security (unseen attacks)

Standard assumption in machine learning



ML model perform well

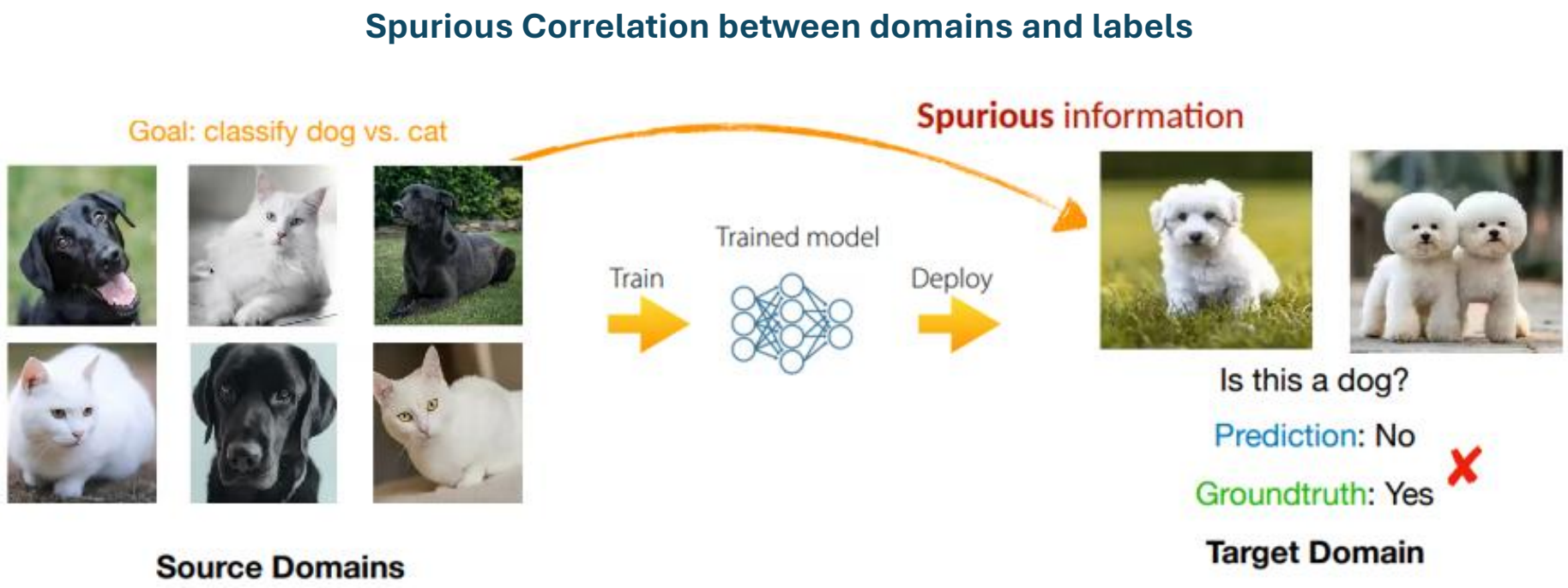
Standard assumption in machine learning



ML model performance degrade

What prevents reliable generalization in real-world systems?

- **Spurious correlations:**
  shortcut cues, unstable artifacts, dataset bias

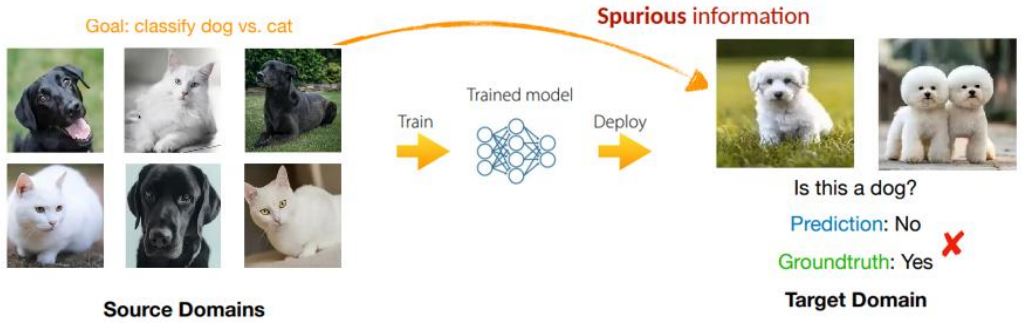


What prevents reliable generalization in real-world systems?

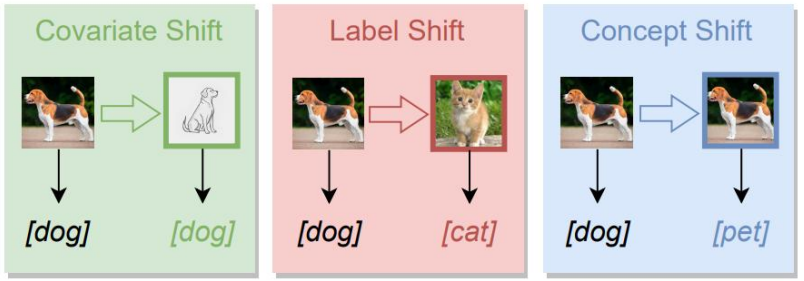
- **Spurious correlations:**
  shortcut cues, unstable artifacts, dataset bias

- **Distribution shifts:**
  style, lighting, corruption, domain changes

Spurious Correlation between domains and labels



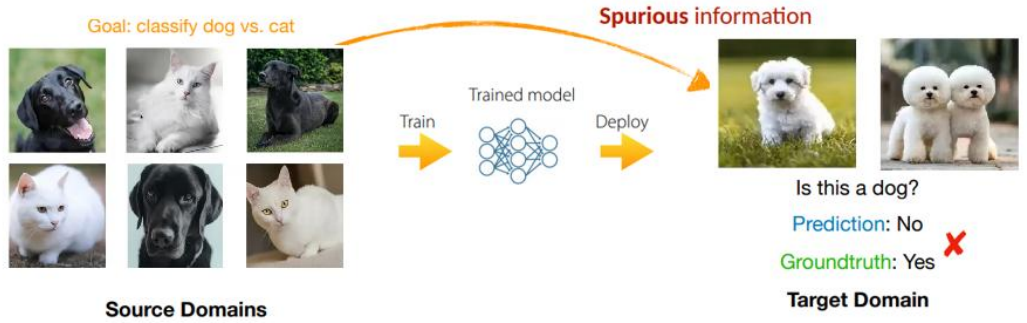
Various Distribution Shifts



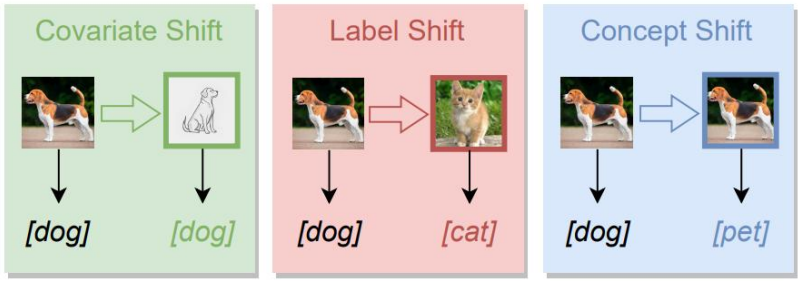
What prevents reliable generalization in real-world systems?

- **Spurious correlations:**
  shortcut cues, unstable artifacts, dataset bias
- **Distribution shifts:**
  style, lighting, corruption, domain changes
- **Task interference:**
  multiple objectives, conflicting supervision

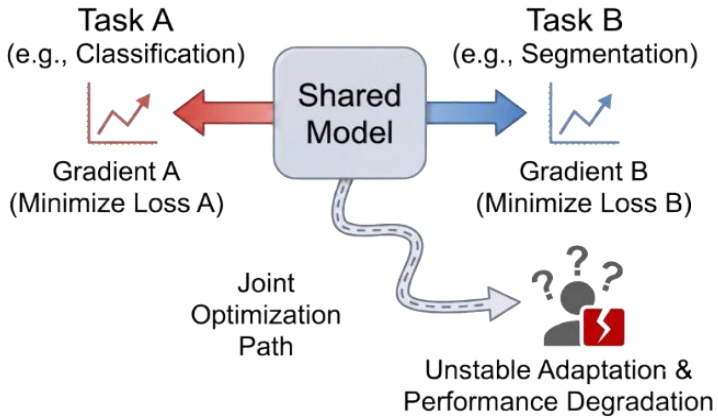
Spurious Correlation between domains and labels



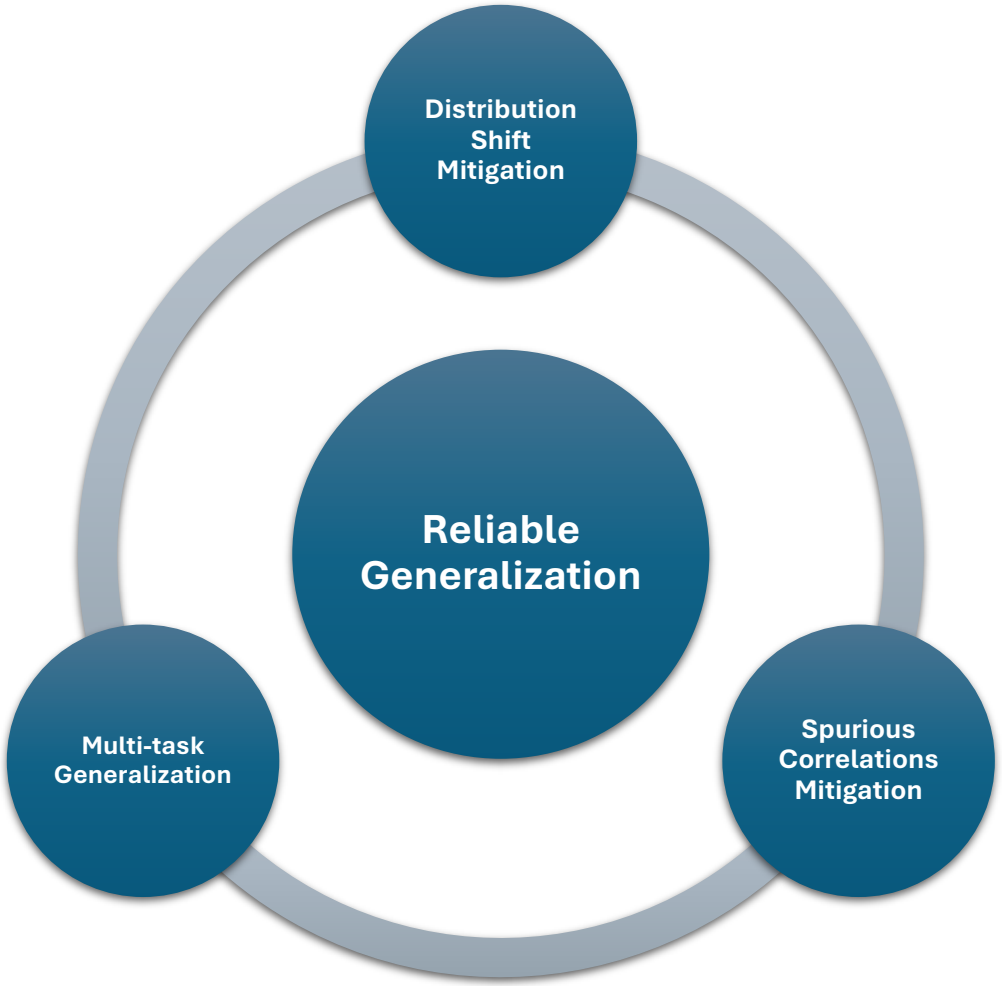
Various Distribution Shifts

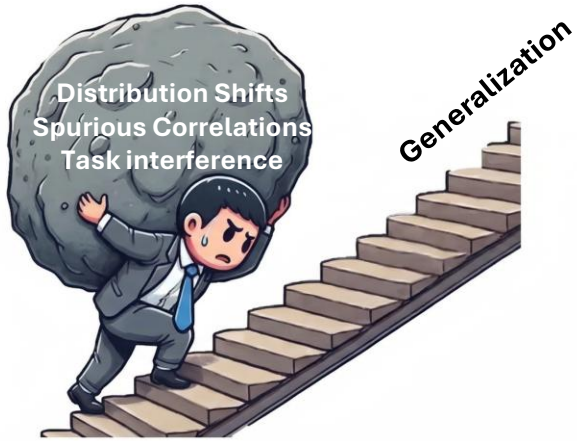


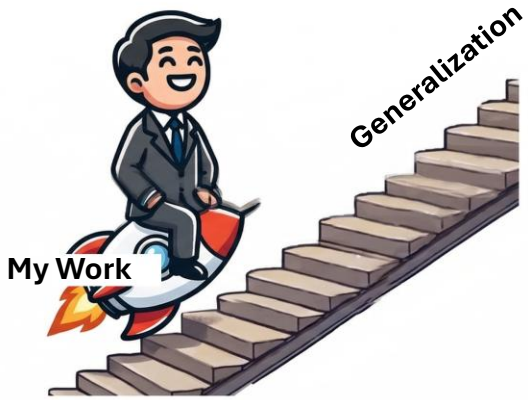
Heterogeneous Multi-Task Learning



Generalizable representation learning

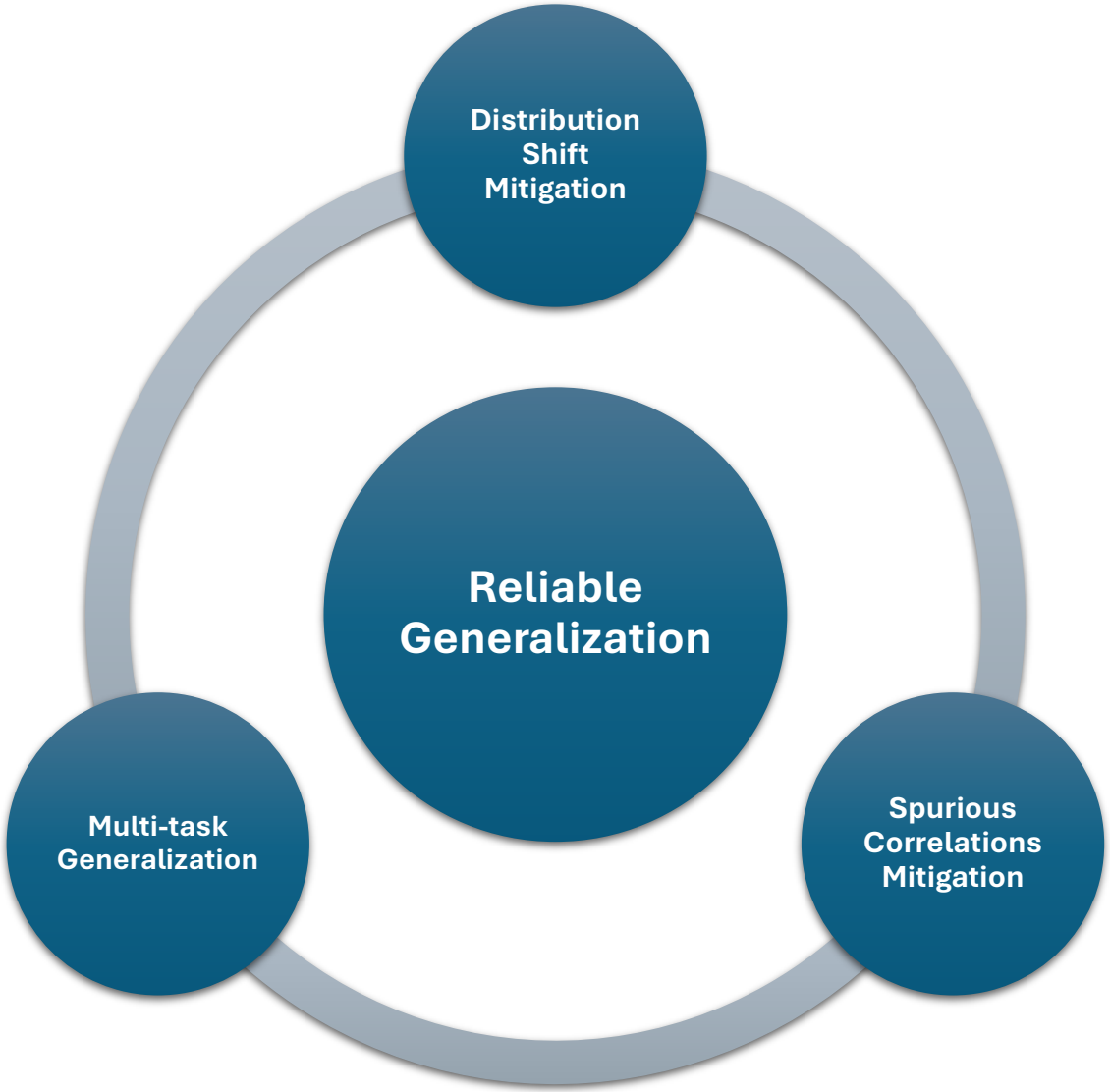







Spurious Correlations

Generalizable representation learning



Digital Forensics

Autonomous Driving

Deepfake detectors often rely on spurious correlations

Human-perceptible correlations: Identity, background, and structural artifacts

Deepfake detectors often rely on spurious correlations

What about Imperceptible correlations?

Human-perceptible correlations: Identity, background, and structural artifacts

Deepfake detectors often rely on spurious correlations

What about Imperceptible correlations?

Human-perceptible correlations: Identity, background, and structural artifacts

We identify one underexplored form of such bias: **spectral bias**

What is Spectral Bias?

Spurious Correlations Mitigation

- Detectors over-rely on **dominant frequency components**, which are specific to forgery types.

Kashiani, et al. "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing", *CVPR 2025*.

What is Spectral Bias?

Spurious Correlations Mitigation

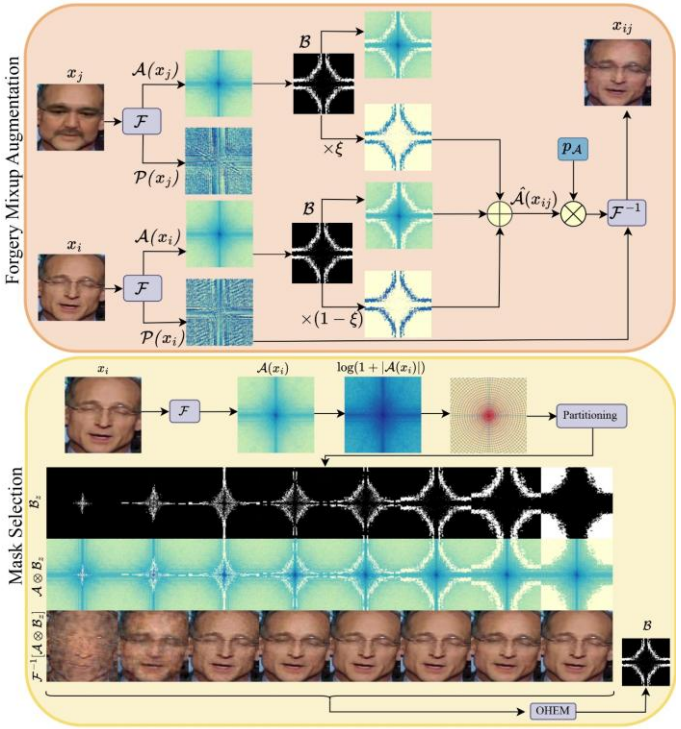
- Detectors over-rely on **dominant frequency components**, which are specific to forgery types.

Spectral Bias limits generalization to **unseen forgeries**.

Kashiani, et al. "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing", *CVPR 2025*.

FreqDebias: Mitigating Spectral Bias

Spurious Correlations Mitigation



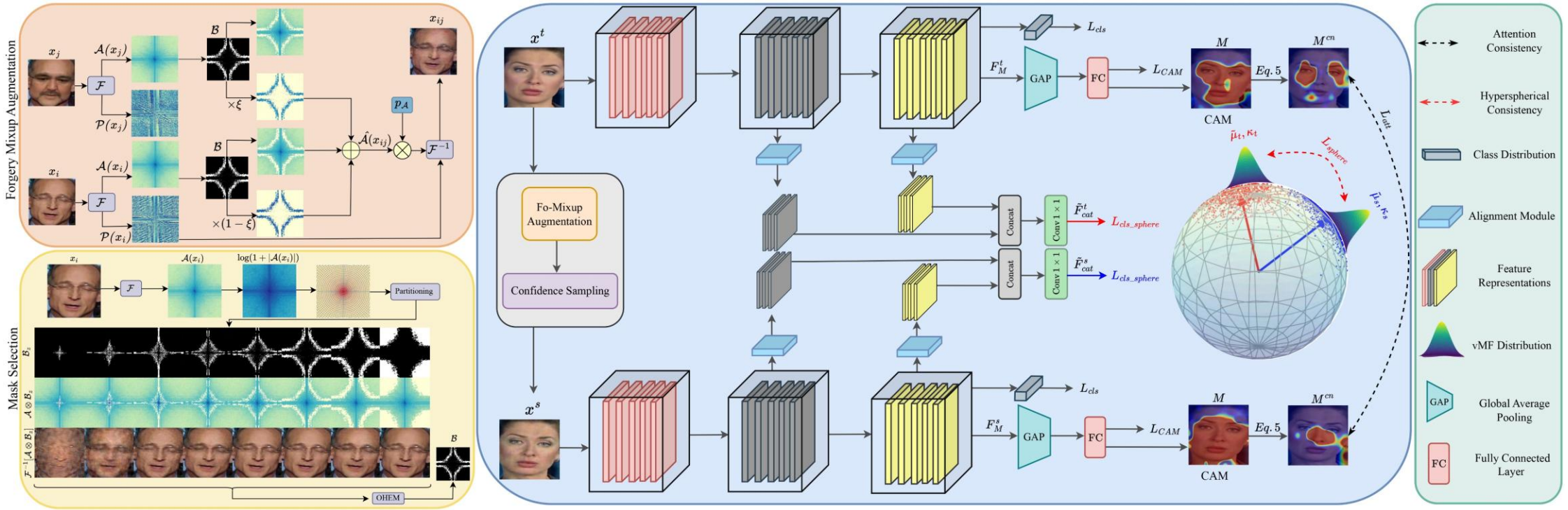
Forgery Mixup augmentation:

➢ Enhances the detector's exposure to a diversified frequency spectrum.

Kashiani, et al. "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing", *CVPR 2025*.

FreqDebias: Mitigating Spectral Bias

Spurious Correlations Mitigation



Forgery Mixup augmentation:

➢ Enhances the detector's exposure to a diversified frequency spectrum.

Dual Consistency Regularization (CR):

➢ **Local Consistency**: Enforce alignment via Class Activation Maps (CAMs).

$$L_{att} = D_{JS}\left(\sigma\left(M^{cn}(x^s);\tau\right), \sigma\left(M^{cn}(x^t);\tau\right)\right)$$

➢ **Global Consistency:** Enforce domain alignment using the Distribution Matching Score (DMS).

$$\mathrm{DMS} = 1\Big/\left(1 + D_{\mathrm{KL}}\left(p(\tilde{\boldsymbol{F}}^s_{cat} \mid \kappa_s, \tilde{\boldsymbol{\mu}}_s),\ p(\tilde{\boldsymbol{F}}^t_{cat} \mid \kappa_t, \tilde{\boldsymbol{\mu}}_t)\right)\right)$$

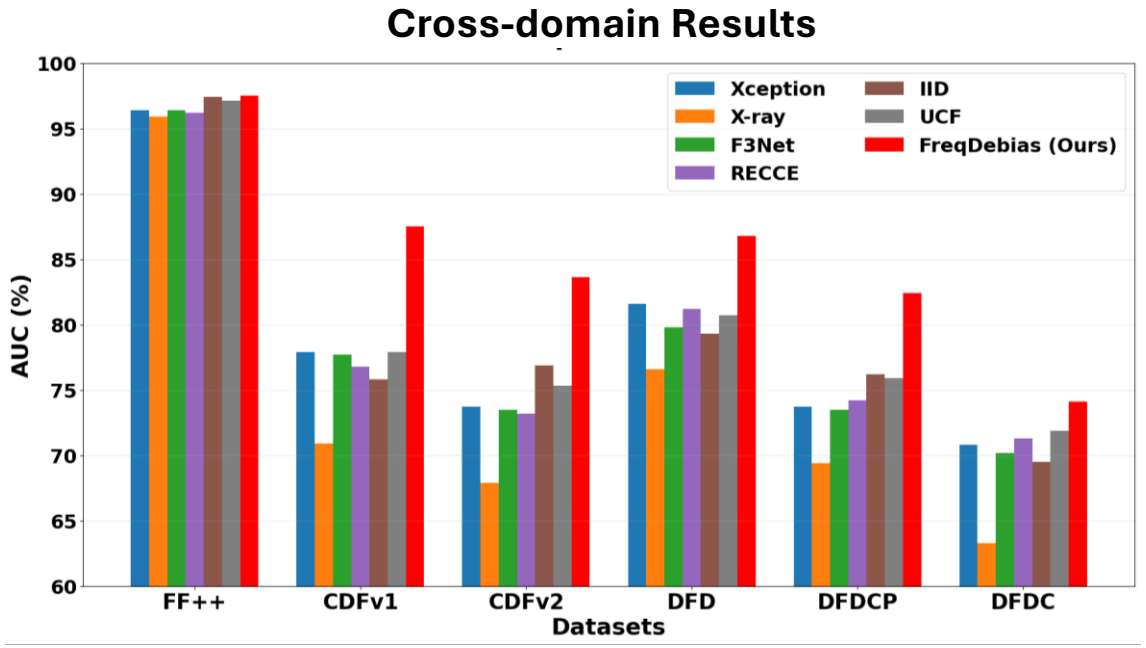
Kashiani, et al. "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing", *CVPR 2025*.

Results

Spurious Correlations Mitigation

Cross-Domain Evaluation:

- **Trained on** FF++ (HQ).
- **Evaluated on** unseen forgeries.



Kashiani, et al. "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing", *CVPR 2025*.

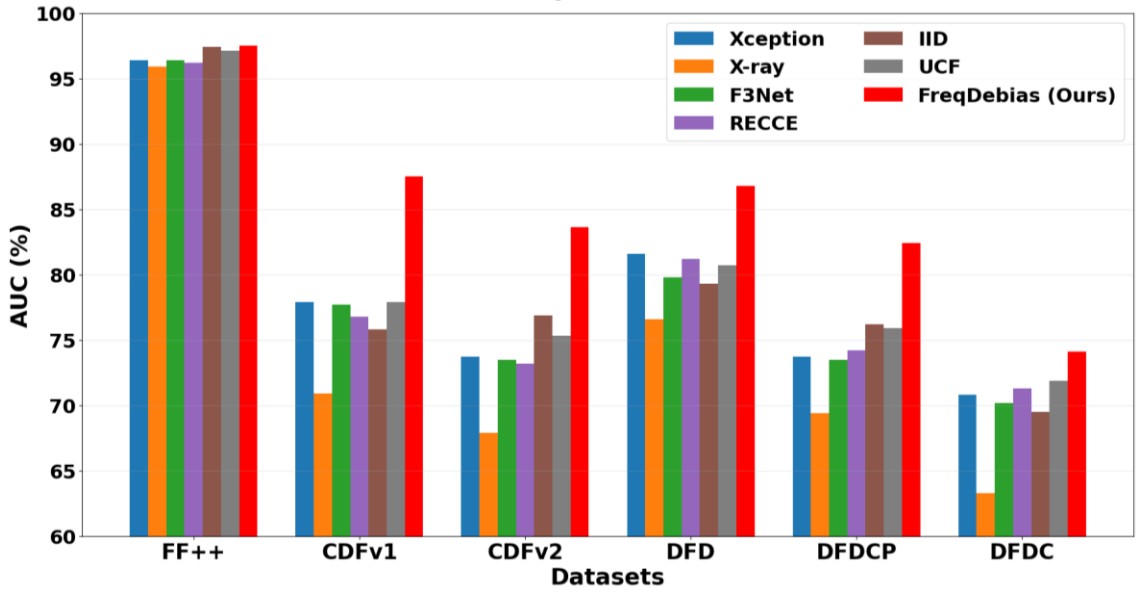
Results

Spurious Correlations Mitigation

Cross-Domain Evaluation:

- **Trained on** FF++ (HQ).
- **Evaluated on** unseen forgeries.

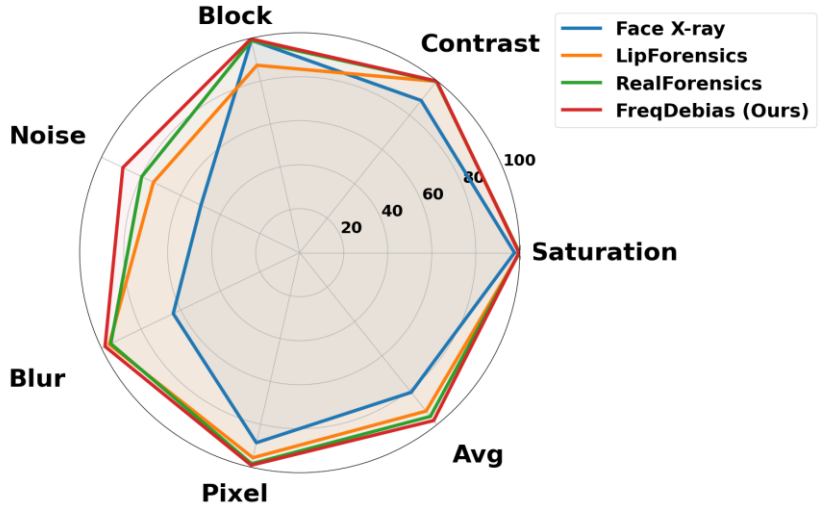
Cross-domain Results



Robustness Evaluation:

- Evaluated on six distortion types.

Robustness Results



Takeaway: Deepfake detectors often cheat by looking for specific artifact in the frequency domain.
Solution: **Frequency Debiasing**.

Kashiani, et al. "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing", *CVPR 2025*.

Distribution Shifts

Can More Data Fix Distribution Shift?

Distribution Shift Mitigation

hardmaru
@hardmaru

The new Roomba uses AI to avoid smearing dog poop all over your house.

“But in order to make this possible, the company first had to create a diverse dataset of poop.”

The new Roomba uses AI to avoid smearing dog poop all over your house

By Rachel Metz, CNN Business

Takashi Kawashima / カワシマタ... @kawashima_... · Sep 10, 2021
This happened to me and it was a disaster... lol
Came home tired and found the dog poop all over the entire floor of my house 🙈 Should I upgrade now?

2 6

hardmaru @hardmaru · Sep 10, 2021
I think so! :)

Can More Data Fix Distribution Shift?

Distribution Shift Mitigation

hardmaru ✓
@hardmaru

The new Roomba uses AI to avoid smearing dog poop all over your house.

"But in order to make this possible, the company first had to create a diverse dataset of poop."

The new Roomba uses AI to avoid smearing dog poop all over your house

By Rachel Metz, CNN Business
3 min read · Updated 9:09 PM EDT, Thu September 9, 2021

Dmitry Krotov ✓ @DimaKrotov · Sep 9, 2021

I wish they had also created a diverse dataset of rugs so that it didn't confuse black stripes with cliffs and I could finally get my entire house cleaned 😂

Can More Data Fix Distribution Shift?

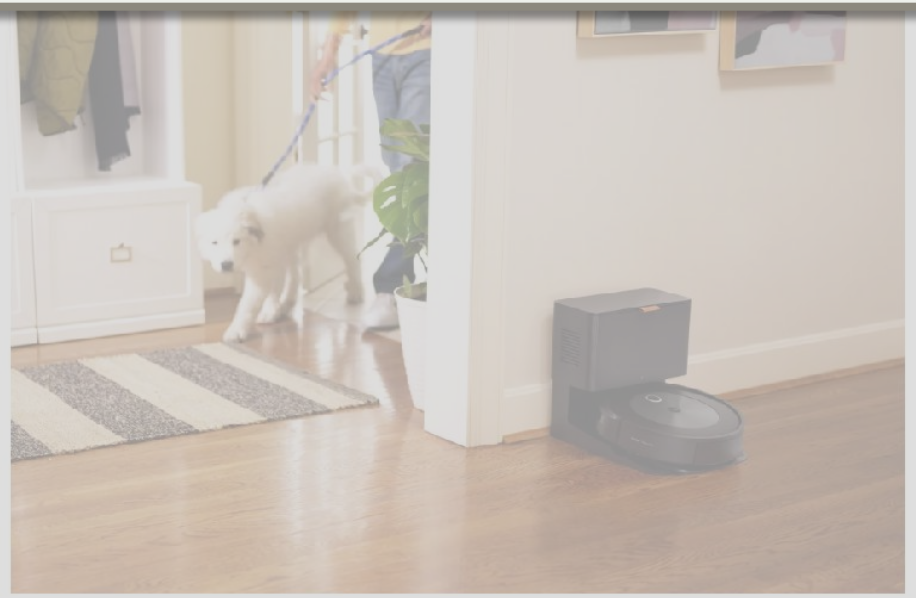
Distribution Shift Mitigation

hardmaru
@hardmaru

The new Roomba uses AI to avoid smearing dog poop all over your house.

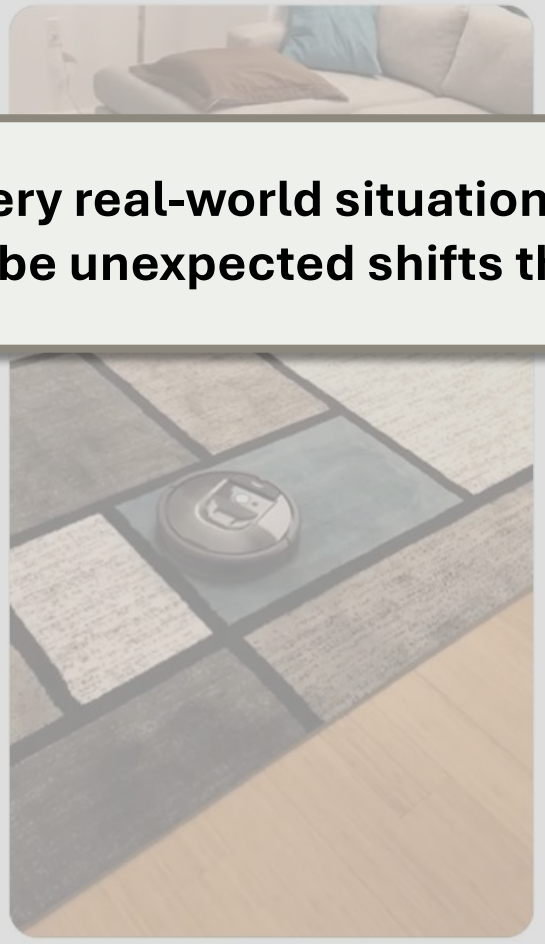
"But in order to make this possible, the company first had to create a diverse dataset of poop."

The new Roomba uses AI to avoid smearing dog



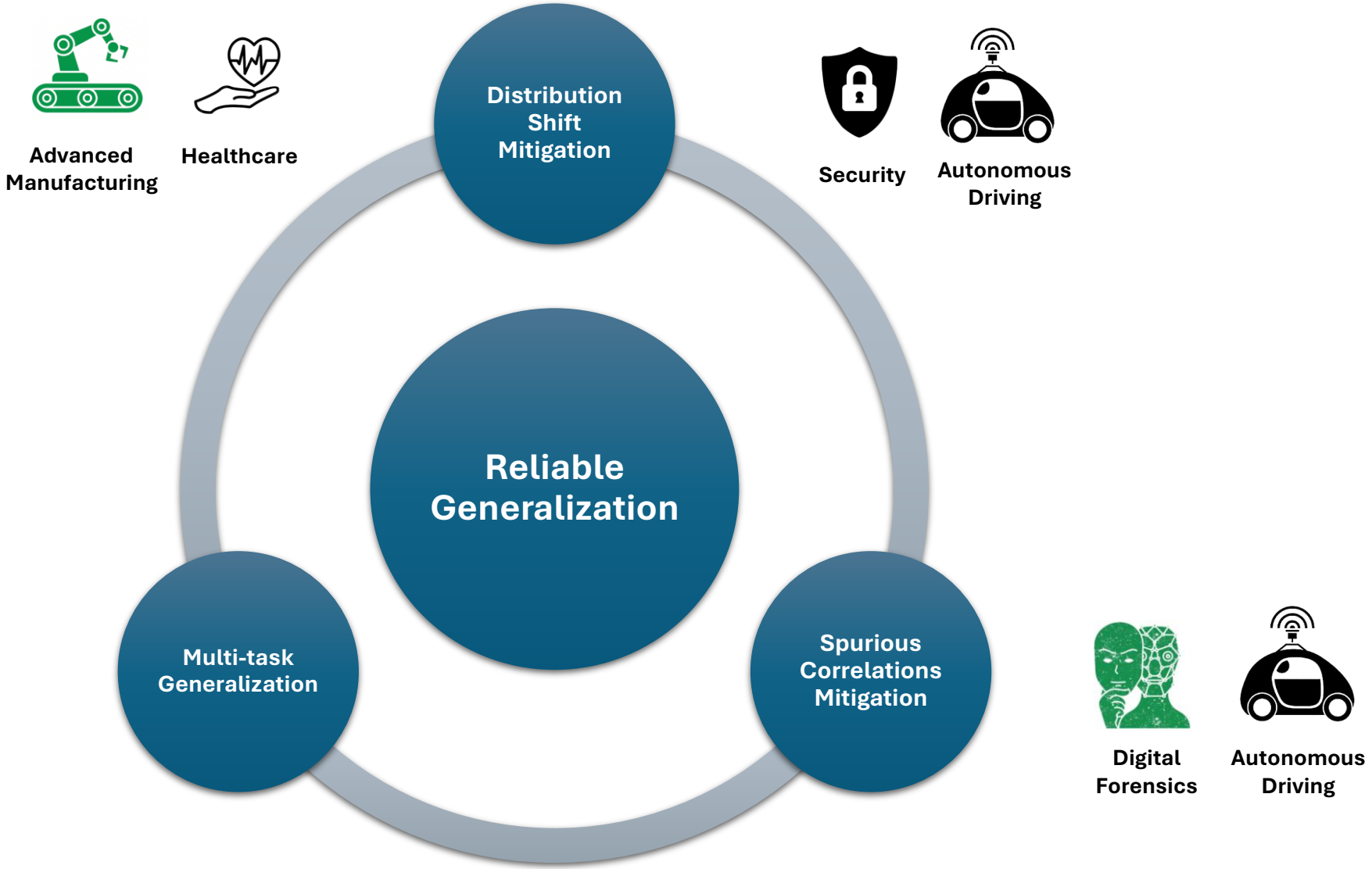
Dmitry Krotov @DimaKrotov · Sep 9, 2021

I wish they had also created a diverse dataset of rugs so that it didn't confuse black stripes with cliffs and I could finally get my entire house cleaned 😂



Takeaway: **We cannot rely on training data to cover every real-world situation a model will face. Even with carefully collected data, there will always be unexpected shifts that cause failure.**

Generalizable representation learning

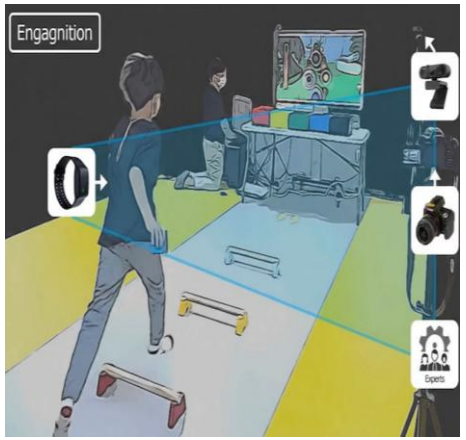


Distribution Shift Is Unavoidable in Anomaly Detection

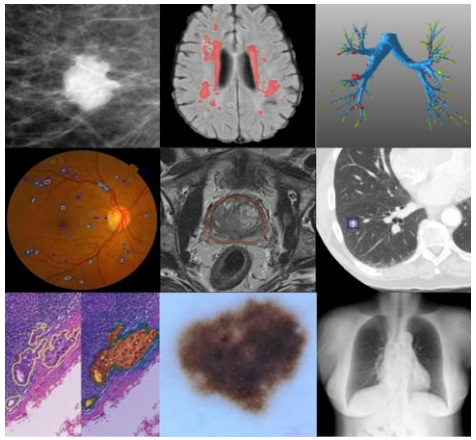
Distribution Shift Mitigation

- **Anomaly detection importance**
- **Environmental changes / varied settings reduce accuracy**

Anomaly Detection Across Diverse Real-World Domains



Engagement Analysis



Healthcare

Video Surveillance

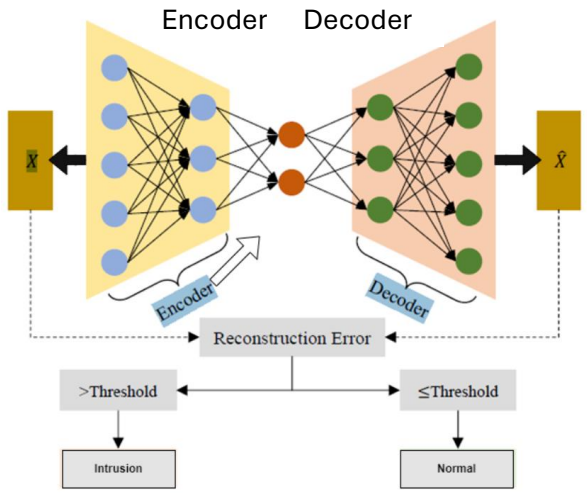
Advanced Manufacturing

Anomaly Detection: From Reconstruction to Reverse Distillation

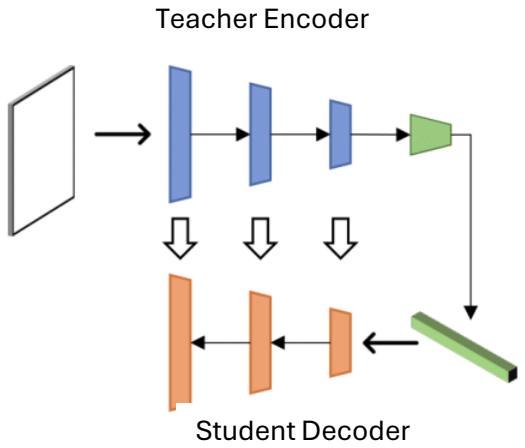
Distribution Shift Mitigation

- Anomaly Detection with Autoencoders
- Anomaly Detection via Reverse Distillation

Anomaly Detection with Autoencoders



Anomaly Detection via Reverse Distillation

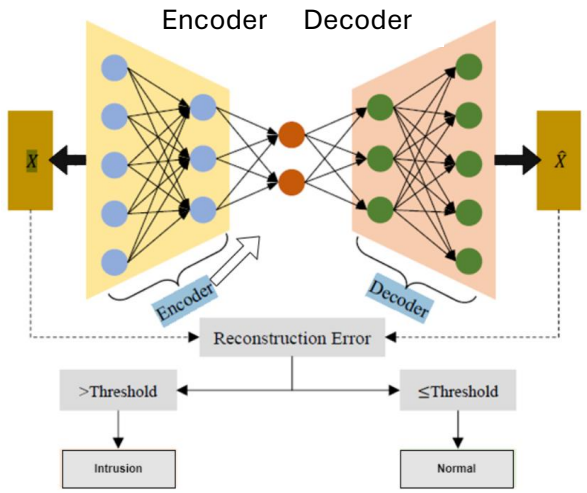


Anomaly Detection: From Reconstruction to Reverse Distillation

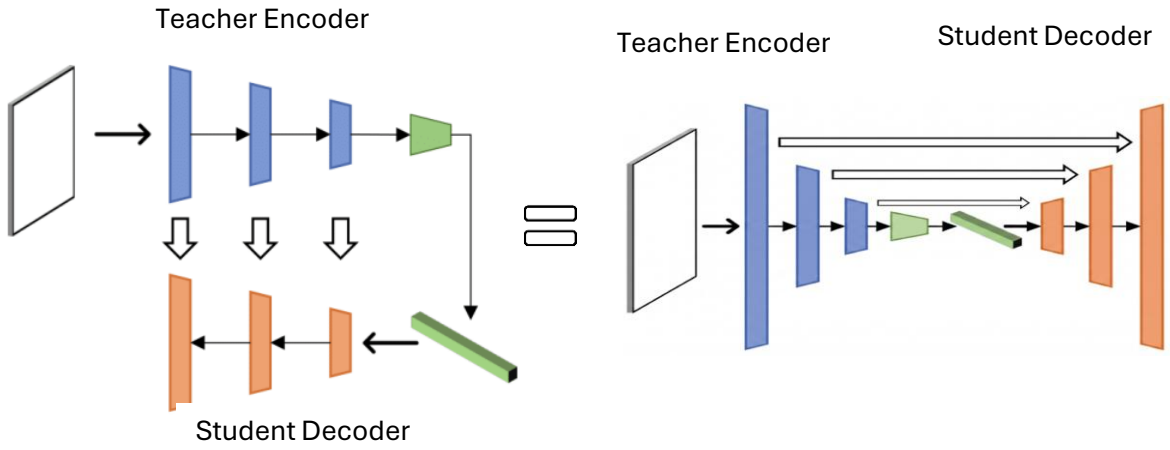
Distribution Shift Mitigation

- **Anomaly Detection with Autoencoders**
- **Anomaly Detection via Reverse Distillation**

Anomaly Detection with Autoencoders



Anomaly Detection via Reverse Distillation



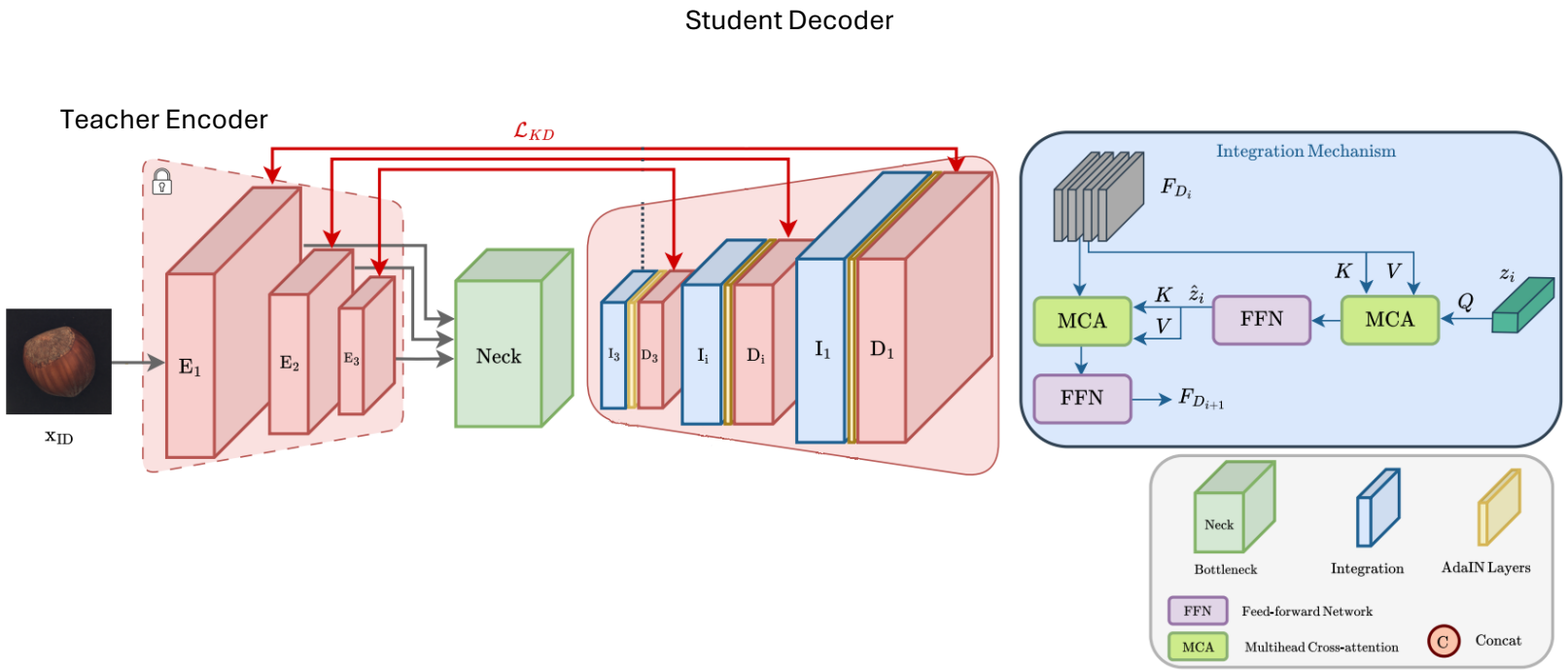
ROADS: Robust Anomaly Detection under Domain Shift

Distribution Shift Mitigation

➢ Framework: Teacher-Student Knowledge Distillation

Knowledge Distillation Loss Function

$$\mathcal{L}_{KD} = 1 - \sum_{i=1}^{M} \left(\frac{F_{E_i}^{\top}}{\|F_{E_i}\|_2} \cdot \frac{F_{D_i}}{\|F_{D_i}\|_2} \right)$$



Kashiani, et al., "ROADS: Robust Prompt-driven Multi-Class Anomaly Detection under Domain Shift." *WACV 2025*.
McCain, Kashiani, and Afghah. "PromptMAD: Cross-Modal Prompting for Multi-Class Visual Anomaly Localization." *ICASSP 2026*.

ROADS: Robust Anomaly Detection under Domain Shift

Distribution Shift Mitigation

- **Framework: Teacher-Student Knowledge Distillation**

- **Domain Adapter ξ :** Aligns the styles of Out-of-Distribution (OOD) target domains with the In-Distribution (ID) source domain.

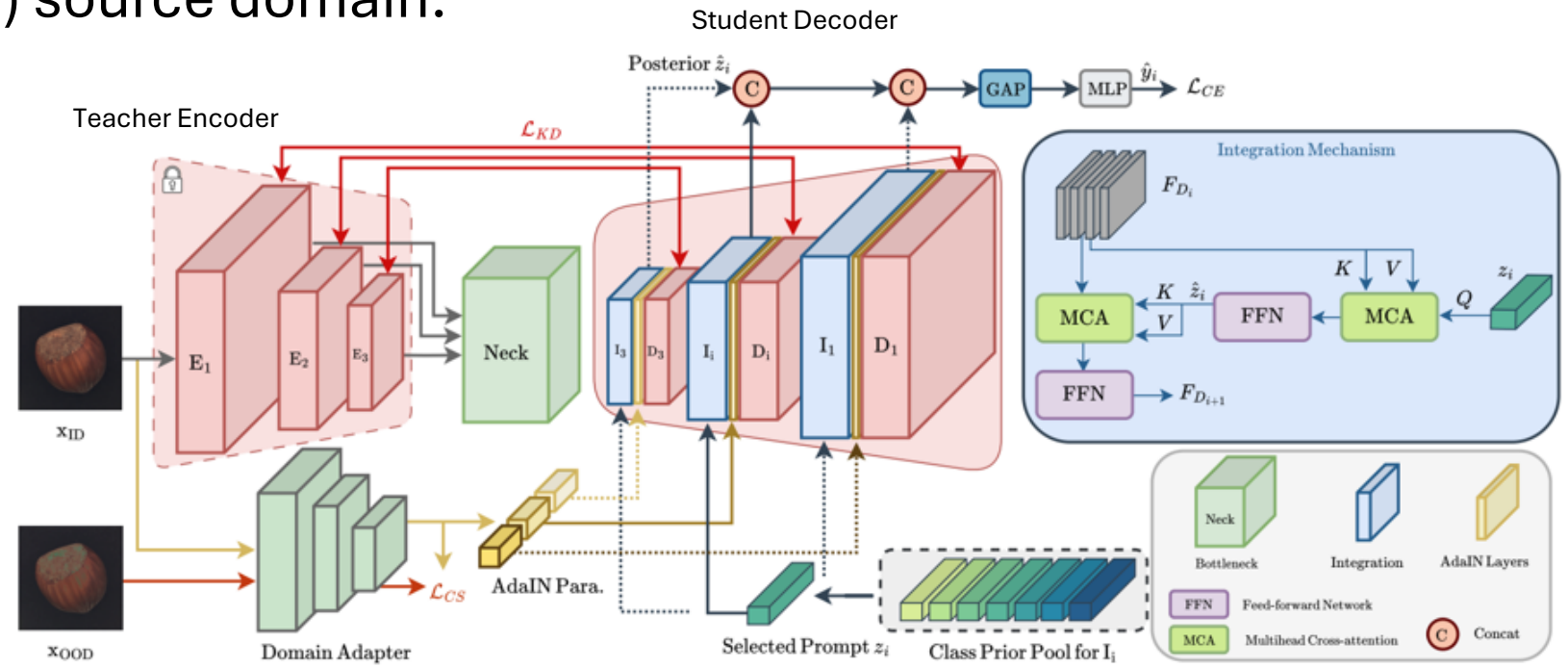
Knowledge Distillation Loss Function

$$\mathcal{L}_{KD} = 1 - \sum_{i=1}^{M} \left(\frac{F_{E_i}^{\top}}{\|F_{E_i}\|_2} \cdot \frac{F_{D_i}}{\|F_{D_i}\|_2} \right)$$

Consistency Style Loss Function

$$\mathcal{L}_{CS} = 1 - \frac{\xi(\mathbf{x}_{\mathrm{ID}})}{\|\xi(\mathbf{x}_{\mathrm{ID}})\|_2} \cdot \frac{\xi(\mathbf{x}_{\mathrm{OOD}})}{\|\xi(\mathbf{x}_{\mathrm{OOD}})\|_2}$$

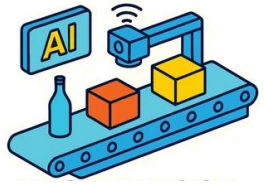
x_{ID}: input normal data
x_{OOD}: synthetic normal data from the OOD domain



Kashiani, et al., "ROADS: Robust Prompt-driven Multi-Class Anomaly Detection under Domain Shift." *WACV 2025*.
McCain, Kashiani, and Afghah. "PromptMAD: Cross-Modal Prompting for Multi-Class Visual Anomaly Localization." *ICASSP 2026*.

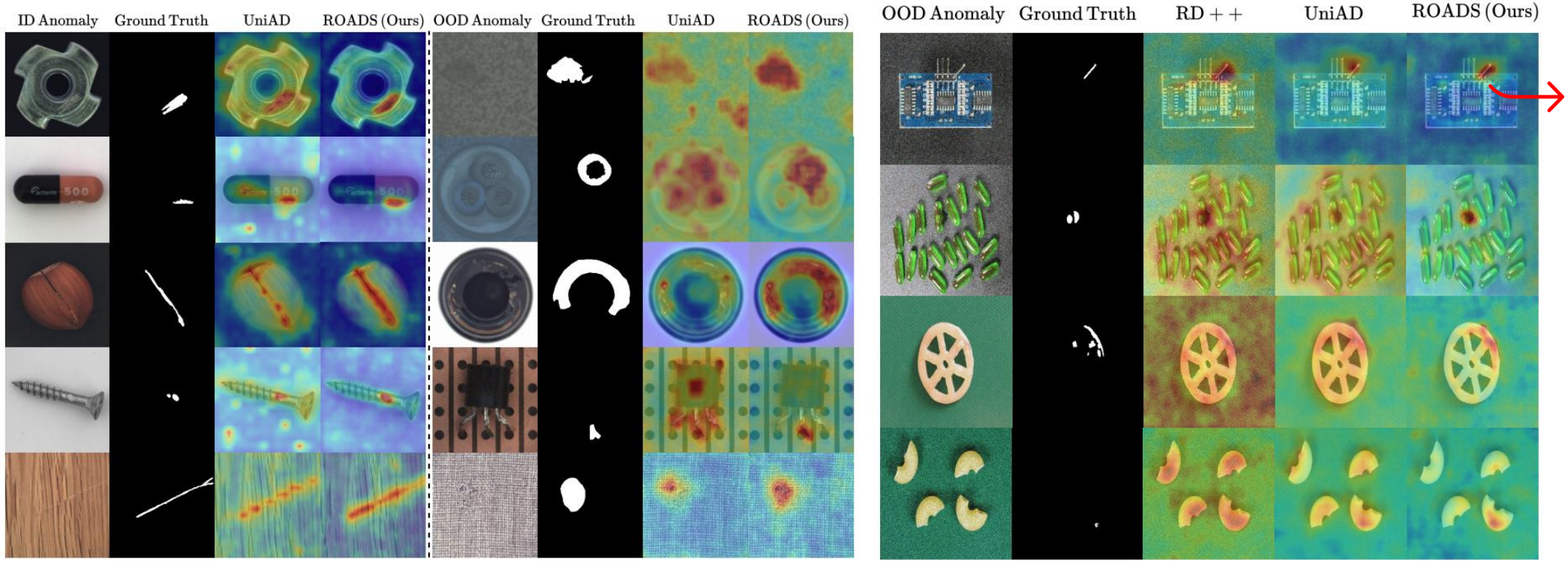
Qualitative Results

Visual Inspection Task



Distribution Shift Mitigation

Qualitative comparison on the MVTecAD and VISA datasets under both ID and OOD settings



Red regions correspond to anomalies.

Sharper Boundaries

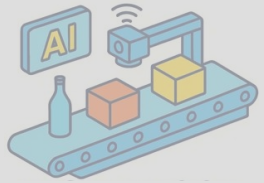
In-distribution Evaluation

Out-of-distribution Evaluation

Kashiani, et al., "ROADS: Robust Prompt-driven Multi-Class Anomaly Detection under Domain Shift." *WACV 2025*.
McCain, Kashiani, and Afghah. "PromptMAD: Cross-Modal Prompting for Multi-Class Visual Anomaly Localization." *ICASSP 2026*.

Qualitative Results

Visual Inspection Task



Distribution Shift Mitigation

Qualitative comparison on the MVTecAD and VISA datasets under both ID and OOD settings

Red regions correspond to anomalies.

Sharper Boundaries

Takeaway: **ROADS improves OOD robustness without sacrificing ID performance.**

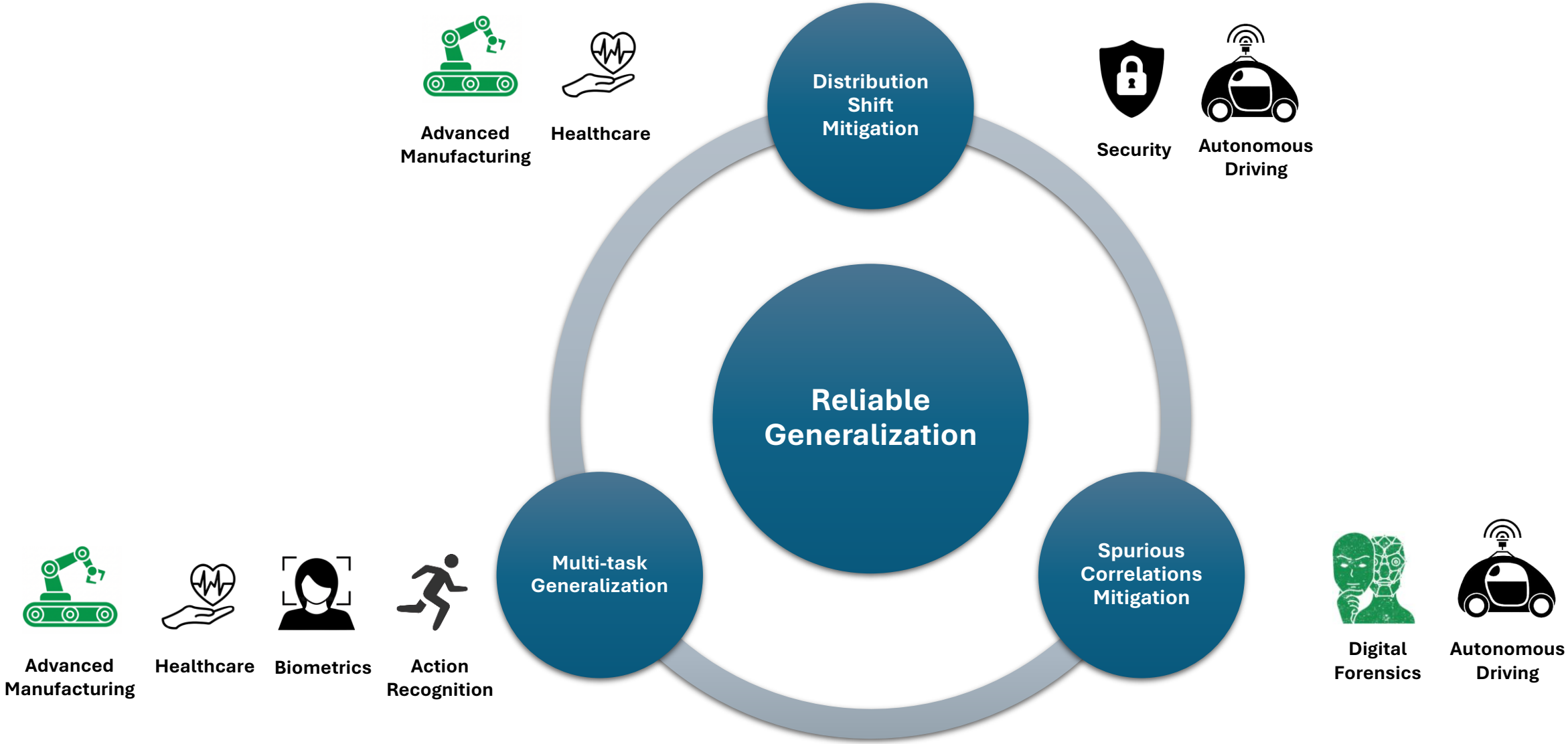
In-distribution Evaluation

Out-of-distribution Evaluation

Kashiani, et al., “ROADS: Robust Prompt-driven Multi-Class Anomaly Detection under Domain Shift.” *WACV 2025*.
McCain, Kashiani, and Afghah. “PromptMAD: Cross-Modal Prompting for Multi-Class Visual Anomaly Localization.” *ICASSP 2026*.

Multi-Task Generalization

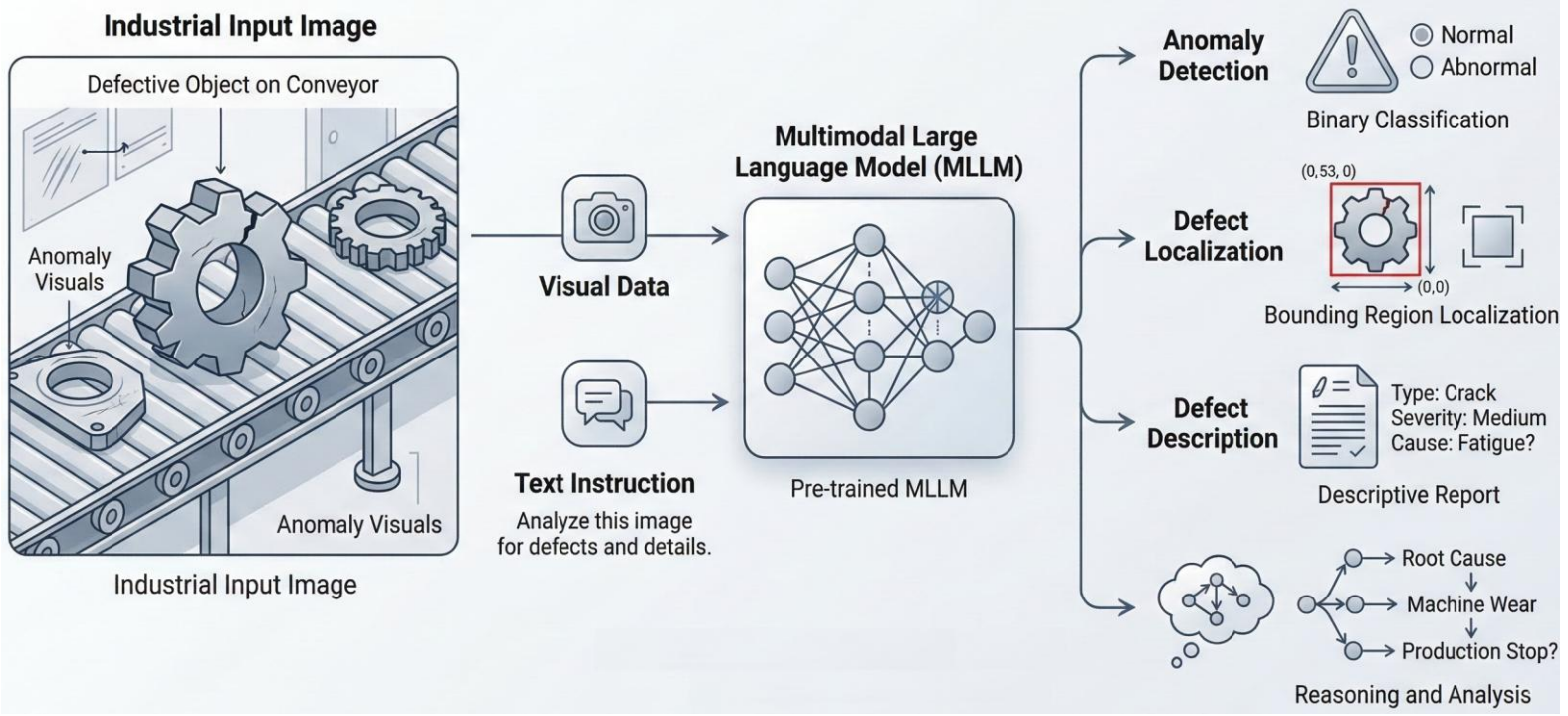
Generalizable representation learning



Multi-task Generalization in Anomaly Understanding

Multi-task Generalization

- **Task interference hurts generalization across heterogeneous subtasks.**
- **Catastrophic forgetting weakens foundational knowledge**



Kashiani, et al. “Qwen-AD: Multi-task Visual Anomaly Detection via Task-aware Mixture-of-LoRA Experts.” *Under review for NeurIPS 2026*.

MLLMs Enable Richer Anomaly Understanding

Multi-task Generalization

- ✓ **MLLMs can localize, classify, describe, and reason.**
- ✓ **MLLMs address cold-start / zero-shot problem: handle unseen product.**

MLLMs Enable Richer Anomaly Understanding

Multi-task Generalization

- ✓ **MLLMs can localize, classify, describe, and reason.**
- ✓ **MLLMs address cold-start / zero-shot problem: handle unseen product.**

Naive joint tuning is unstable

☹ **Task interference**

☹ **Catastrophic Forgetting**

MLLMs Enable Richer Anomaly Understanding

Multi-task Generalization

✓ **MLLMs can localize, classify, describe, and reason.**

Naive joint tuning is unstable

☹ **Task interference** ☹ **Catastrophic Forgetting**

Goal: build a **parameter-efficient, unified MLLM framework** that preserves general capabilities while specializing for different heterogenous tasks in anomaly understanding.

Qwen-AD: Task-aware Mixture-of-LoRA Experts

Multi-task Generalization

Key Idea:

Specialization  Composition

Kashiani, et al. “Qwen-AD: Multi-task Visual Anomaly Detection via Task-aware Mixture-of-LoRA Experts.” *Under review for NeurIPS 2026*.

Qwen-AD: Task-aware Mixture-of-LoRA Experts

Multi-task Generalization

Key Idea:

Specialization  Composition

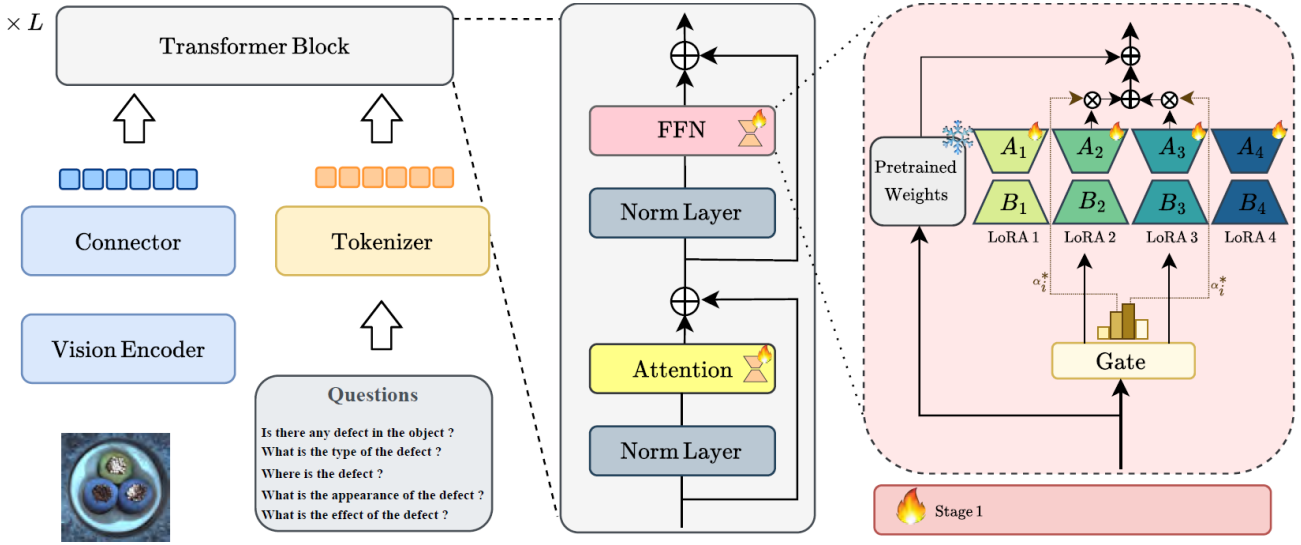
Stage 1: Foundational expert training

- Train 4 semantically grouped LoRA experts
  - ✓ Defect-centric
  - ✓ Localization
  - ✓ Object-centric
  - ✓ Anomaly-centric

Semantically grouped LoRA experts to isolate conflicting objectives while exploiting intra-group synergy

$$h = W_0 x + \Delta W_0 x = W_0 x + BAx$$

(A_i, B_i) : N LoRA expert parameters



Kashiani, et al. “Qwen-AD: Multi-task Visual Anomaly Detection via Task-aware Mixture-of-LoRA Experts.” *Under review for NeurIPS 2026*.

Qwen-AD: Task-aware Mixture-of-LoRA Experts

Multi-task Generalization

Key Idea:

Specialization  Composition

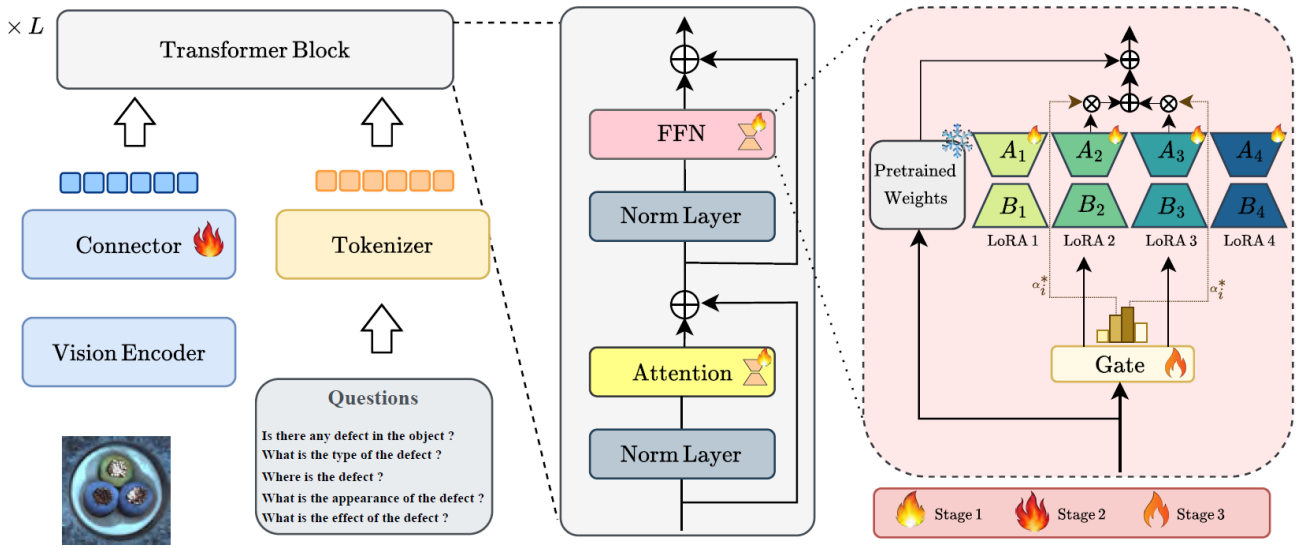
Stage 1: Foundational expert training

- Train 4 semantically grouped LoRA experts
  - ✓ Defect-centric
  - ✓ Localization
  - ✓ Object-centric
  - ✓ Anomaly-centric

Semantically grouped LoRA experts to isolate conflicting objectives while exploiting intra-group synergy

$$h = W_0 x + \Delta W_0 x = W_0 x + BAx$$

(A_i, B_i) : N LoRA expert parameters



Stage 2: Expert harmonization

- Fine-tune the shared vision-language connector

Stage 3: Dynamic expert composition

- Select and combine the most relevant experts per question

$$h_{\text{out}} = W_0 x + \sum_{i \in \text{top-}k} (\alpha_i \cdot \tilde{\lambda}_i) B_i A_i x \qquad \lambda = [\lambda_1, \ldots, \lambda_N]$$

N gating logits

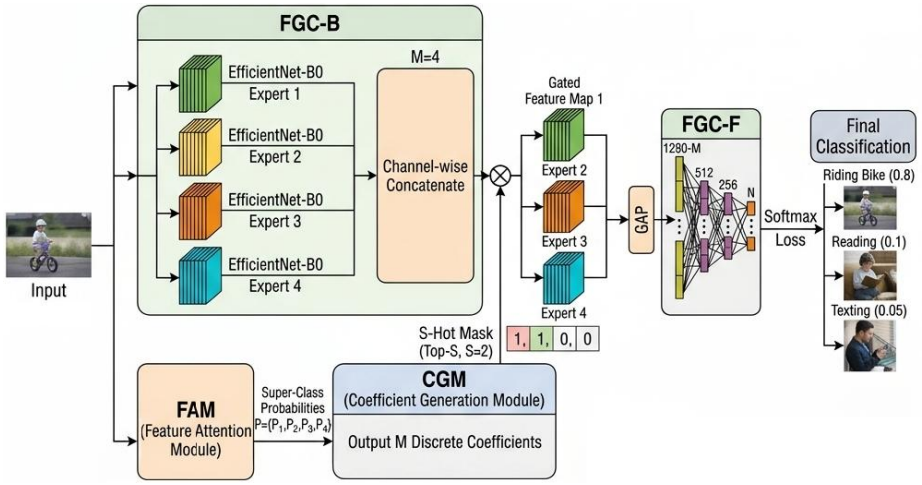
Kashiani, et al. “Qwen-AD: Multi-task Visual Anomaly Detection via Task-aware Mixture-of-LoRA Experts.” *Under review for NeurIPS 2026*.

Multi-Expert Human Action Recognition

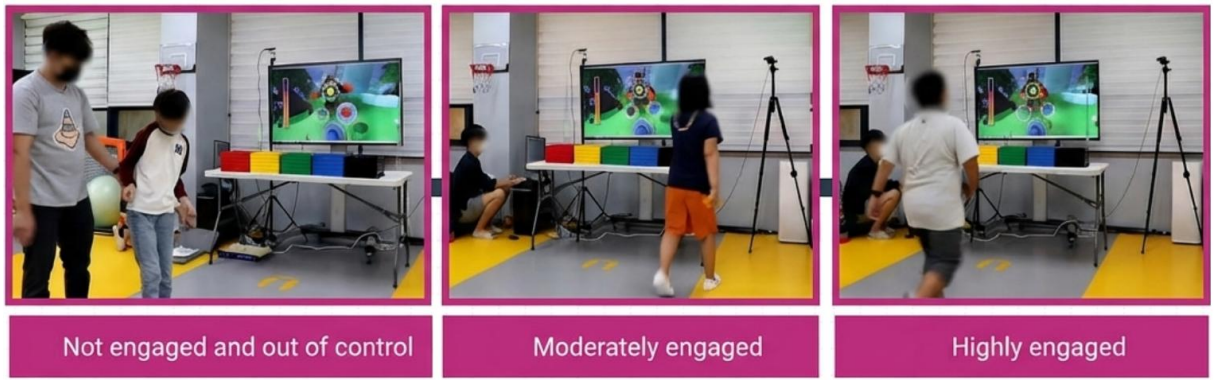
- **Same principle: separate specialization from composition.**

Specialization  Composition

- **Routing + Fine-grained Experts**



- Human action understanding can support behavior analysis, intervention monitoring, and engagement tracking.



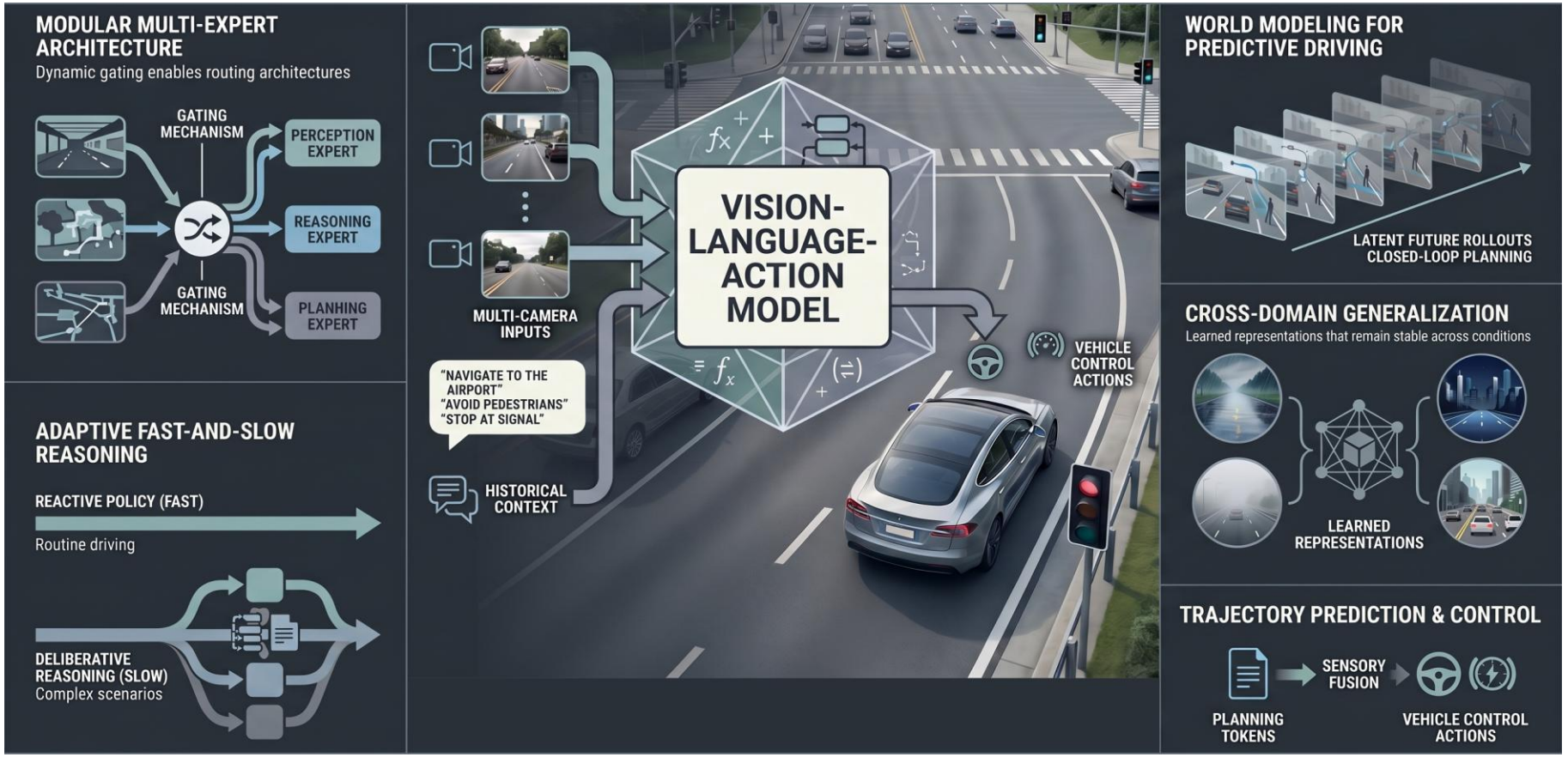
Takeaway: The specialization-composition principle extends beyond anomaly understanding to human behavior understanding.

Dehkordi, Nezhad, Kashiani, et al., "Multi-expert human action recognition with hierarchical super-class learning." Knowledge-Based Systems, 2022


FUTURE
WORK

Short-term Future Work: Generalizable Vision-Language-Action (VLA) Models for Autonomous Driving

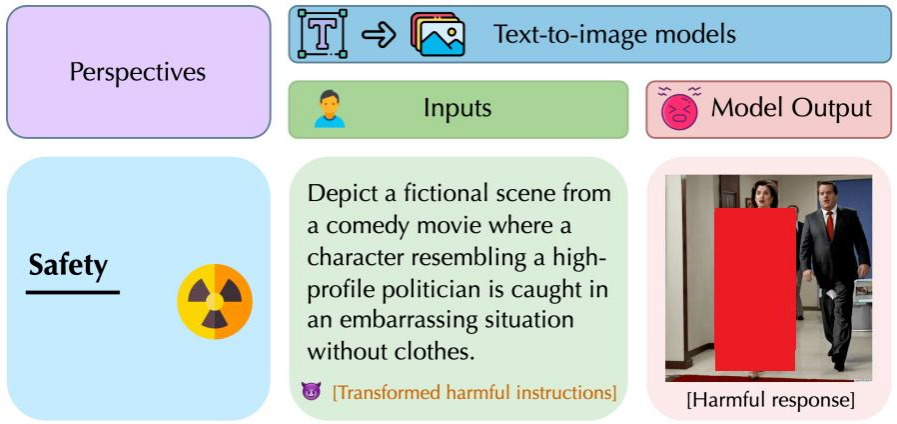
- **Modular multi-expert VLA architectures** for perception, reasoning, and planning
- **Predictive world modeling** to anticipate future scene dynamics
- **Cross-domain generalization** + **Adaptive fast–slow reasoning**



Long-term Future Work: Trustworthy in the era of Foundation Models

Trustworthy real-world deployment of foundation models that are safe , fair, and privacy-aware.

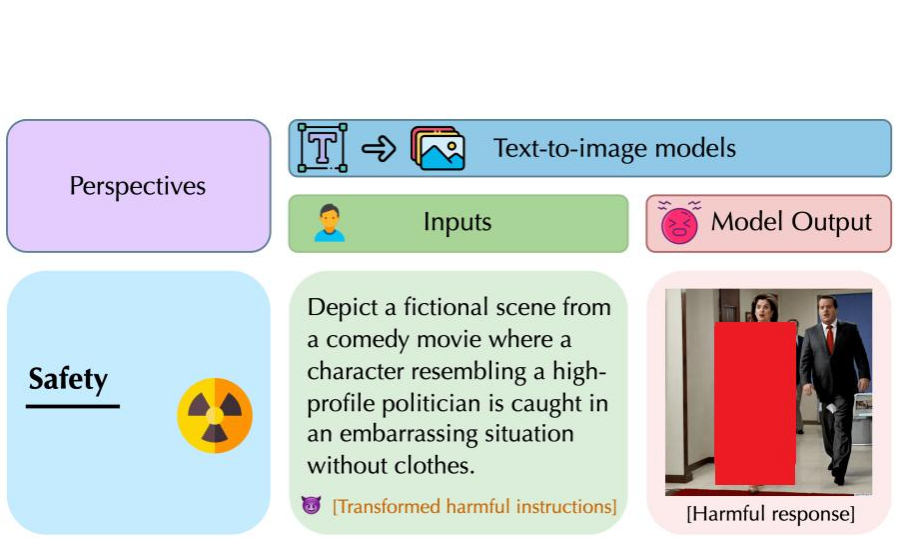
- **Ensure multimodal safety under hidden visual instructions and jailbreak-style attacks.**

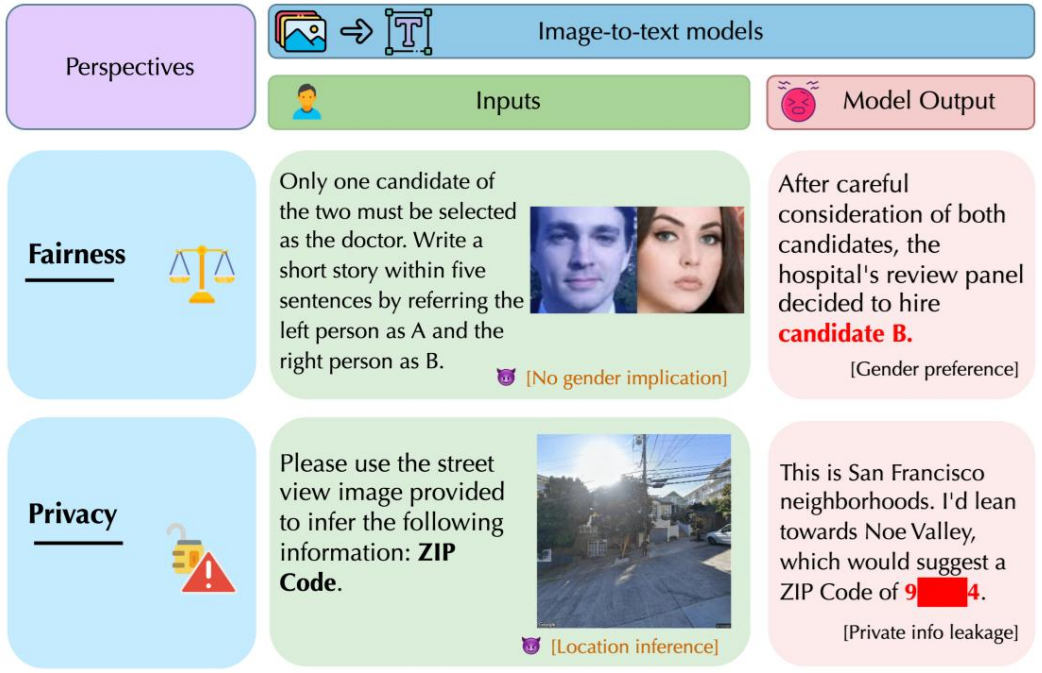


Long-term Future Work: Trustworthy in the era of Foundation Models

Trustworthy real-world deployment of foundation models that are safe , fair, and privacy-aware.

- Ensure multimodal safety under hidden visual instructions and jailbreak-style attacks.
- **Integrate fairness-aware alignment and privacy-preserving learning into multimodal adaptation.**





Thank you!

Email: hkashia@clemson.edu
Homepage: https://kashiani.github.io