Unlocking Visual Anomaly Detection

The Evolution of Visual Anomaly Detection
VLMs, MLLMs, and Generalizable Inspection

Hossein Kashiani

Clemson University

December 3, 2025

Overview

- **The Imperative:** The Trillion Dollar Problem.
- **Problem Definition:** Visual Anomaly Detection.
- **Key Challenges:** Localization, Complexity, Camouflage, Shifts.
- **The Classical Era:** Limitations of Reconstruction & Distillation.
- **ROADS:** Robust Multi-Class AD under Domain Shift (WACV 25).
- **The Semantics Pivot (VLM):**
  - WinCLIP (CVPR 23)
  - AnomalyCLIP (ICLR 24)
- **The Reasoning Leap (MLLM):**
  - AnomalyGPT (AAAI 24)
- **Evaluation:** The MMAD Benchmark (ICLR 25).
- **Future Directions:** Where do we go from here?

The Trillion Dollar Problem: Where Anomalies Matter

- **Where Anomalies Matter:**
  - **Manufacturing & QC:** Defective products, process drift, equipment wear.
  - **Cybersecurity:** Intrusions, fraud, unusual traffic patterns.
  - **Healthcare:** Abnormal imaging findings, vital sign outliers, rare conditions.
  - **Other Domains:** Finance, transportation, energy, environmental monitoring.
- **The Economic & Safety Stakes:**
  - Global manufacturing defects alone cost trillions annually.
  - **Critical Impact:** Safety (automotive/aerospace), patient health (pharma), and brand reputation.
- **The Unsupervised Imperative:**
  - Manual inspection cannot scale to high-speed lines (1000+ items/min).
  - **Data Scarcity:** Defects are rare ($< 5\%$) and highly variable; we cannot supervise every type.

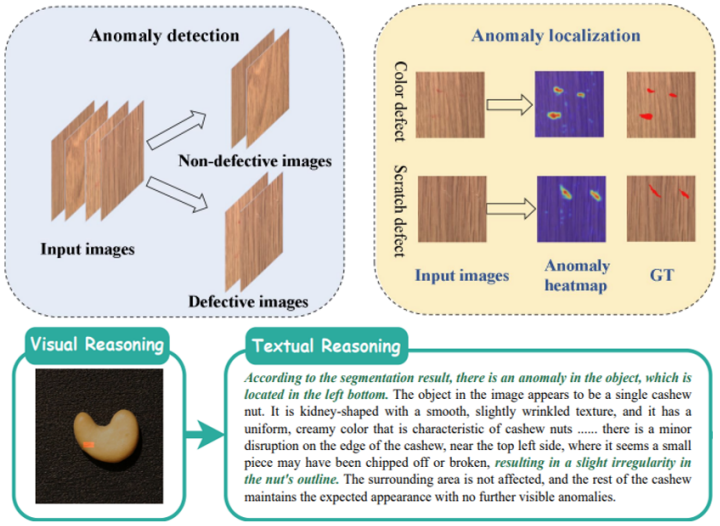
Problem Definition: Visual Anomaly Detection

Core Task

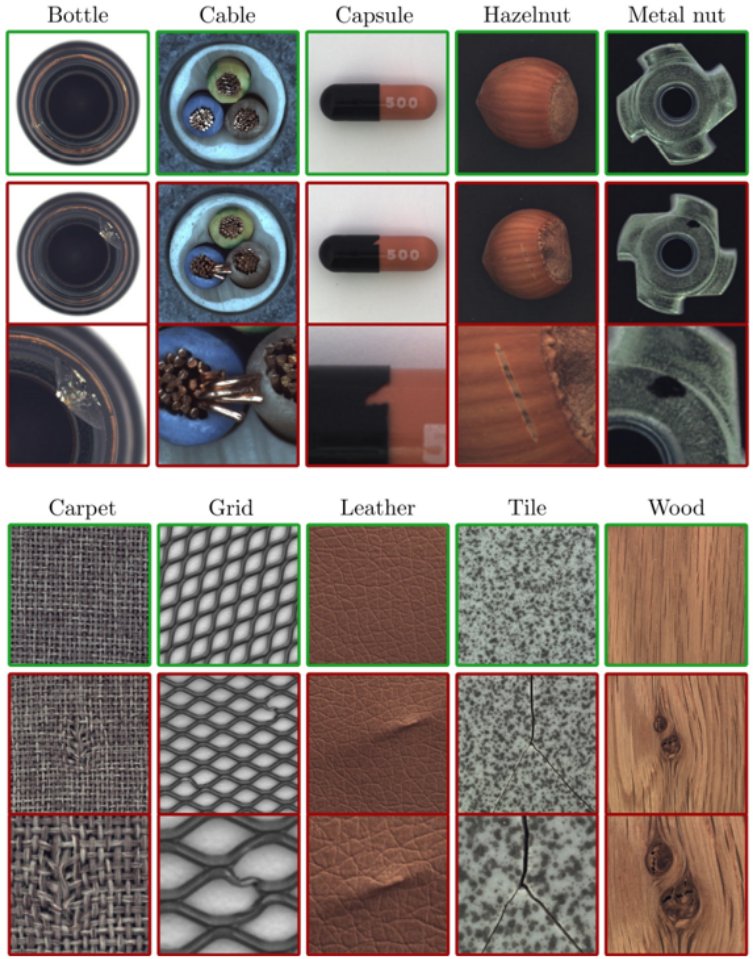
- Train on **normal only** (or zero images).
- Test: Identify deviations (pixel-level).

Goal of Modern VAD:

- Move from "One-Class-One-Model" to **Zero-Shot Generalization**.
- Detect unseen defects on unseen products without retraining.



Objectives in Visual Anomaly Detection

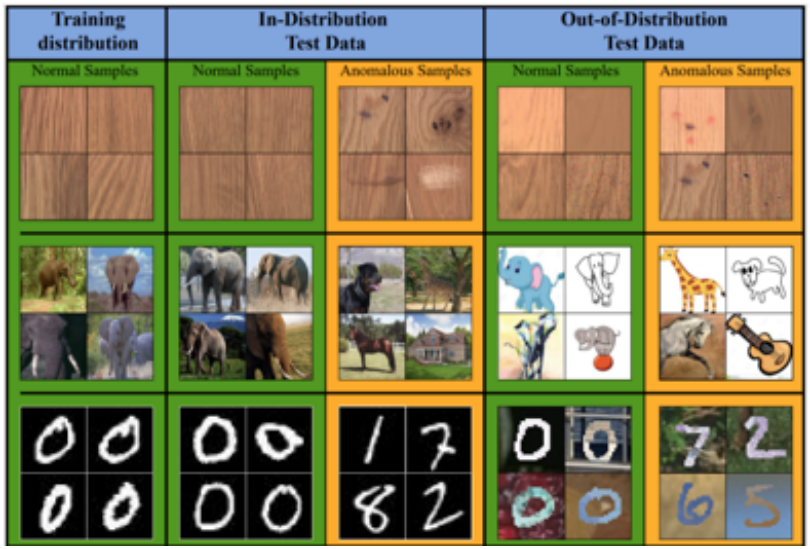


Normal vs. Anomalous (MVTec-AD)

Key Challenges in Visual Anomaly Detection

Technical Hurdles

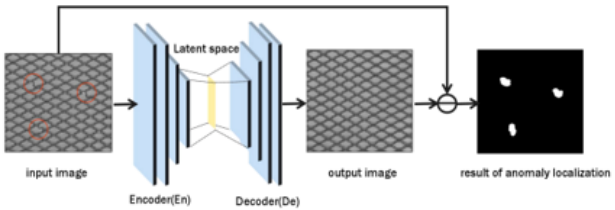
- **Precise Localization:**
  Requires pixel-level accuracy and sharp boundaries.
- **Multi-Class Complexity:**
  Handling diverse objects vs. textures in a single model.
- **Extreme Class Imbalance:**
  Rare anomalies ($< 1\%$); models tend to predict all-normal.
- **Domain Shifts:**
  Changes in illumination, sensors, background, and process.



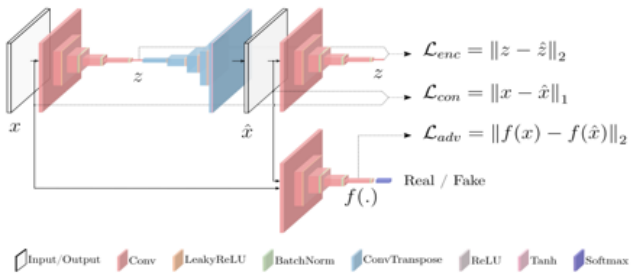
Domain Shift Challenge

The Classical Era: Reconstruction & Distillation

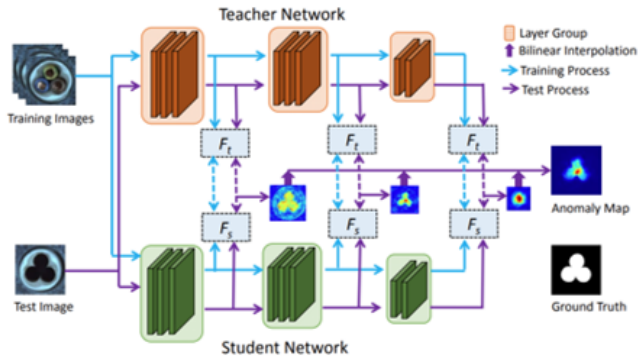
- **Reconstruction (AE/GAN):**
  - If I can't reconstruct it, it's a defect.
  - Issue: Sometimes reconstructs anomalies too.
- **Knowledge Distillation:**
  - Student mimics Teacher on normal data only.
  - Issue: Fails under domain shift.



Vanilla Auto-encoder



GANomaly (GAN-Based)



Teacher-Student Knowledge Distillation

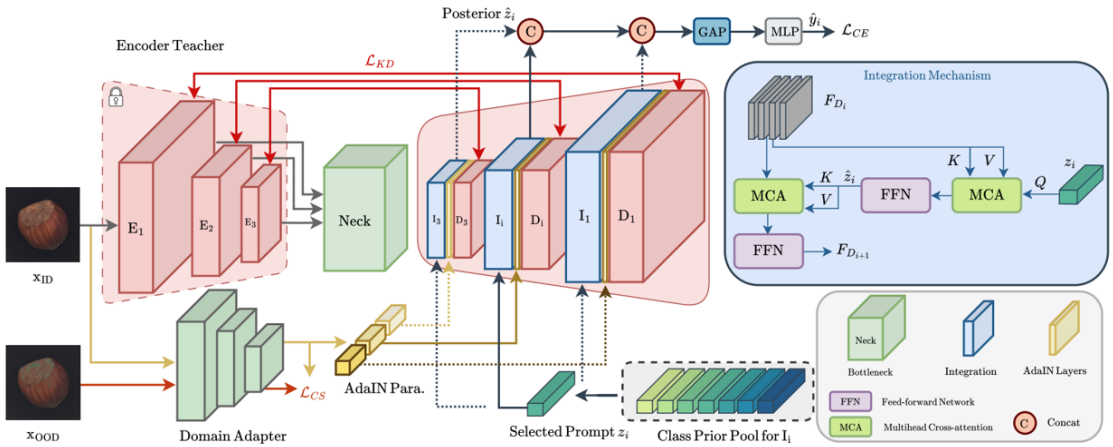
The Generalization Gap

These methods fail the *Cold Start* problem. They require training data for *every* new product category.

ROADS: Robust Multi-Class AD under Domain Shift (WACV 2025) [1]

Domain Adapter with AdaIN

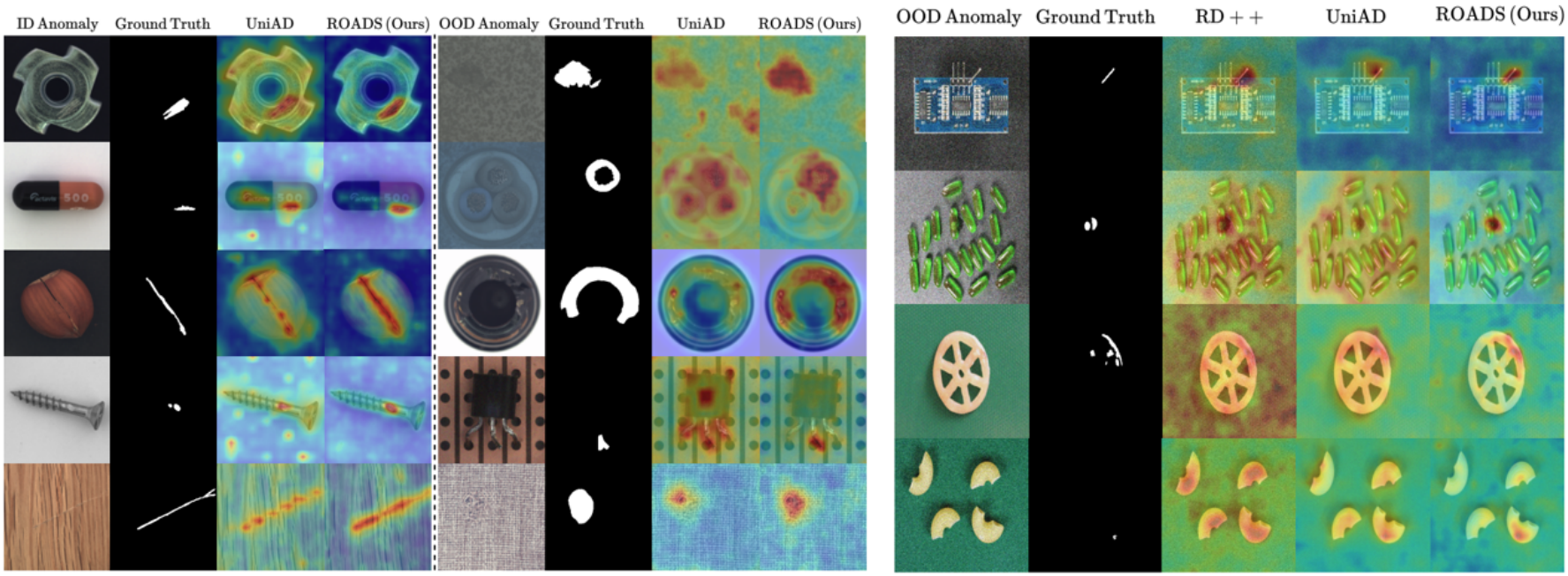
- **Challenge:** Standard unified detectors fail when test conditions differ from training (e.g., lighting changes, sensor noise).
- **Mechanism:** Uses Adaptive Instance Normalization (AdaIN) layers within the student decoder.
- **Style Alignment:** A dedicated adapter network predicts *domain-invariant style codes*.
- **Result:** Dynamically aligns the statistics (mean/variance) of OOD input features to match the Normal source distribution.



ROADS Architecture

ROADS: Visual Results on MVTec and VISA

- The qualitative results illustrates the clear superiority of our method over UniAD in OOD scenarios.
- While UniAD struggles to accurately localize anomalies under distribution shifts, ROADS consistently identifies anomalies, demonstrating greater robustness in such conditions.



Visualization of anomaly maps.

Part I: The Semantic Pivot (VLMs)

From detecting pixel errors to understanding damaged objects.

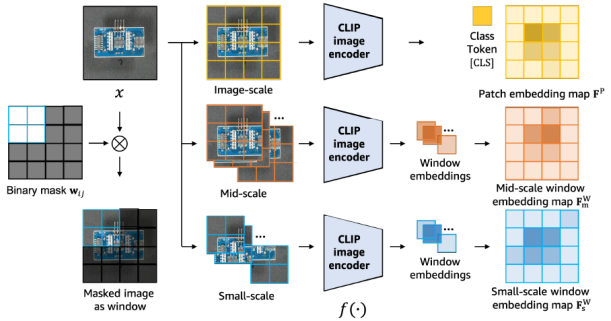
- Leveraging CLIP for Zero-Shot Detection.
- WinCLIP, and AnomalyCLIP.

WinCLIP: The Pioneer (CVPR 2023) [2]

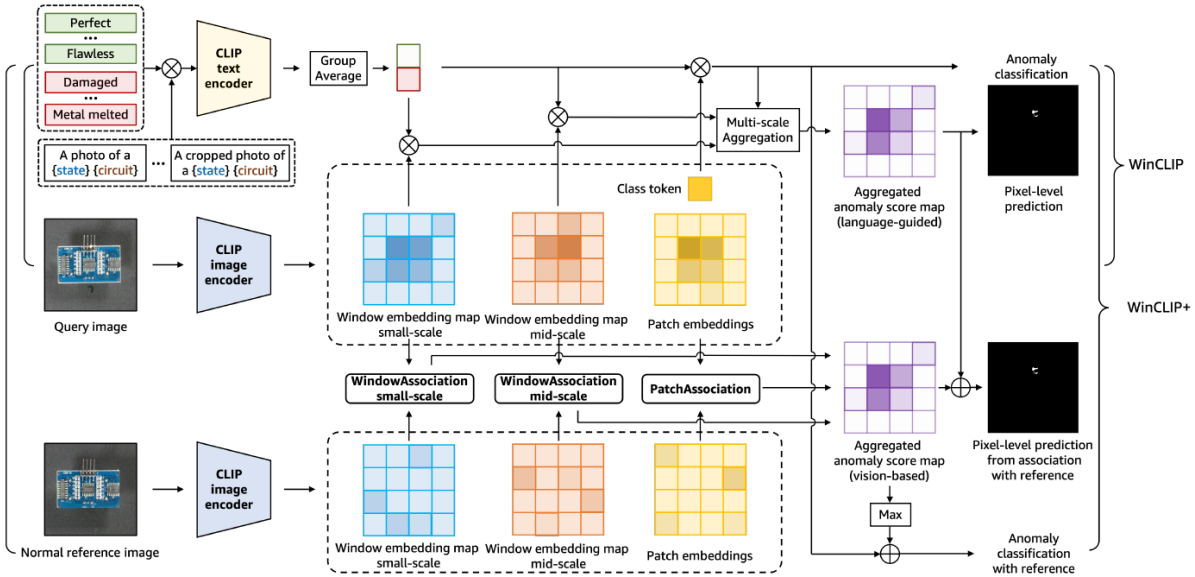
Intuition: Normality is a state defined by language.

Methodology

1. **Compositional Prompt Ensemble:**
   - Normal: A photo of a [flawless] [bottle]
   - Anomaly: A photo of a [damaged] [bottle]
2. **Window-based CLIP:**
   - Slide local windows and feed them into CLIP for text-aligned embeddings.



WinCLIP feature extraction



WinCLIP Architecture

AnomalyCLIP: Fixing the Focus (ICLR 2024) [3]

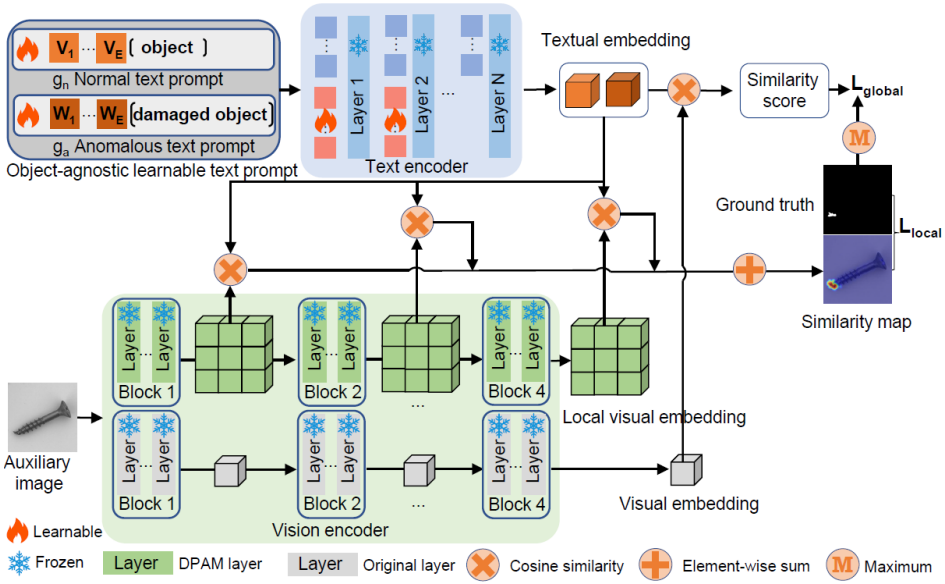
The Problem with CLIP:

- CLIP focuses on *Object semantics* ("This is a cat"), not *State semantics* ("This cat is sick").
- Embeddings are dominated by the object identity.

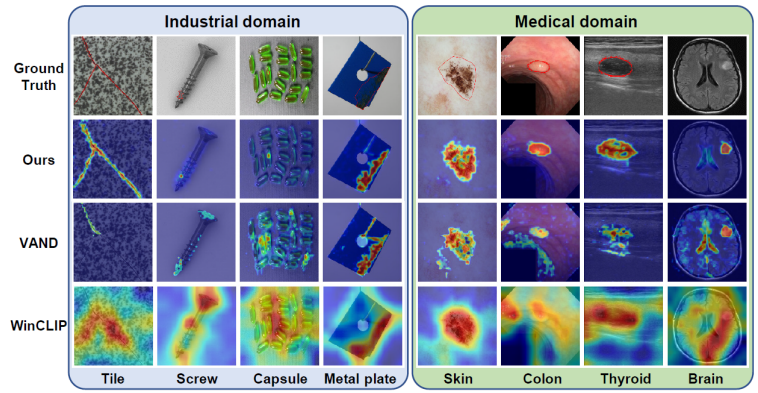
The Solution:

- **Object-Agnostic Prompting:** Learn generic "[V1]... object" vs "[W1]... damaged object" prompts.

Results: Zero-shot transfer to new industrial and medical datasets without target retraining



AnomalyCLIP Architecture



Segmentation Visualization

Part II: The Reasoning Leap (MLLMs)

From Heatmaps to Explanations and Manufacturing Context.

- AnomalyGPT: Conversational Detection

AnomalyGPT: The Conversation Begins (AAAI 2024) [4]

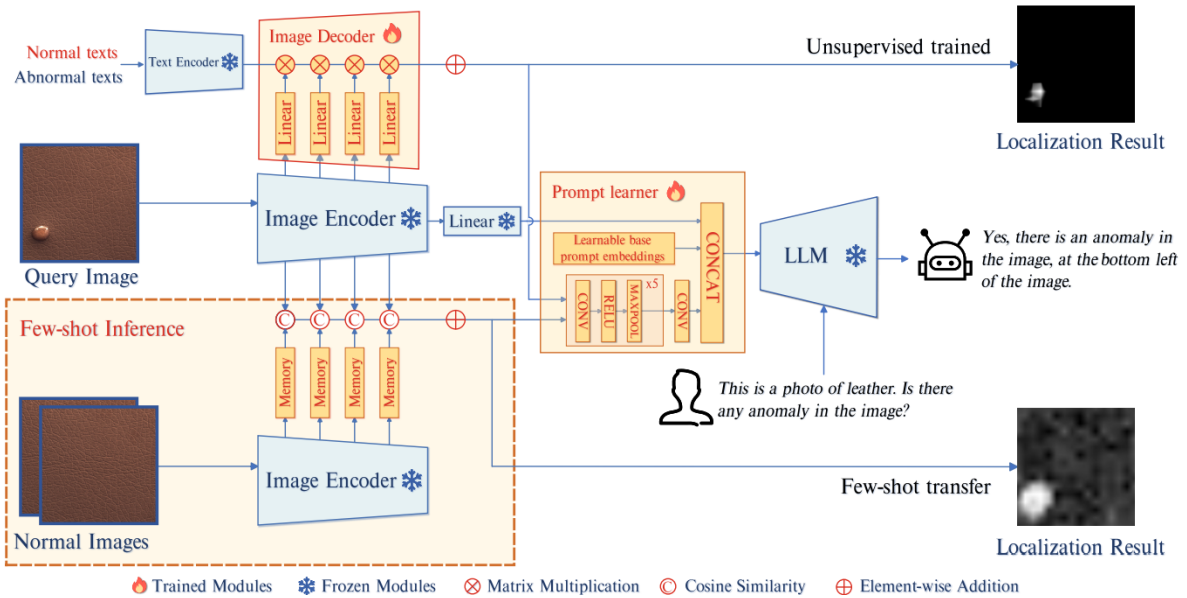
Motivation

- VLMs give a score/heatmap, but no explanation.
- Thresholding is brittle (picking 0.5 vs 0.8).

Method

- **LVLM Backbone:** ImageBind + Vicuna LLM.
- **Pixel-Decoder:** Lightweight module generating a mask.
- **Prompt Learner:** Converts mask info into prompt embeddings for the LLM.

Result: Threshold-free. Just ask: "Is there any anomaly in the image?"



AnomalyGPT Framework

MMAD Benchmark: The Reality Check (ICLR 2025) [5]

Are general MLLMs ready for the factory floor?

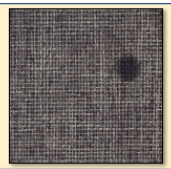
The Benchmark:

- 7 Tasks: Localization, Classification, Description, Analysis, etc.
- Tested GPT-4o, Gemini, LLaVA.

Anomaly Discrimination

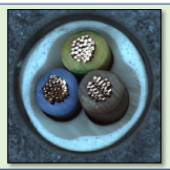
Is there any defect in the object?
A: Yes.
B: No.

Defect Classification



There is a defect in the object. What is the type of the defect?
A: A tear.
B: Discoloration.
C: Wrinkling.
D: Loose thread.

Object Classification



What kind of product is in the image?
A: A section of garden hose.
B: A cross-section of a tri-phase electrical cable.
C: A bundle of fiber optic cables.
D: A piece of computer hardware.

Defect Localization

There is a defect in the object. Where is the defect?
A: Top left candle
B: Top right candle
C: Bottom left candle
D: Bottom right candle

Object Analysis

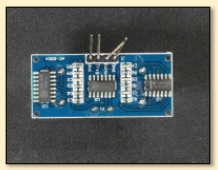
What are the subcomponents of the breakfast box?
A: Oranges, nectarine, granola, nuts, and banana slices
B: Oranges, apples, cereal, and dried fruit
C: Granola, yogurt, and berries
D: Bread, cheese, and vegetables

What is the likely purpose or emphasis of the traditional design elements on the cigarette box?
A: To convey modernity
B: To highlight the product's origin
C: To attract younger consumers
D: To emphasize the product's quality

Defect Description

There is a defect in the object. What is the appearance of the defect?
A: The cap is slightly ajar on one side
B: The cap is completely missing
C: The bottle is dented
D: The label is peeling off

Defect Analysis

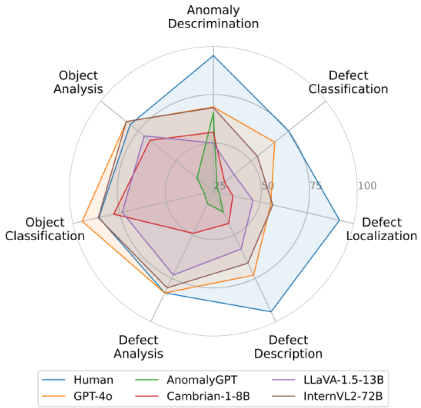


There is a defect in the object. What is the potential effect of the defect?
A: Reduced performance
B: Increased power consumption
C: Improper insertion
D: Shorter lifespan

MMAD Benchmark: The Reality Check (ICLR 2025) [5]

Two proposed boost methods:

- **Large Gap:** High accuracy on Object Classification ($> 90\%$), low on Defect Analysis ($< 50\%$).
- **Hallucination:** Models often invent defects that aren't there.



Results of 5 representative MLLMs and Human.

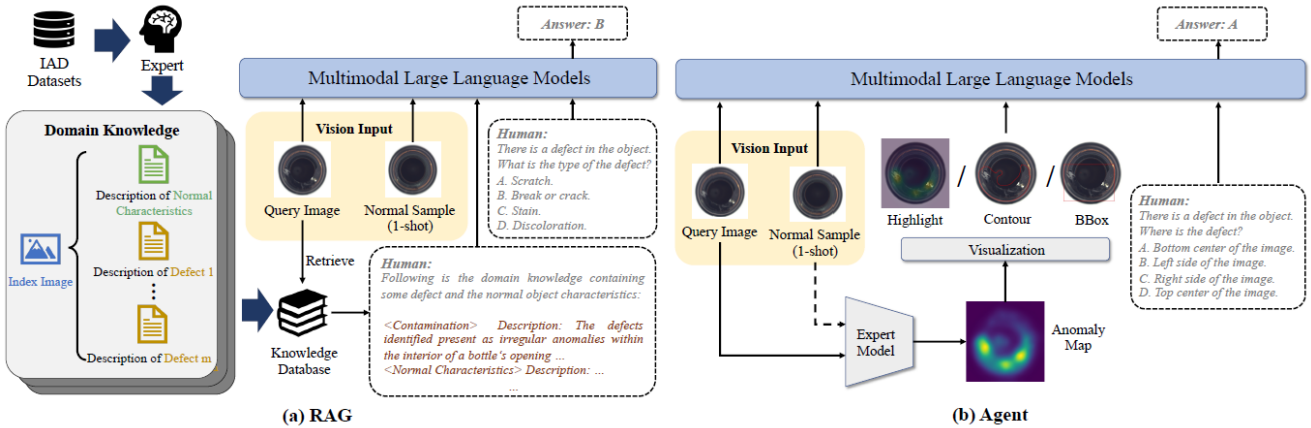


Illustration of two proposed boost methods.

Conclusion and Future Directions

Summary

- **Evolution:** Classical Processing $\rightarrow$ Semantic Understanding $\rightarrow$ Causal Reasoning.
- **VLMs** solved the *Cold Start* (Zero-Shot) problem.
- **MLLMs** are solving the Explainability problem.

Open Challenges

- **Latency:** MLLMs are too slow for 50ms production cycles.
- **Trust:** The most critical issue in MLLM-based VAD is Hallucination.
- **Agents:** Finally, the future VAD lies in Agentic AI.

References I

Kashiani et al., *ROADS: Robust Prompt-driven Multi-Class Anomaly Detection under Domain Shift*, WACV 2025.

Jeong et al., *WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation*, CVPR 2023.

Zhou et al., *AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection*, ICLR 2024.

Gu et al., *AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models*, AAAI 2024.

Jiang et al., *MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection*, ICLR 2025.