Suspect, Verify, and Conclude: Training-free Guidance for MLLMs for Industrial Anomaly Detection and Reasoning

Hosik Hwang*, Dohoon Kim*, Sungyong Baik
Deparment of Data Science, Hanyang University

Introduction

Problem Statement

Visual Anomaly Detection (VAD) identifies abnormal patterns in images, critical for manufacturing, medical diagnosis, and security.

Challenge: Existing MLLM-based methods struggle to balance three requirements simultaneously:

❶ **Accurate classification**
❷ **Precise localization**
❸ **Interpretable explanations**

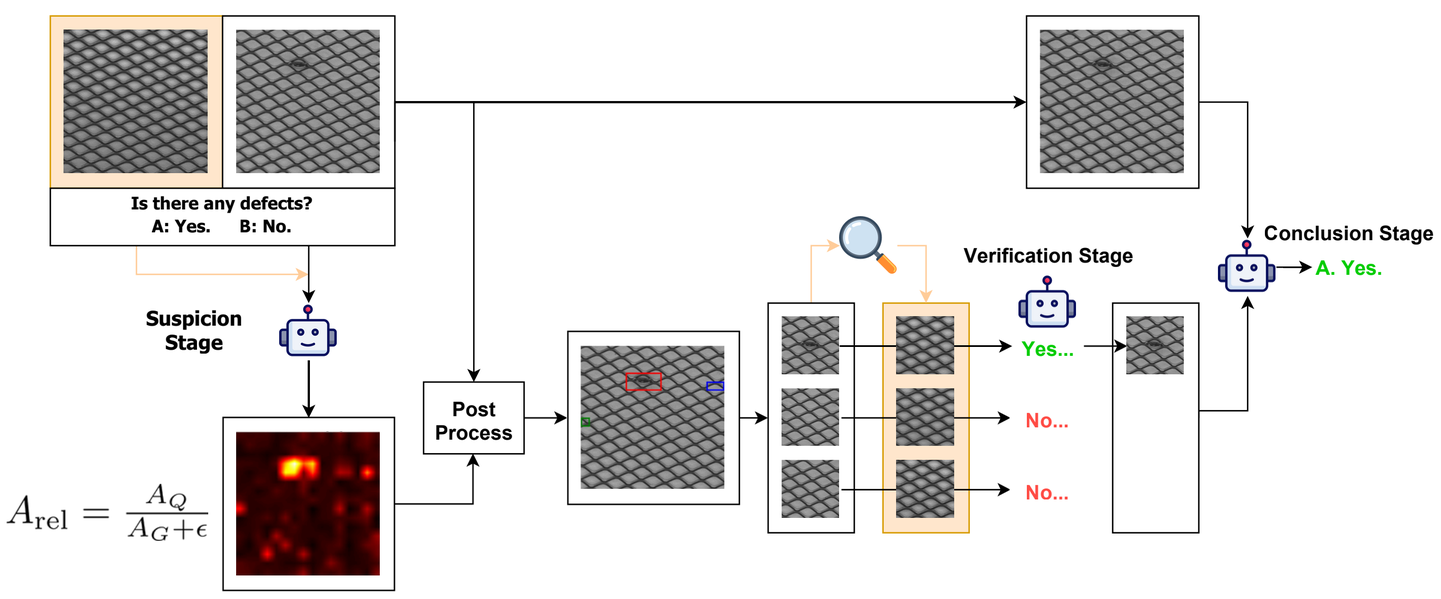
Gap: Most methods excel at ONE but fail to balance all three.

Key Idea: Bridging the Alignment Gap in MLLMs

▶ **Insight 1: The Attention-Generation Discrepancy**
- **Observation:** MLLMs often look at the correct defect location,
- even when they verbally hallucinate "Normal."
- **Strategy:** Decouple the roles: Use Relative Attention for localization and Generative Reasoning for the final decision.

▶ **Insight 2: Context-Driven Grounding**
- **Observation:** Vague zero-shot prompts lack specific industrial criteria, leading to errors.
- **Strategy:** Inject explicit Normal Standards (Knowledge & Reference Images) to ground the model's reasoning in visual evidence.

SVC-AD: Unified Pipeline



SVC-AD: A Unified 4-Stage Pipeline

Stage 1: Knowledge Condensation
- Extracts essential normal characteristics
- from domain knowledge.

Stage 2: Suspicion
- Identifies anomalous regions via Relative Attention.

Stage 3: Verification
- Validates candidates against normal reference crops.

Stage 4: Conclusion
- Generates final decisions and
- structured explanations.

Results

Evaluation on **MMAD** Dataset, MLLM Benchmark for Anomaly Detection.
All experiments are conducted in a 1-shot setting. We extract internal attention from layer 20 for Qwen and 14 for LLaVA, for the suspicion stage.

Model	Scale	Shot	Method	Anomaly	Defect				Object		Average
				Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
LLaVA-1.5	7B	One-shot	Baseline	44.55	43.32	34.26	55.35	72.88	**73.60**	65.38	56.46
			SVC-AD (Ours)	**58.58**	**54.32**	**44.48**	**59.21**	**74.31**	72.28	**66.89**	**61.44**
Qwen2.5-VL	7B	One-shot	Baseline	68.87	55.48	60.60	66.52	78.99	**93.02**	90.18	72.45
			SVC-AD (Ours)	**72.75**	**62.59**	**62.37**	**67.35**	**84.54**	90.26	**94.16**	**76.29**

Conclusion

Key Contributions

▶ ***Relative Attention Mechanism***
Isolates anomaly-relevant regions by contrasting task-aware vs. generic attention

▶ ***Retrieval-Based Verification***
Validates suspects against normal reference samples from memory bank

▶ ***Architecture-Agnostic Framework***
Works on LLaVA-1.5 (linear projection) and Qwen2.5-VL (complex adapter)

Practical Advantages

✓ No retraining needed when production specs change

✓ Works with 1-shot - minimal data requirements

✓ Interpretable - natural language explanations for human-AI collaboration

Future Work

▶ ***Diverse MLLM Model Experiments***
Various models like InternVL, LLaVA series (LLaVA-NeXT, LLaVA-OV, etc.)

▶ ***Real-Time Optimization***
Attention caching, prompt compression, parallelization for edge deployment

▶ ***Hyperparameter tuning for performance improvement***
Adaptive selection of layer, thresholds, etc.

▶ ***Domain Expansion***
Medical imaging, security surveillance, agricultural quality assessment