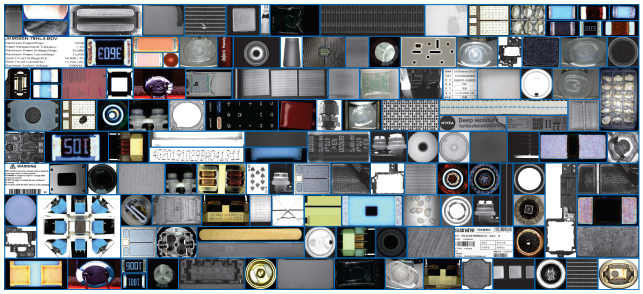
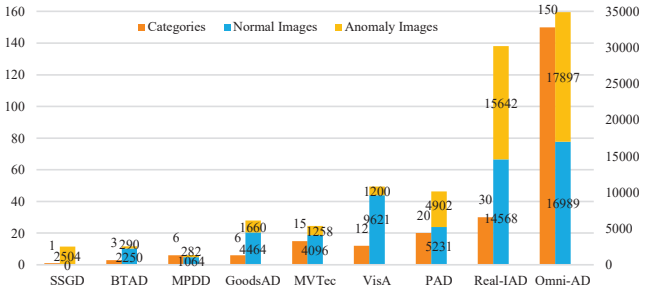
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Omni-AD: A Large-scale and Versatile Benchmark for Industrial Anomaly Detection

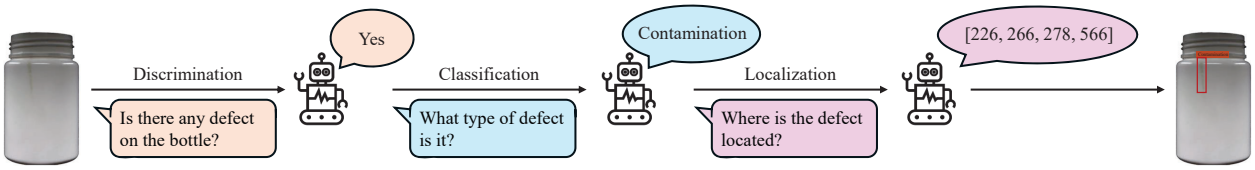
Dahu Shi[1,2] Chengshen He[3] Shaochen Zhang[4] Bo Qian[3]
Xiaochen Quan[2] Wencong Zhang[2*] Xing Wei[3]
[1]CCAI, Zhejiang University [2]Hikrobot Co., Ltd. [3]Xi'an Jiaotong University
[4]The Chinese University of HongKong, Shenzhen



(a) Visual overview of the proposed Omni-AD dataset.



(b) Quantitative comparison of popular IAD datasets.



(c) The defined subtasks for assessing anomaly detection abilities of MLLMs, *i.e.*, discrimination, classification, and localization.

Figure 1. **Overview of the proposed Omni-AD benchmark.** (a) presents a gallery of samples from the proposed Omni-AD dataset. (b) provides a comparative analysis of popular IAD datasets in terms of category diversity and image scale. (c) depicts three well-designed subtasks for evaluating the performance of MLLMs in industrial anomaly detection.

Abstract

Industrial Anomaly Detection (IAD) has attracted significant attention and witnessed rapid development. However, the advancement in this field is hindered by two key issues: the performance saturation of existing benchmarks, limiting discriminative evaluation of different IAD methods, and the absence of benchmarks tailored to assess recent multi-modal large language models (MLLMs) in anomaly detection. To this end, we present Omni-AD, a comprehensive IAD benchmark featuring: i) ***Large scale****: The dataset consists of approximately 35K images (6× larger than MVTec AD) with 150 product categories (10× larger than MVTec AD) spanning 16 industrial sectors, delivering unprecedented category diversity and greater dataset scale compared with existing datasets. ii)* ***Versatility****: The benchmark supports both conventional unsupervised and emerging MLLM-based IAD evaluation protocols. The latter is achieved by defining three subtasks of progressive difficulty, with two structured as visual question answering (VQA) and one as visual grounding. iii)* ***Challenge****: Extensive experimental results of state-of-the-art methods reveal that the Omni-AD benchmark is more challenging than existing benchmarks, which can drive the future development of the IAD field.*

1. Introduction

Industrial anomaly detection (IAD) is a critical computer vision task focused on identifying and localizing manufacturing defects, thereby enhancing both product quality and operational efficiency in production lines with stringent quality control. However, this task is inherently challenging due to its open-set nature, characterized by scarce defect data and unpredictable defect types.

*Corresponding author.
†Project page: https://omni-ad.github.io

In recent years, the mainstream anomaly detection algorithms [3, 10, 11, 16, 19, 23, 28, 35, 36, 38, 40, 41] have primarily adopted the unsupervised learning paradigm. These algorithms require training only on defect-free images, enabling the model to detect unknown defects while significantly reducing the costs associated with defect data collection and manual annotation. Although the IAD field has made significant progress under the unsupervised paradigm, it still faces two major obstacles. On the one hand, the performance of current top-performing algorithms on existing datasets tends to be saturated, making it impossible to distinguish the advantages of different algorithms. On the other hand, multimodal large language models (MLLMs) have demonstrated impressive capabilities across multiple domains, but their performance in the field of industrial anomaly detection has not been well studied. Therefore, this paper presents Omni-AD, a larger-scale and more challenging anomaly detection dataset, and natively supports the evaluation benchmark of MLLMs.

For the unsupervised IAD, the most widely used benchmark dataset, MVTec AD [5], has promoted the birth of a large number of algorithms [6–8, 12, 23, 25, 28]. However, the performance of state-of-the-art algorithms on this benchmark has approached saturation with technological progress. For instance, recent methods has achieved over 99% in both image-level and pixel-level AUROC metrics on the MVTec dataset, making it difficult to distinguish the advantages of different algorithms. As a consequence, new datasets like VisA [43] and Real-IAD [32] have been successively introduced. The existing datasets, however, are manually constructed and suffer from limited diversity in product and defect categories, which notably impedes their transfer from laboratory research to actual industrial scenarios. To bridge this gap, Omni-AD collects data from production lines with real-world defect inspection needs, encompassing 150 different product categories across 16 diverse application scenarios, totaling 34,886 images. As shown in Fig. 1a and Fig. 1b, our dataset offers notable advantages over previous datasets in terms of both scale and object diversity, posing substantial challenges to state-of-the-art unsupervised IAD methods. We conduct extensive experiments on Omni-AD alongside existing datasets using these approaches. The results demonstrate that Omni-AD is more challenging and enables a more discriminative evaluation of advanced methods.

For the evaluation of emerging MLLM-based IAD methods, MMAD [18] was the first to propose a multiple-choice VQA-style benchmark, providing a preliminary assessment of MLLMs' defect understanding capabilities. However, such multiple-choice VQA–style evaluation approach is not fully aligned with the practical requirements of IAD. In comparison, Omni-AD organizes defect detection into three progressively challenging subtasks: Defect Discrimination, Defect Classification, and Defect Localization. As shown in Fig. 1c, this benchmark directly evaluates MLLMs on three core questions relevant to real-world scenarios: "Is there a defect?", "What is the type of the defect?", and "Where is the defect located?". Specially, we formulate defect discrimination and defect classification as VQA tasks, and defect localization as a visual grounding task. To generate semantically rich QA pairs, we design a dedicated pipeline. Firstly, GPT-4o [15] is guided to enrich instance-level prior knowledge with defect-related information. Subsequently, knowledge from all instances of the same product is consolidated to produce product-level global information. By integrating instance- and product-level priors, GPT-4o acquires a comprehensive understanding of the product and its defects, which in turn allows it to generate high-quality QA pairs. Using these QA pairs, we conduct extensive evaluations of leading open-sourced MLLMs, including LLaVA-NeXT [20], InternVL-3.5 [33], and Qwen3-VL [2]. Experimental results show that current MLLMs exhibit modest capabilities in defect discrimination and classification, indicating their potential for industrial anomaly detection. However, MLLMs generally perform poorly on defect localization, indicating that precise localization remains a considerably more challenging task in IAD scenarios.

The main contributions of this work are summarized as follows:

- We introduce Omni-AD, a large-scale dataset that outperforms existing ones in both scale and diversity. It comprises approximately 35K pixel-wise annotated images from 150 product categories spanning 16 industrial sectors and presents a wide range of defect sizes, types, and aspect ratios.
- Toward a comprehensive benchmark for emerging MLLM-based IAD, we define three subtasks with increasing difficulty (*i.e.*, anomaly discrimination, classification, and localization) and construct a pipeline to create corresponding question-answer pairs with rich semantics.
- Extensive experimental results confirm that Omni-AD is more challenging than existing datasets, enabling better differentiation of IAD algorithms and revealing that current general-purpose MLLMs still have substantial room to improve in anomaly detection.

2. Related Work

2.1. Anomaly Detection Datasets

In the field of industrial anomaly detection, the evolution of benchmark datasets has fundamentally shaped methodological advancements. Early benchmarks like DAGM [26] and KolektorSDD [30] targeted specific industrial domains and are commonly used for supervised anomaly detection. The emergence of MVTec AD [5] marked a paradigm shift by establishing an unsupervised learning setting, where mod-

els are trained solely on normal samples. Comprising 5,354 images across 10 objects and 5 textures, this benchmark has driven advances in unsupervised anomaly detection. Following this, datasets such as BTAD [27] (2,540 images across 3 product classes) and MPDD [17] (1,346 images with 6 metal parts) are introduced, yet they remain limited in both category coverage and image volume. While VisA [43] and PKU-GoodsAD [39] represent steps toward larger-scale data, their category and industrial scenarios coverage remains constrained. Meanwhile, PAD [42] proposes Multi-pose Anomaly Detection (MAD) dataset and PIAD [34] extends PAD by further incorporating illumination variations, proposing a new dataset with synthetic and real samples to benchmark pose- and illumination-agnostic anomaly detection. More recently, Real-IAD [32] provides 150K multi-view industrial images spanning 30 categories, while MANTA [9] offers a large-scale multi-view visual-text dataset of 137K images across 38 categories specifically tailored for tiny objects. Despite these progresses, neither dataset is sourced from actual production lines, limiting their ability to fully represent the complexity and variability of real-world industrial inspection scenarios.

In contrast, the proposed Omni-AD dataset is collected from actual industrial production lines and covers 150 product categories across 16 industrial sectors, which expands the coverage of industrial scenarios. This dataset bridges this gap and provides more challenging and realistic benchmarks for the development of advanced IAD algorithms.

2.2. MLLMs in Industrial Anomaly Detection

MLLMs have recently gained significant traction, yet most existing benchmarks are not tailored to industrial anomaly detection. Current benchmarks predominantly focus on single-image processing scenarios (e.g., MMBench [22], MMVP [31]) while prioritizing general-purpose capabilities over domain-specific industrial needs. Specialized visual benchmarks exhibit similar limitations: TextVQA [29] emphasizes textual recognition in images, while MATH-Vision and MATHVISTA [24] target mathematical reasoning in visual contexts. Even cross-domain benchmarks like MMMU [37] conspicuously exclude industrial applications, contrasting with the medical domain where dedicated MLLM benchmarks (VQAMed [4], PathVQA [14]) have emerged.

MMAD [18] pioneered a comprehensive and quantitative evaluation of MLLMs for industrial anomaly detection. It defines seven subtasks for defect and object analysis, generating 39,672 semantically rich multiple-choice questions from four public datasets. While MMAD provides a preliminary assessment of MLLMs' understanding of defect detection, its task formulation does not fully align with real-world industrial requirements. Therefore, the proposed Omni-AD benchmark assesses MLLMs on three progressively challenging defect understanding subtasks: Defect Discrimination, Defect Classification, and Defect Localization, providing a more practical and direct evaluation of their capabilities in IAD.

Dataset	#Category	#Image			Seg. Label
		Normal	Abnorm	All	
SSGD [13]	1	0	2,504	2,504	✗
BTAD [27]	3	2250	290	2,540	✓
MPDD [17]	6	1,064	282	1,346	✗
MVTec AD [5]	15	4,096	1,258	5,354	✓
VisA [43]	12	9,621	1,200	10,821	✓
GoodsAD [39]	6	4,464	1,660	6,124	✓
PAD [42]	20	5,231	4,902	10,133	✓
PIAD [34]	30	7,102	4,166	11,268	✓
Real-IAD† [32]	30	14,568	15,642	30,210	✓
Omni-AD	**150**	**16,989**	**17,897**	**34,886**	✓

Table 1. **Comparison with mainstream IAD datasets.** †: compared with the single-view Real-IAD dataset to ensure that each image corresponds to a distinct object instance and unique defects.

3. Omni-AD Dataset

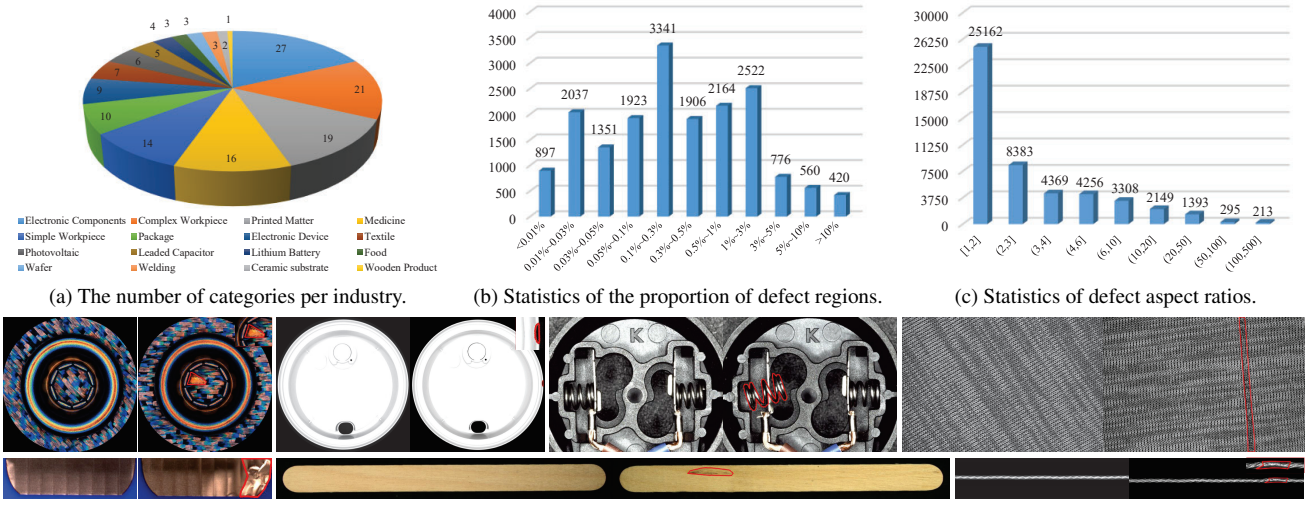
In this section, we introduce details of the Omni-AD dataset on data collection (Sec. 3.1), data annotation (Sec. 3.2) and statistics (Sec. 3.3).

3.1. Data Collection

We collected real-world data from various production lines, containing approximately 35K high-resolution normal and abnormal images. As shown in Fig. 2a, there are a total of 150 product categories from a wide range of 16 industries, covering various materials such as metal, plastic, ceramic, paper, fabric, wood, *etc*.

Object Alignment. Since the original image may contain irrelevant background, we then use template matching algorithm to align the foreground object as much as possible for rigid product categories. In addition, we use image cropping to make the RoI (region-of-interest) located near the center of the image for deformable and texture product categories. This process weakens the negative impact of background interference and object misalignment on model performance, especially for reconstruction-based anomaly detection algorithms [8], thereby making the performance comparison of various algorithms fairer.

Anomaly Definition. Unlike manually generated defect data [5, 32, 42, 43], real-world data does not have predefined defect types. In other words, the types and locations of anomalies are unknown before annotation. Therefore, we define a unique set of defect categories A_{cat} for each product category, based on the manufacturer's description of the product's usage and quality requirements. Subsequently, we



(a) The number of categories per industry. (b) Statistics of the proportion of defect regions. (c) Statistics of defect aspect ratios.

(d) Visualization of representative defects.

Figure 2. **Statistics of the proposed Omni-AD Dataset.** (a) demonstrate that the Omni-AD dataset covers a broad spectrum of 16 industries, including 150 product categories in total. (b) shows statistics of the proportion of defective regions occupying the entire image. (c) displays statistics of the aspect ratio of the minimum bounding box of the defect. (d) visualizes several pairs of normal and abnormal images, with small defects highlighted in the close-up figure in the upper right corner.

manually assign the collected data into normal images and anomaly images according to the defined defect types.

3.2. Data Annotation

Annotating fine-grained anomalies across a wide array of industrial scenarios is not a trivial task. To ensure high-quality and consistent annotations, we design a cyclic two-stage annotation pipeline to perform pixel-wise defect annotation.

Human-Led Annotation. In the first stage, a team of professional annotators was tasked with identifying defect types and outlining their polygonal masks. To ensure quality, we established a control set of 500 meticulously annotated images. All annotators were required to pass a qualification test on this set, achieving a discrepancy rate below a strict threshold before beginning formal annotation. The team was divided into annotators and inspectors. Images for each product category were evenly split into three jobs and annotated sequentially. Upon completion of a job, an inspector reviewed the annotations. Jobs containing any errors were rejected and returned to the original annotator for correction. This review process served to correct individual mistakes and enforce consistent application of the annotation rules across all images of the current product category.

Model-Assisted Annotation. In the second stage, the manually-labeled data for each product category are split evenly into two subsets, each containing equal numbers of normal and anomalous images. We then train a supervised semantic segmentation model (*e.g.*, U-Net) to identify potential annotation errors. The assigned annotator examines discrepancies between the model's predictions and the manual labels. These inconsistent regions are manually reviewed and corrected. This process of model-assisted review and manual correction is repeated until the discrepancy rate falls below a predefined threshold, resulting in a refined and high-quality dataset.

3.3. Statistics

Through the stringent annotation pipeline described above, we constructed the Omni-AD dataset. It contains 34,886 images from 150 product categories, with each category exhibiting 1 to 7 distinct defect types, and includes 16,989 normal and 17,897 anomalous samples. Fig. 1 shows examples of each product category and visualizes the contour of defect regions with red lines.

We present a detailed comparison with the mainstream datasets, as shown in Tab. 1. The number of product categories in our dataset surpasses those in the mainstream MVTec AD [5] and VisA [43] datasets by an order of magnitude and a factor of 12.5, respectively (*i.e.*, 150 *vs*. 15 and 12). Furthermore, the total number of images in our dataset is approximately 6 times that of MVTec AD and 3 times that of VisA. Compared with the recent Real-IAD [32] dataset, our dataset outperforms its single-view counterpart in both the variety of product categories and the volume of images. As shown in Fig. 2b, the proportion of defect regions is very wide, indicating that the dataset contains both very small and large defects. As shown in Fig. 2c, the defect aspect ratios in our dataset exhibit a wide range, demonstrating the diversity of defect morphology. The ratio of normal to ab-

normal images in our dataset is approximately 1:1 to ensure that each product category has abundant defects, which is better than the MVTec AD and VisA datasets and comparable to the Real-IAD dataset. Several representative normal-abnormal image pairs are shown in Fig. 2d. To summarize, the proposed Omni-AD dataset offers significant advantages over existing datasets, notably in its scale (number of images and product categories), broader industry distribution, and greater defect diversity. The comprehensive evaluation results (described in Sec. 5) show that our dataset poses great challenges to existing anomaly detection methods and can drive the development of the IAD field.

4. Benchmark

Building upon the proposed Omni-AD dataset, we establish two evaluation protocols, including the conventional Unsupervised IAD and the newly-designed MLLM-Based IAD, to encourage the community to evaluate and compare their methods.

4.1. Unsupervised IAD Protocol

The conventional unsupervised IAD protocol is a well-studied anomaly detection setting, where models are trained exclusively on normal samples and subsequently evaluated on both normal and abnormal samples. Following the common practices illustrated in MVTec AD [5], VisA [43], and Real-IAD [32] datasets, we assign normal images to both the training and testing sets, while reserving all anomalous images exclusively for the testing set.

Metrics. We utilize commonly-used evaluation metrics, as summarized in the survey [21], to evaluate the performance of unsupervised anomaly detection methods, For anomaly classification, image-level Area Under the Receiver Operating Characteristic (*i.e.*, I-AUROC) and Area Under the Precision-Recall curve (*i.e.*, I-AUPR) are adopted. For anomaly segmentation, we employ pixel-level AUROC (*i.e.*, P-AUROC) as well as Area Under Per-Region-Overlap curve (*i.e.*, P-AUPRO) to better account for varying defect sizes and the imbalance issue between normal pixels and abnormal pixels.

4.2. MLLM-Based IAD Protocol

With the rapid advancement of MLLMs, these models have exhibited impressive general capabilities and the potential to assist or even replace humans in a variety of tasks, sparking paradigm shifts across multiple domains. A pivotal question arises for the IAD community: Can MLLMs excel in this field? The potential to provide a general-purpose solution stands in stark contrast to conventional approaches, which are inherently limited by their need to be tailored to specific scenes, products, or defect types, thus suffering from poor generalization. To thoroughly assess the anomaly detection abilities of MLLMs, we propose three subtasks on the Omni-AD dataset: Defect Discrimination, Defect Classification, and Defect Localization. These subtasks, in ascending order of difficulty, are closely aligned with practical industrial inspection demands.

4.2.1. Task Definitions

In particular, Defect Discrimination and Defect Classification are cast as VQA tasks, while Defect Localization is formulated as a visual grounding paradigm. These three subtasks are described as follows:

Defect Discrimination. Given an image or a specific bounding box, together with a question enriched with product- and defect-level semantic context, the model must decide whether the corresponding content is anomalous. The response is constrained to a binary output: *Yes* or *No*.

Defect Classification. Given an input image or a specific bounding box, together with a question, the model is presented with several candidate options and is required to select the corresponding defect type. The ground truth answer should be the *specific anomaly type* or *Not any*.

Defect Localization. For the Defect Localization task, given a question, the model is required to produce a JSON-formatted output containing the precise location and category of the detected anomaly. The output must be structured as: {'box' : $[x_{min}, y_{min}, x_{max}, y_{max}]$, 'label':'*defect type*'}.

4.2.2. Semantically Rich QA Construction

In Omni-AD, each sample is annotated with structured metadata describing the product and its defects. However, this information, though well-formatted, lacks the semantic richness required for direct MLLM evaluation. To address this, we design a hierarchical data enrichment process (Fig. 3) built upon the MMAD pipeline [18], but enhanced with semantic guidance and hierarchical consolidation.

Prior Enrichment. To enrich each instance's prior knowledge, we construct the visual prompt comprising two components: the original image enhanced with edge-highlighted defect contours derived from its mask annotation, and a sampled normal image of the same product. This visual prompt explicitly conveys defect characteristics and supports contrastive analysis between defective and normal regions. Leveraging this prompt along with the instance's structured metadata, we then direct GPT-4o [15] to fill in missing semantic attributes—preventing it from generating irrelevant generic content due to its general-purpose training. The enrichment incorporates high-level semantics such as usage scenarios, functional context, defect morphology, formation mechanisms, and impact, resulting in *Enriched Instance Prior Knowledge*.

Knowledge Consolidation. For each product, the enriched priors from all its instances are aggregated. GPT-4o is then employed to distill this collection into consistent *Product Prior Knowledge* by filtering out instance-specific

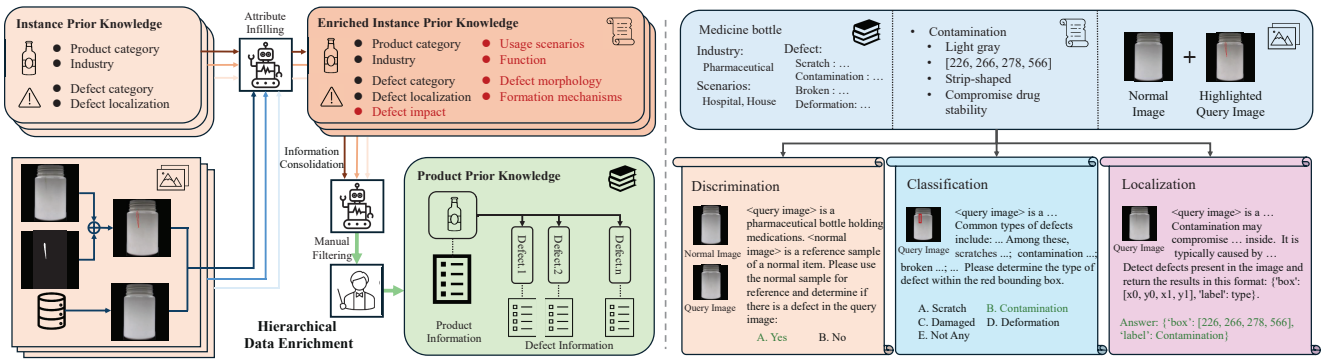


Figure 3. **Overview of the QA pair generation pipeline.** We leverage GPT-4o [15] to semantically enrich the defect data and summarize global prior knowledge for each product category. By incorporating both product-level and instance-level prior knowledge, GPT-4o is able to generate Question-Answer (QA) pairs with richer defect-related semantic information after comprehensively understanding the product and defect details.

variations such as positional or background noise. This process yields global knowledge characterizing the product and its typical defects, which is subsequently refined manually to ensure reliability.

QA Pairs Generation. The visual prompts, enriched instance priors, and consolidated product knowledge collectively form the contextual information that supports GPT-4o in generating Question-Answer (QA) pairs, enabling it to better recognize and understand the product and its corresponding defects. Subsequently, based on this contextual information along with task-specific definitions and formatting requirements, GPT-4o produces semantically rich QA pairs. Unlike simple templated questions, these questions generated through our proposed pipeline provide comprehensive descriptions of products and defects, serving as meaningful references for the evaluated MLLMs in their responses.

4.2.3. Evaluation Settings

Employing the aforementioned pipeline, we generate a total of 12K QA data items and 3K grounding data items. For each subtask, we conduct evaluations under two settings: zero-shot and one-shot. In the zero-shot scenario, the model makes predictions based only on the query image. In contrast, the one-shot setting allows the model to refer to the most similar sample from the normal set as additional visual guidance.

Metrics. For the Defect Discrimination and Defect Classification subtasks, we designed image-level and box-level questions: the former queries about the image as a whole, while the latter directs the model's attention to specific regions. For both types of tasks, we adopt VQA accuracy as the evaluation metric. For Defect Localization, we employ Recall, Precision, and the F1 score instead, as MLLMs typically lack reliable confidence estimation. These localization metrics are calculated by comparing predicted and ground-truth bounding boxes at an IoU threshold of 0.5, with the F1 score representing their harmonic mean.

5. Experiments

We conduct extensive experiments under two evaluation protocols, including the evaluation of conventional unsupervised IAD methods and the assessment of the anomaly inspection capabilities of up-to-date multi-modal large language models (MLLMs).

5.1. Evaluation on Unsupervised IAD

Comprehensive experiments are conducted to compare the performance of our Omni-AD and several popular IAD datasets, such as MVTec AD [5], VisA [43], and Real-IAD [32]. MVTec AD [5] is the most widely used benchmark dataset for unsupervised industrial anomaly detection. It contains ∼5K images over 10 objects and 5 texture categories with different types of defects, captured under highly controlled illumination conditions. VisA [43] is twice as large as MVTec AD in terms of the number of images, while comprising 12 object categories, slightly fewer than those in MVTec AD. It features a complex structure, multiple objects, and large variations in object locations. Real-IAD [32] consists of ∼30K images (single-view version) with 30 object categories. The number of images and object categories has increased compared to previous datasets. In comparison, the proposed Omni-AD in this paper comprises 150 product categories spanning a wide range of 16 industrial sectors, including ∼35K images. Our dataset exceeds all previous datasets in terms of both quantity and object diversity, posing great challenges to the current state-of-the-art methods.

Experimental Details. For performance comparison on these datasets, we select representative IAD methods from three dominant paradigms: embedding-based methods (*e.g.*, Padim [7], PatchCore [28]), reconstruction-based methods (*e.g.*, RD4AD [8], Dinomaly [12]) and

Datasets	MemoryBank-based		Reconstruction-based		Data-Aug-based	
	PatchCore [28]	PaDim [7]	RD4AD [8]	Dinomaly [12]	SimpleNet [23]	GLASS [6]
Image-Level Metrics (I-AUROC / I-AUPR)						
MVTec AD [5]	99.1 / 99.5	94.9 / 95.9	98.5 / 99.3	99.6 / 99.8	99.6 / 99.7	99.9 / 99.8
VisA [43]	94.7 / 93.0	85.1 / 86.2	96.0 / 96.9	98.7 / 98.9	97.1 / 97.6	98.8 / 99.4
Real-IAD [32]	90.4 / 93.6	86.6 / 89.8	87.6 / 93.2	89.3 / 86.8	91.7 / 93.8	92.3 / 94.1
Omni-AD	87.8 / 92.9	82.5 / 88.4	83.1 / 90.3	84.5 / 91.5	82.1 / 89.5	83.1 / 91.2
Pixel-Level Metrics (P-AUROC / P-AUPRO)						
MVTec AD [5]	98.1 / 93.5	97.3 / 92.7	97.8 / 93.9	98.4 / 94.8	98.1 / 91.0	99.3 / 96.8
VisA [43]	98.5 / 91.8	96.2 / 79.7	98.5 / 91.1	98.7 / 94.5	98.2 / 90.7	98.8 / 92.8
Real-IAD [32]	96.9 / 92.5	95.6 / 90.0	96.2 / 94.4	98.8 / 93.9	92.3 / 88.9	93.5 / 86.4
Omni-AD	95.1 / 81.1	93.9 / 81.6	92.0 / 80.1	93.0 / 71.4	90.4 / 76.7	85.1 / 74.8

Table 2. **Unsupervised Anomaly Detection and Localization: performance comparison on Omni-AD and mainstream IAD datasets.** Averaged results are reported over all categories on each dataset. A lower value indicates that the anomaly detection dataset is more challenging. The worst result is marked with a dark gray background, while the second-worst is marked with a light gray background.

data-augmentation-based methods (*e.g.*, SimpleNet [23], GLASS[6]). PatchCore [28] and PaDiM [7] both utilize pre-trained features but differ in modeling normality. PatchCore employs a memory bank of core-set patch embedding, whereas PaDiM uses a parametric Gaussian model. At test time, they similarly detect anomalies by computing patch-level distances from the established normal distribution. RD4AD [8] and Dinomaly [12] introduce specific architectural designs, *i.e.*, reverse distillation and novel transformer-oriented components, to mitigate the known issue of shortcut learning. Otherwise, anomalous samples also tend to be well reconstructed, resulting in the missed detection of abnormal areas. SimpleNet [23] and GLASS [6] reformulate the unsupervised IAD task into a supervised learning framework. This is achieved by training on real normal images with synthetic anomalous samples or features, enabling discriminative model training and inference. We run PatchCore [28], PaDim [7], RD4AD [8], and Dinomaly [12] with the open-sourced Anomalib [1]. As for SimpleNet [23] and GLASS [6], we adopt the official implementation. The input image are resized to 256×256. Other hyperparameters of each algorithm are all set to default values.

Results. The experimental results of all methods on MVTec AD [5], ViaA [43], Real-IAD [32] and our Omni-AD are shown in Tab. 2. On the one hand, it can be observed that both the image-level and pixel-level performance on our dataset exhibit noticeable performance degradation compared to the other three datasets. In particular, multi-class algorithms (*i.e.*, Dinomaly [12]) that claim performance comparable to single-class methods on MVTec AD exhibit a notable drop in performance on our dataset. This demonstrates that the Omni-AD dataset is more challenging than existing datasets due to its unprecedented diversity in product categories and defect types. On the other hand, the performance differences among various methods are more pronounced on the Omni-AD Dataset. For example, on MVTec AD, all methods achieve I-AUROC scores between 94.9%∼99.9%, and P-AUPRO scores between 91.0%∼96.8%. On Real-IAD, the ranges decrease to 86.6%∼92.3% and 86.4%∼94.4%, respectively. In contrast, on Omni-AD, I-AUROC and P-AUPRO span 82.1%∼87.8% and 71.4%∼81.6%, respectively, demonstrating a wider performance spread across methods.

5.2. Evaluation on MLLM-Based IAD

With the aim of providing a thorough and quantitative evaluation of MLLMs' defect detection abilities, we then assess several models across three defined subtasks: Defect Discrimination, Defect Classification, and Defect Localization.

Experimental Details. Specifically, we locally deploy several leading open-source models on Nvidia H20 GPUs, including LLaVA-NeXT [20], InternVL-3.5 [33], Qwen3-VL-Instruct [2] and Qwen3-VL-Thinking [2]. During evaluation, responses that fail to provide a valid choice among the given options in the VQA-based defect discrimination and classification tasks are directly judged as incorrect. For the defect localization task in the form of visual grounding, if the model fails to produce an output in the required JSON format containing defect category and bounding box information, it is considered as a missed detection.

Results. The results are shown in Tab. 3. Firstly, the performance of these models decrease successively on the three subtasks with increasing difficulty. All models consistently demonstrate stronger discrimination and classification capabilities at the box level than at the image level. This enhancement in performance suggests that these MLLMs can extract more meaningful features and make more reli-

Model	#Param	Setting	Discrimination		Classification		Localization		
			Img-Acc	Box-Acc	Img-Acc	Box-Acc	Recall	Precision	F1
Random Output	-	-	50.00	50.00	20.00	20.00	N/A	N/A	N/A
Human Expert	-	-	95.67	99.33	91.67	96.00	75.65	83.37	79.32
LLaVA-NeXT [20]	34B	0-shot	59.43	55.16	38.84	37.70	N/A	N/A	N/A
		1-shot	59.17	55.29	34.80	34.31	N/A	N/A	N/A
InternVL-3.5 [33]	30B-A3B	0-shot	52.23	61.94	40.44	48.06	9.57	10.66	10.08
		1-shot	58.61	68.49	44.05	49.30	8.82	8.68	8.75
	38B	0-shot	57.58	66.31	40.70	50.67	10.81	9.43	10.07
		1-shot	57.54	67.61	43.96	51.97	8.68	7.14	7.84
Qwen3-VL-Instruct [2]	8B	0-shot	61.81	68.66	59.60	62.37	19.66	14.56	16.73
		1-shot	64.40	71.81	62.37	65.95	14.60	13.99	14.29
	30B-A3B	0-shot	62.38	69.09	62.22	67.89	19.59	18.67	19.12
		1-shot	<u>64.65</u>	<u>72.19</u>	**65.94**	**70.97**	17.31	13.03	14.84
Qwen3-VL-Thinking [2]	8B	0-shot	57.86	67.09	54.65	59.02	18.98	18.06	18.51
		1-shot	63.05	69.85	57.14	58.80	17.96	18.54	18.25
	30B-A3B	0-shot	62.85	69.05	57.22	65.95	**29.56**	**23.53**	**26.20**
		1-shot	**69.38**	**73.32**	<u>64.29</u>	<u>69.31</u>	<u>28.54</u>	<u>23.29</u>	<u>25.64</u>

Table 3. **MLLM-Based Anomaly Discrimination, Classification, and Localization: performance comparison of different MLLMs on the Omni-AD benchmark.** Each subtask is evaluated under 0-shot and 1-shot settings. Img-Acc denotes image-level accuracy, where the entire query image is fed into the model. Box-Acc denotes box-level accuracy, where a specific region of the query image is highlighted before being fed into the model. For each metric, the best is in **bold** while the second-highest is <u>underlined</u>. N/A: Not Available.

able judgments when focusing on a specific region of interest. On the most challenging localization subtask, the evaluated models struggle with visual grounding, achieving the lowest scores (*e.g.*, the best-performing model, Qwen3-VL-30B-A3B-Thinking, scores only 26.20 F1). LLaVA-NeXT fails to generate responses in the required JSON format, and thus its results are marked as N/A. Overall, all evaluated models surpass the random baseline, validating that current general-purpose MLLMs possess a modest ability to detect anomalies. Secondly, Qwen3-VL-30B-A3B-Instruct and Qwen3-VL-30B-A3B-Thinking demonstrate superior overall performance among all evaluated models, outperforming LLaVA-NeXT and InternVL-3.5 across all metrics. Notably, in the defect localization task, the Qwen3-VL-30B-A3B-Thinking model outperforms its Instruct counterpart (F1 score: 26.20 *vs*. 19.12). This noticeable advantage demonstrates that MLLMs incorporating a thinking mode possess stronger visual grounding capabilities, highlighting the promise of thinking-enabled MLLMs for future development. Thirdly, most models show clear gains in defect discrimination and classification when moving from 0-shot to 1-shot setting, demonstrating their ability to perform pairwise comparisons between the query image and normal template image. In contrast, most methods exhibit a slight performance drop on the more challenging localization subtask, highlighting its unique challenge.

Human Evaluation. We also conduct a human evaluation using 100 examples randomly sampled from each subtask. Three industrial anomaly detection researchers are employed as human experts and the averaged performance are reported. As shown in Tab. 3, humans substantially outperform MLLMs across all seven metrics. This significant gap not only underscores the inherent challenges of the Omni-AD benchmark but also highlights the considerable room for improvement in MLLM-based anomaly detection.

6. Conclusion

In this work, we collect and meticulously annotate a large-scale IAD dataset, termed Omni-AD. Based on this dataset, we evaluate state-of-the-art IAD methods in real-world industrial quality inspection. The relatively low metrics indicate that these methods still fall short of practical requirements, consistent with the gap between strong performance on existing datasets and inferior performance in real-world applications. Additionally, to reveal the potential of current MLLMs in industrial quality inspection, we generate 12K VQA data and 3K visual grounding data to benchmark their capabilities in terms of defect discrimination, classification, and localization. Comprehensive results show that MLLM-based anomaly detection still has substantial room for improvement. We believe Omni-AD can advance the development of IAD.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62572385.

References

[1] Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A deep learning library for anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1706–1710. IEEE, 2022. 7

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 7, 8

[3] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 128–138, 2024. 2

[4] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, Bucharest, Romania, 2021. CEUR-WS.org. 3

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 2, 3, 4, 5, 6, 7

[6] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *European Conference on Computer Vision*, pages 37–54. Springer, 2024. 2, 7

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer, 2021. 6, 7

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022. 2, 3, 6, 7

[9] Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25518–25527, 2025. 3

[10] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107, 2022. 2

[11] Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6447–6458, 2023. 2

[12] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20405–20415, 2025. 2, 6, 7

[13] Haonan Han, Rui Yang, Shuyan Li, Runze Hu, and Xiu Li. Ssgd: A smartphone screen glass dataset for defect detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[14] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 3

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 5, 6

[16] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 2

[17] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021. 3

[18] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: The comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5

[19] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022. 2

[20] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Im-

proved reasoning, ocr, and world knowledge, 2024. 2, 7, 8

[21] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024. 5

[22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 3

[23] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20402–20411, 2023. 2, 7

[24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3

[25] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9974–9983, 2025. 2

[26] Fred. A. Hamprecht Matthias Wieler, Tobias Hahn. Weakly supervised learning for industrial optical inspection. https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection, 2007. 2

[27] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 3

[28] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022. 2, 6, 7

[29] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[30] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020. 2

[31] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 3

[32] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. 2, 3, 4, 5, 6, 7

[33] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 7, 8

[34] Kaichen Yang, Junjie Cao, Zeyu Bai, Zhixun Su, and Andrea Tagliasacchi. Piad: Pose and illumination agnostic anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4734–4743, 2025. 3

[35] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 2

[36] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 2

[37] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 3

[38] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2

[39] Jian Zhang, Runwei Ding, Miaoju Ban, and Linhui Dai. Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation. *IEEE Robotics and Automation Letters*, 9(3):2008–2015, 2024. 3

[40] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 2

[41] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16699–16708, 2024. 2

[42] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection. *Advances in Neural Information Processing Systems*, 36:44558–44571, 2023. 3

[43] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022. 2, 3, 4, 5, 6, 7