



# Triad: Empowering LMM-based Anomaly Detection with Expert-guided Region-of-Interest Tokenizer and Manufacturing Process

Yuanze Li<sup>1†</sup> Shihao Yuan<sup>1,2†</sup> Haolin Wang<sup>1</sup> Qizhang Li<sup>1,2</sup>  
Ming Liu<sup>1(✉)</sup> Chen Xu<sup>2</sup> Guangming Shi<sup>2</sup> Wangmeng Zuo<sup>1,2</sup>

sqliyz@hit.edu.cn, csshiahao@outlook.com, why\_cs@outlook.com, csqizhang@gmail.com  
csmliu@outlook.com, xc.xc@qq.com, gmshi@xidian.edu.cn, wzmzuo@hit.edu.cn

<sup>1</sup>Harbin Institute of Technology, <sup>2</sup>Pengcheng Lab, Guangzhou

## Abstract

Although recent methods have tried to introduce large multimodal models (LMMs) into industrial anomaly detection (IAD), their generalization in the IAD field is far inferior to that for general purposes. We summarize the main reasons for this gap into two aspects. On one hand, general-purpose LMMs lack cognition of defects in the visual modality, thereby failing to sufficiently focus on defect areas. Therefore, we propose to modify the AnyRes structure of the LLaVA model, providing the potential anomalous areas identified by existing IAD models to the LMMs. On the other hand, existing methods mainly focus on identifying defects by learning defect patterns or comparing with normal samples, yet they fall short of understanding the causes of these defects. Considering that the generation of defects is closely related to the manufacturing process, we propose a manufacturing-driven IAD paradigm. An instruction-tuning dataset for IAD (InstructIAD) and a data organization approach for Chain-of-Thought with manufacturing (CoT-M) are designed to leverage the manufacturing process for IAD. Based on the above two modifications, we present **Triad**, a novel LMM-based method incorporating an expert-guided region-of-interest tokenizer and manufacturing process for industrial anomaly detection. Extensive experiments show that our Triad not only demonstrates competitive performance against current LMMs but also achieves further improved accuracy when equipped with manufacturing processes. Source code, training data, and pre-trained models will be publicly available at <https://github.com/tzjtatata/Triad>.

## 1. Introduction

Industrial Anomaly Detection (IAD) plays a crucial role in modern manufacturing, where continuous iteration of product design and diverse inspection criteria call for robust and

adaptable inspection solutions. Although large multimodal models (LMMs) have recently shown remarkable success on general-purpose tasks, their performance in IAD remains noticeably below expectations. We attribute this discrepancy to two key factors.

First, current LMMs are typically trained to align vision and language modalities for broad semantic understanding, yet they lack the specialized ability to identify and focus on potential defect regions. Industrial defects often appear subtly against complex backgrounds or in close proximity to functional components, making a purely global, coarse alignment insufficient. To address this challenge, we propose an expert-guided region-of-interest tokenizer (EG-RoI). Inspired by AnyRes module in LLaVA [16, 22, 23], EG-RoI extracts high-resolution and potentially anomalous regions of interest identified by existing IAD methods. By explicitly highlighting anomalous and normal regions during training, we enable the model to learn targeted visual cues indicative of attributes of products and defects. During inference, the high resolution suspicious regions predicted by vision experts further serve as key references for more precise anomaly detection.

Second, prevailing IAD approaches treat defects as isolated visual phenomena, neglecting their intrinsic relationship to manufacturing (MFG) workflows. In reality, anomalies arise from process deviations (material impurities, assembly errors, or equipment malfunctions) that propagate through production stages. Existing methods, which rely on image-level comparisons or defect taxonomies, fail to leverage this causal knowledge. To bridge this gap, we introduce a manufacturing-driven industry anomaly detection paradigm that weaves relevant manufacturing processes into the reasoning process. Our framework incorporates: (1) An instruction-tuning dataset with attribute-rich captions, InstructIAD, spanning a variety of products and defect types from existing IAD datasets; (2) Chain-of-Thought with Manufacturing (CoT-M), a data organi-



Figure 1. Main workflow of Triad. As shown in the figure, the query image is first passed through EG-RoI with the suspicious regions predicted by the vision expert. These regions are cropped to keep a higher resolution than the query image and then encoded along with it. The text input includes a basic prompt, the manufacturing process for the specific product (this kind of cable for this case) and a question about anomaly detection. In this case, Triad finds out that the issue is the wrong insulation color, as the manufacturing process mentioned that the cable is composed of three different color wires.

zation strategy that synthesizes defect causality by linking anomalies to specific manufacturing steps. CoT-M expands InstructIAD by (i) editing normal product captions at the attribute level to simulate product evolution and defect occurrence and (ii) generating Chain-of-Thought processes grounded in manufacturing steps via GPT. When captions are unavailable, CoT-M employs a checklist-style template mirroring human inspector workflows. Ultimately, we propose **Triad**, a novel LMM-based method incorporating an expert-guided region-of-interest tokenizer and manufacturing process for industrial anomaly detection. We evaluate Triad on three benchmarks, MVTec-AD [2], WFDD [6], and PCB-Bank [33], and assess results under 0-/1-shot settings. Experiments show that Triad-ov-7B not only achieves competitive performance against both general-purpose and domain-specific LMMs but also demonstrates further improvements when accounting for manufacturing processes. Notably, with a single reference image (1-shot), Triad-ov-7B achieves 94.1% on MVTec-AD, surpassing Qwen2-VL-72B [31] by 2.1%. Extensive analyses show the ability of Triad to comprehend complex manufacturing workflows and extend to unseen processes.

In summary, the contributions of this paper include,

- Triad, a novel LMM equipped with an expert-guided region-of-interest tokenizer and manufacturing-aware Chain-of-Thought reasoning, achieving state-of-the-art 0-/1-shot anomaly detection performance.
- A data organization strategy CoT-M and a human-annotated instruction-tuning dataset InstructIAD, augmenting LMMs with causal defect understanding.
- Comprehensive zero-/one-shot benchmarks showing Triad’s superiority over general-purpose and domain-specific LMMs, with analyses demonstrating its ability to generalize to unseen manufacturing processes.

## 2. Related Works

### 2.1. Industrial Anomaly Detection

Traditional unsupervised IAD methods typically fall into two categories. Reconstruction-based methods [20, 33, 36, 39] regenerate defect-free images to highlight deviations, while feature-embedding methods [14, 25, 28, 29] compare query embeddings against normal feature patterns. Although effective, these methods usually require a dedicated model for each product, limiting their adaptability in dynamic production settings.

Recent works have leveraged vision-language models for zero-shot and few-shot anomaly detection by aligning visual features with textual prompts. For instance, WinClip [12] measures similarity using hand-crafted text prompts, and later methods [7, 8] enhance these prompts with visual cues or learned representations. However, the limited reasoning capabilities of their base models still present challenges. Our work addresses this gap by introducing long CoT answers during training to fully harness the reasoning power of large multimodal models.

### 2.2. Large Multimodal Models for IAD

Recent progress in LMMs has led to significant advances across a variety of vision-language tasks, driven by improvements in visual-encoder architectures such as Q-Former [17], adaptive visual encoding [34], and spatial-aware visual sampling [38]. Several IAD-centric methods attempt to incorporate such models by supplying textual guidance. For instance, Customizable-VLM [32] tailors off-the-shelf LMMs to anomaly detection using class-specific prompts and normality descriptions, while MMAD [13] employs a defect-aware strategy to integrate a database of normal product features and defect descriptions. Although these textual cues enhance performance, they still fall short

of capturing the underlying interactions among manufacturing steps that directly lead to defects.

The most closely related efforts include AnomalyGPT [10] and Myriad [19], which fine-tunes LMMs to the IAD domain. AnomalyGPT uses an image decoder to predict pixel-level anomaly maps and then applies a prompt learner to guide large language models, but its reliance on coarse-level image tokens and predicted masks can lead to loss of fine-grained details. Myriad leverages pretrained IAD models as vision experts to resample fine-grained features via Q-Former [17], effectively transferring general-purpose visual perception into the IAD domain. However, their reliance on simulated anomaly data and limited vision-language alignment undermines their ability to capture fine-grained visual attributes and understand manufacturing processes, ultimately leading to suboptimal performance. To bridge these gaps, our work, Triad, first enhances attribute perception by introducing the CoT-M and EG-RoI, allowing manufacturing-driven anomaly detection.

### 3. Methods

#### 3.1. Problem definition

Manufacturing-driven industry anomaly detection aims to determine whether an industrial product exhibits anomalies by jointly analyzing visual features and detailed product-specific manufacturing process descriptions. Given an input image and a textual description of the manufacturing process, the task requires the model to output a binary decision (normal or anomalous).

While standard IAD tasks (e.g., [10, 19]) primarily focus on visual cues, manufacturing-driven IAD leverages intricate manufacturing process details, such as multi-step production procedures, to support a deeper analysis. This paradigm demands accurate visual attribute recognition and sophisticated reasoning that accounts for the interplay between visual cues and contextual manufacturing information, allowing more precise detection of defects arising from complex process deviations.

#### 3.2. Training Objectives

Based on the problem defined before, our input, noted as  $X$ , is composed of vision part  $X_V$  and language part  $X_L$ , with language annotation  $Y$  having  $n$  tokens  $(y_1, y_2, \dots, y_n)$  from our training dataset, including InstructIAD and CoT-M.  $X_V$  is composed of query image  $X_{img}$  and extra patches produced by EG-RoI.

$$X_V = EG-RoI(X_{img}, Expert(X_{img})). \quad (1)$$

Our fine-tuning procedure follows the common setting of Supervised Fine-Tuning (SFT), including AdamW [26] op-

timizer and Cross Entropy loss function  $\mathcal{L}_{SFT}$ .

$$\mathcal{L}_{SFT} = - \sum_{i=1}^n \log \pi_{\theta}(y_i | X, y_1, y_2, \dots, y_{i-1}), \quad (2)$$

where  $\theta$  represents the parameters of LMMs and  $\pi_{\theta}(y|X)$  represents the probability of LMM generating the response  $y$  given input  $X$ . The whole training procedure could be formulated as optimize  $\mathcal{L}_{SFT}(LMM(X_V, X_L, \theta), Y)$ . Our fine-tuning involves all parameters of LMM and has only one stage with all tasks mixed from both InstructIAD and CoT-M. Technique details will be demonstrated in Sec. 4.

#### 3.3. InstructIAD Dataset

We first collect an instruction-tuning dataset with human-annotated attribute-rich captions and coarse labels from existing IAD datasets, RealIAD [30], VisA [41], and MVTec-LOCO AD [3], as shown in the left part of Fig. 2. InstructIAD comprises 9,444 abnormal and 13,578 normal samples spanning 40 different products in RealIAD and VisA. From these, we manually annotate 2,098 samples (1,049 normal and 1,049 defect images) across all 40 classes, providing fine-grained, attribute-level captions describing products (e.g., color, shape, layout, material, texture) and defects (e.g., location, orientation, shape, color). InstructIAD supports three key tasks: i) Anomaly detection – a binary (Yes/No) classification task for samples that only have normal or abnormal labels, ii) Attribute-level caption – similar to standard image caption but with a heightened focus on detailed product and defect attributes, and iii) Anomaly analysis – a two-part task where the model predicts whether an image is normal or abnormal, then provides an explanation grounded in the relevant visual attributes. Explanations are generated by LLaMA3 [9] from the attribute-level descriptions. Together, these tasks build a logical path from fine-grained visual attributes to anomaly detection. Please refer to Sec. C in the supplementary material for details.

#### 3.4. Chain-of-Thought with Manufacturing

To achieve manufacturing-aware reasoning, we introduce CoT-M: a data organization strategy that infuses Chain-of-Thought (CoT) reasoning with manufacturing processes. CoT-M is built upon InstructIAD and aims to enhance anomaly detection by joint reasoning between product attributes and manufacturing-related information.

CoT-M extends InstructIAD along two axes: (i) product and defect diversity and (ii) manufacturing awareness. The data-generation pipeline (Fig. 2) adapts to the information available for each sample through three complementary modes: (a) Images with captions (Fig. 2, middle): The original attribute-level caption and its MFG are fed into GPT, which interleaves visual details with manufacturing steps to yield a coherent chain-of-thought (CoT). The outcome is an annotated triplet (image, manufacturing process,

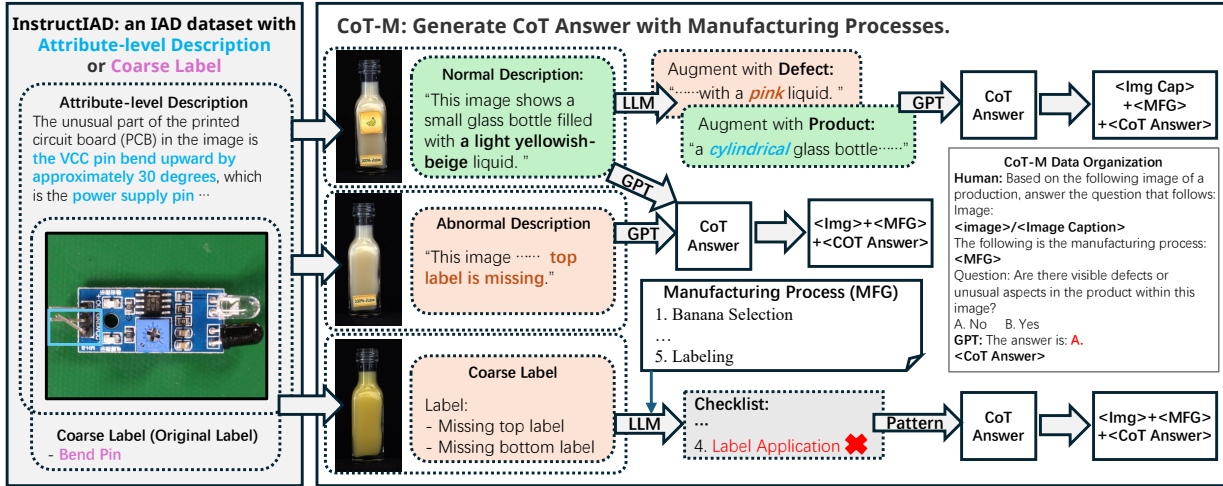


Figure 2. Overview of the proposed CoT-M data organization pipeline. Given normal product descriptions and manufacturing process information (top row), the large language model expands product and defect types by editing product attributes, generating textual CoT data. When incorporating visual input, there are two scenarios: images with detailed textual descriptions (middle row) and images annotated only with coarse labels (bottom row), collectively forming multimodal CoT data.

reasoning trajectory); (b) Images without captions (Fig. 2, bottom): For categories such as “juice bottle” in MVTEC-LOCO AD, caption information is absent. We therefore invoke a checklist-style template that emulates a human inspector. Each coarse defect label is cross-referenced with an LLM-generated MFG checklist to pinpoint the faulty visual attribute; the completed checklist itself serves as the CoT explanation. (c) Text-only exemplars (Fig. 2, top): Starting from a normal product caption in InstructIAD, we synthetically augment the product by altering attributes (e.g., colour, component type, quantity) and stochastically injecting defect descriptions. GPT then generates the reasoning trajectory conditioned on the augmented caption and its MFG, producing a purely textual CoT pair. After generating the accompanying reasoning trajectories with the augmented descriptions and manufacturing processes, we manually filter out any hallucinated or erroneous samples. The final output, including an augmented caption, corresponding manufacturing processes, and generated reasoning process forms a new textual CoT answer pair.

### 3.5. Expert-guided RoI module

In LMMs, the image is encoded into tokens by a visual encoder such as CLIP [27] and SigLIP [37], thus the encoder resizes the resolution of the input images to a fixed size. To offset the effect of limited resolution, LLaVA-Next [22] has proposed a technique called AnyRes by entering extra large patches of the original image. This technique improves performance and has been further developed in LLaVA-OneVision [16]. However, dividing the whole image into patches as extra patches is memory-consuming and redundant for IAD, as most patches of a defective image do not include defects. Accordingly, we propose the EG-RoI mod-

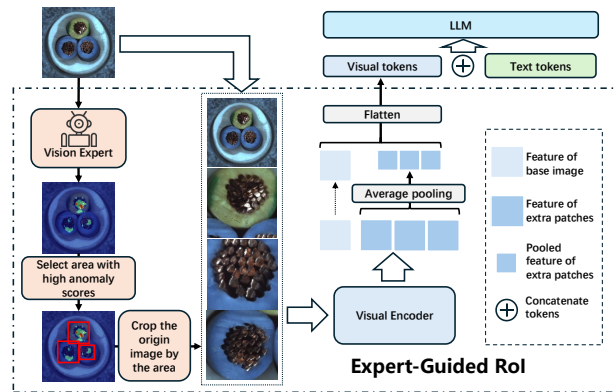


Figure 3. The whole workflow of EG-RoI, with cropped potential defective area encoded as extra patches, Triad could have more detailed visual information to inspect.

ule (see Fig. 3), which improves the recognition of product attributes and defects. During training, ground-truth defect regions together with randomly sampled normal regions are supplied to the module. During the evaluation, we can leverage established zero-shot IAD vision experts—MuSc [18], AnomalyCLIP [40], and April-GAN [7]—to supply informative suspicious regions to EG-RoI. Anomaly maps from the vision experts are normalized and binarized; pixels scoring  $> 0.9$  define suspicious regions. Each region is enclosed in a fixed-size box cropped from the image to preserve resolution. To comply with the language model’s context-length limit, heavily overlapping boxes are merged, and the total number of regions is capped at four. The resized original image forms the base view, while cropped regions serve as auxiliary patches. Both the base image and the patches are processed by the image encoder; patch to-

Table 1. Zero-shot anomaly detection performance with **manufacturing process** (+MFG Proc.) on MVTec-AD [2] and WFDD [6] datasets. The best results are in **bold** while the second best are underlined.

Model	Params	MVTec-AD		WFDD	
		0-shot	+ MFG Proc.	0-shot	+ MFG Proc.
GPT-4o [11]	-	82.2%	67.9% (14.3%↓)	78.5%	77.3% (1.2%↓)
Qwen2-VL [31]	2B	77.0%	46.7%(30.3%↓)	70.6%	45.2%(25.4%↓)
LLava-1.6 [22]	7B	76.9%	75.9% (1.0%↓)	63.8%	64.0% (0.2%↓)
MiniCPM-V [34]	8B	62.3%	51.6%(10.7%↓)	70.3%	52.1%(18.2%↓)
LLaVA-OneVision-si [16]	7B	77.7%	60.6%(17.1%↓)	65.2%	61.4% (3.8%↓)
LLaVA-OneVision-ov [16]	7B	<u>91.0%</u>	80.8%(10.2%↓)	79.8%	<u>80.3%</u> (0.5%↑)
Qwen2-VL [31]	7B	84.4%	61.1%(23.3%↓)	74.4%	61.4%(13.0%↓)
Qwen2-VL [31]	72B	87.1%	79.5% (7.6%↓)	<b>81.1%</b>	74.2% (6.9%↓)
LLaVA-OneVision-ov [16]	72B	87.3%	75.5%(11.8%↓)	75.0%	74.6% (0.4%↓)
Myriad [19]	7B	79.3%	81.5% (2.5%↑)	60.5%	61.7% (1.2%↑)
Triad-llava-1.6	7B	85.0%	<u>87.5%</u> (2.5%↑)	67.3%	69.9% (2.6%↑)
Triad-ov	7B	<b>91.2%</b>	<b>92.6%</b> (1.4%↑)	<u>80.2%</u>	<b>81.1%</b> (0.9%↑)

kens are average-pooled to save memory, flattened, and concatenated with the base-image tokens. The combined visual tokens are then projected into the textual embedding space and passed as a single sequence to the language model. By supplying these high-resolution local views, EG-RoI allows the LMM to conduct a more detailed MFG-driven inspection.

## 4. Experiments

Our experiments mainly focus on how manufacturing processes boost LMMs in IAD tasks through our method, including common results from general LMMs and LMM-based IAD methods and ablation experiments of different industrial contexts.

**Implementation Details** Our method is implemented on both LLaVA-1.6 (the earlier version of LLaVA-NeXT [22]) and LLaVA-OneVision-ov (the checkpoint after the “one-vision stage” of LLaVA-OneVision [16]), referring to them as Triad-llava-1.6 and Triad-ov, respectively. We build our approach on the LLaVA architecture by integrating our expert-guided region-of-interest (EG-RoI) tokenizer. Specifically, in the LLaVA-1.6 version, the original AnyRes module is replaced with the EG-RoI module. In contrast, for LLaVA-OneVision, we append the suspicious regions to the output of the AnyRes module to fully leverage its native anomaly detection capabilities.

To preserve the generalization of the base models, we supplement our IAD-related instruction data by sampling 12K pairs from the original fine-tuning datasets of both LLaVA-1.6 and LLaVA-OneVision. Moreover, since LLaVA-1.6 lacks inherent multi-image processing capabilities, we constructed a simple dual-image caption dataset using the COCO subset [21] from the ShareGPT4V dataset [5]. This dataset is exclusively used in the 1-shot setting to provide basic multi-image support. Given the distinct objectives of 0-/1-shot anomaly detection, we offer

separate versions of Triad for each setting, with data organized by CoT-M with specific instructions. For the 1-shot version of Triad-llava-1.6, we integrate the zero-shot model with the one-shot model using a Confidence Voting Mechanism (see Sec. E.1) to mitigate its multi-image inability.

Both Triad-llava-1.6 and Triad-ov were trained on 4xA800 80G GPUs with a mega-batch size of 128. For Triad-llava-1.6, we set the per-device batch size to 8 with a 4-step gradient accumulation. In the case of Triad-ov, due to the context length increasing from 4096 to 32768 tokens, the per-device batch size was reduced to 1, and gradient accumulation was increased to 32 steps. All other settings follow those established for LLaVA-1.6 and LLaVA-OneVision. With these configurations, 0-shot fine-tuning requires approximately 3 hours, while 1-shot fine-tuning takes about 5 hours for Triad-llava-1.6 and roughly twice that for Triad-ov.

**Evaluation Details** We evaluate Triad using images from MVTec-AD [2], WFDD [6], and PCB-Bank [33] for quantity and quality results, complemented by product-specific manufacturing processes. Since the original datasets do not provide meta-information about their products, we employ ChatGPT4 to generate manufacturing processes based on each product’s name and a caption describing a normal (non-defective) item. In real-world industrial applications, these manufacturing process details would typically be supplied directly by the factory.

Our evaluation comprises 21 different products and a total of 3003 multiple-choice questions following recent multimodal benchmarks [15, 24, 35], spanning both object and texture categories. To broaden the complexity of anomaly detection scenarios, we incorporate WFDD and PCB-Bank, which introduce diverse texture-based and object-based defect types, respectively. Similarly, we use multiple-choice accuracy as the metric.

For zero-shot evaluation, the instruction follows the same format as the anomaly detection task shown in Fig. 2. For one-shot evaluation, general LMMs that support multi-image directly input the query and a normal reference image with a similar prompt with zero-shot, only the question is replaced with:

*The second image shows an acceptable product. Compared to the acceptable product, find out whether there are defects in the product in the first image.*

LMM-based IAD method AnomalyGPT [10] and Myriad [19] are tested according to their own instructions. Because of their low instruction following ability, we retrieve the keyword *yes* or *no* for accuracy calculation. Detailed demonstration can be found in Sec. B.

**Baseline** In this study, we perform extensive evaluations against several robust baselines, including AnomalyGPT [10] and Myriad [19]. General-purpose LMMs exhibit enhanced reasoning capabilities as a result of be-

ing fine-tuned on complex reasoning data spanning diverse domains, including mathematics, chart interpretation, and document analysis. For evaluation, we benchmark our approach against state-of-the-art general-purpose LMMs, including Qwen2-VL [31], MiniCPM-V [34], the LLaVA series [16, 22], and the closed-source GPT-4o [11].

#### 4.1. Main results

**Quantity Results** The zero-shot anomaly detection results are presented in Tab. 1. Although state-of-the-art LMMs such as LLaVA-OneVision-ov-7B [16] and Qwen2-VL-72B [31] demonstrate strong zero-shot performance on MVTec-AD, WFDD, and PCB-Bank (see Sec. A.1), our Triad-ov-7B remains highly competitive. Notably, by integrating manufacturing processes, Triad-ov-7B surpasses LLaVA-OneVision-ov-7B by 1.6%, highlighting the advantage of manufacturing-aware reasoning in IAD. In contrast, most general-purpose LMMs fail to integrate manufacturing information, resulting in a significant performance drop. Interestingly, we also observe that Myriad benefits from manufacturing processes, likely due to its built-in visual enhancement mechanisms.

To evaluate the adaptability of our approach, we provide two versions of Triad based on different LMM backbones, LLaVA-1.6-7B and LLaVA-OneVision-ov-7B. When starting with the relatively weak LLaVA-1.6-7B model, Triad yields substantial gains of 8.1% (from 76.9% to 85.0% on MVTec-AD) and 3.5% (from 63.8% to 67.3% on WFDD). Moreover, equipping Triad-llava-1.6-7B with manufacturing processes adds a further 2.5% and 2.6% improvement, respectively. Similarly, Triad-ov-7B outperforms the original LLaVA-OneVision-ov-7B and gains an additional 1.4% and 0.9% boost from manufacturing-aware reasoning on MVTec-AD and WFDD, respectively. These results confirm that our approach generalizes effectively across different model architectures and continues to enhance performance even at higher baseline accuracies (e.g., above 90%).

Table 2 shows 1-shot results. Providing a single reference image generally degrades the performance of most general-purpose LMMs due to their limited instruction-following capabilities. Two exceptions, LLaVA-OneVision-ov-7B (91.5% vs. 91.0% in 0-shot) and Qwen2-VL-72B (92.0% vs. 87.3% in 0-shot), still cannot fully utilize manufacturing information for anomaly detection (accuracy down by 5.7% and 1.6%, respectively). Meanwhile, Triad-llava-1.6-7B achieves comparable performance to both AnomalyGPT and Myriad under 1-shot conditions, and significantly outperforms them with manufacturing processes (88.4% vs. 85.4% for Myriad and 80.9% for AnomalyGPT). Leveraging LLaVA-OneVision-ov-8B—a carefully tuned multi-image baseline—Triad-ov-8B achieves 92.9% on MVTec-AD in the 1-shot setting, further improving to 94.1% when incorporating manufacturing processes.

Table 2. One-shot anomaly detection performance with **manufacturing process** on MVTec-AD [2] The best results are in **bold** while the second best are underlined.

Model	Params	MVTec-AD	
		1-shot	+ MFG Proc.
GPT-4o [11]	-	77.6%	72.5% (5.1%↓)
Qwen2-VL [31]	2B	32.1%	30.7% (1.4%↓)
LLava-1.6 [22]	7B	72.3%	76.0% (3.7%↑)
LLaVA-OneVision-ov [16]	7B	91.5%	85.2% (5.7%↓)
Qwen2-VL [31]	7B	81.7%	84.5% (2.8%↑)
Qwen2-VL [31]	72B	<u>92.0%</u>	90.4% (1.6%↓)
AnomalyGPT <sup>1</sup> [10]	7B	86.1%	80.9% (5.2%↓)
Myriad [19]	7B	87.4%	85.4% (2.0%↓)
Triad-llava-1.6	7B	87.7%	88.4% (0.7%↑)
Triad-ov	7B	<b>92.9%</b>	<b>94.1%</b> (1.2%↑)

These results suggest that even a strong base model can benefit from the Manufacturing processes. Overall, our findings demonstrate that Triad offers robust and flexible anomaly detection across both zero-shot and few-shot scenarios, while the manufacturing-driven IAD paradigm consistently delivers strong performance gains.

**Qualitative results.** Figure 4 compares IAD with and without manufacturing processes. In the standard IAD setting, Triad and GPT-4o both identify a board-level defect. However, Myriad incorrectly flags a different (normal) component, and GPT-4o infers inconsistent soldering without a clear visual cue. When manufacturing process information is included, Triad accurately detects the missing button by referring to the white square silkscreen mark, which indicates the location of the button. This demonstrates Triad’s manufacturing-aware reasoning capacity, even when handling previously unseen manufacturing descriptions.

Figure 5 presents three representative cases. In the first example (Fig. 5a), a surface discoloration might typically be viewed as a defect (as GPT-4o incorrectly predicts); however, Triad recognizes that the discoloration corresponds to the cranberry additive described in the manufacturing process, leading to a correct classification. Thanks to the EG-RoI module’s ability to capture fine-grained visual cues and the data organization via CoT-M, Triad confidently aligns the observed appearance with the relevant manufacturing steps. In the second example (Fig. 5b), a pill containing “yellow speckles” is mistakenly introduced into a production pipeline that should only produce pills with “red speckles.” Triad correctly identifies the inconsistency between the visual observation and the manufacturing procedures. GPT-4o, though aware that red speckles are the standard, fails to accurately match the color of the speckles in question. This underscores the necessity of precise attribute recognition for reasoning. Finally, the third example (Fig. 5c) highlights Triad’s resilience to vision expert errors, where

<sup>1</sup>AnomalyGPT adds hints mentioning defect types for each product in MVTec-AD test set.

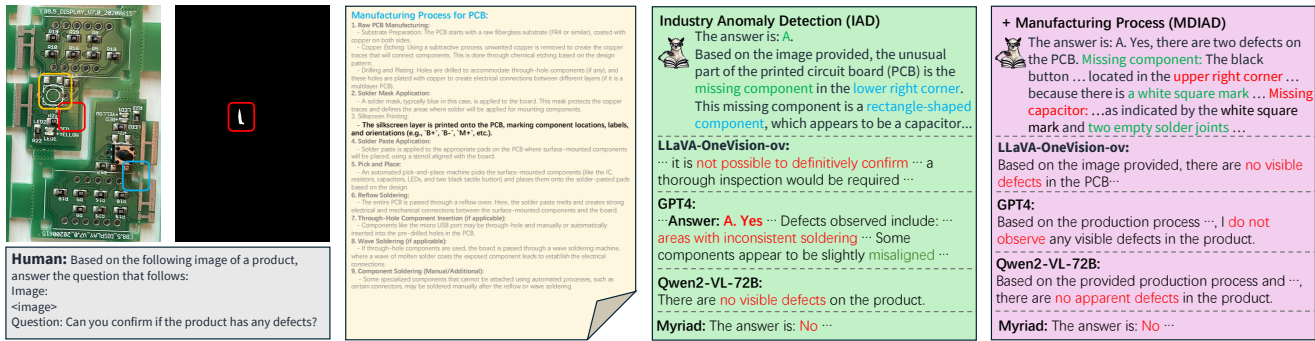


Figure 4. Comparison with and without manufacturing processes on a PCB7 example from PCB-Bank [33]. The responses from state-of-the-art LMMs are provided. Expert’s detection is indicated by a red bounding box, while the ground truth is shown in yellow. An additional component identified by Triad is highlighted with blue bounding boxes. Best viewed in color.

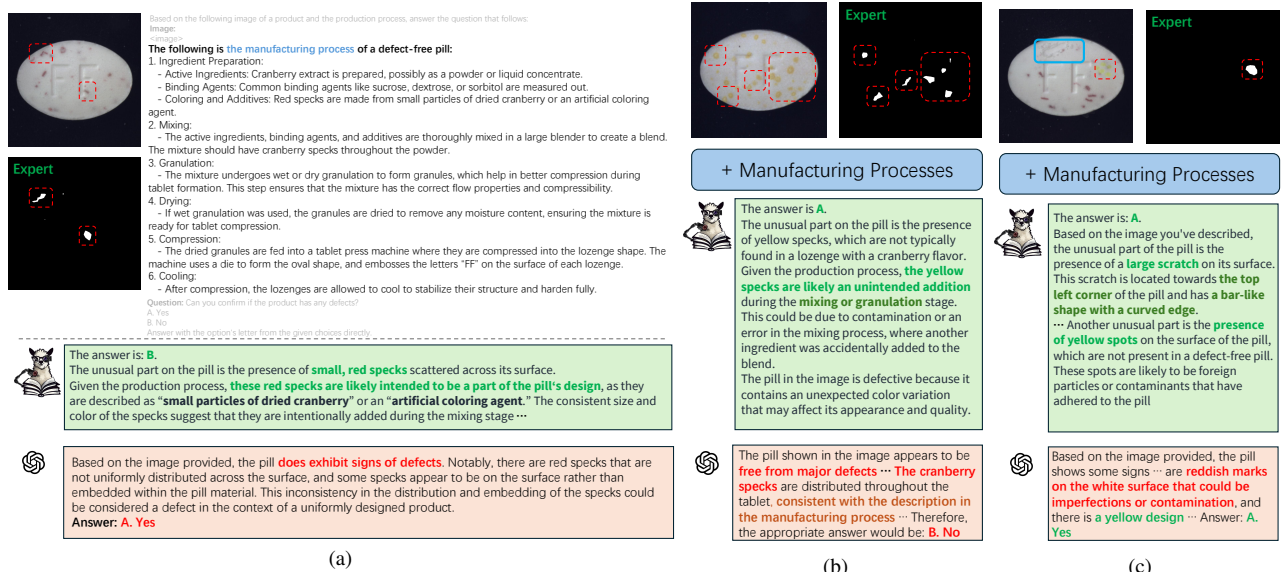


Figure 5. Qualitative evaluation of Triad-llava-ov-7B on MVTEC-AD [2]. MuSc [18] is used as vision expert. The related manufacturing process is shown in (a). Three representative examples from the “Pill” class are shown: (a) a normal pill exhibits red speckles on its surface which are a natural outcome of the manufacturing process and could be easily misjudged as defects. (b) Pills with unintended additive variations indicate a defect. (c) a case where Triad correctly predicts a defect despite errors in the expert outputs.

it fails to detect a scratch on the surface. Triad still identifies the scratch independently and also notes a yellow discoloration, demonstrating its capacity to rectify external region proposal mistakes by integrating visual analysis with contextual defect reasoning.

Overall, these observations illustrate Triad’s ability to integrate manufacturing knowledge with detailed visual attributes, thereby achieving more accurate and context-sensitive anomaly detection.

4.2. Ablations

**Effect of InstructIAD.** InstructIAD is placed as an instruction-tuning dataset for IAD, which aims to align the visual features and textual descriptions, leading to a performance gain compared with the base model (LLaVA-1.6-

mistral-7B). Notably, as shown in Tab. 3, the performance with manufacturing processes also increases, empirically demonstrating the intrinsic relationship between attribute-level recognition and manufacturing-aware defect analysis.

**Effect of CoT-M.** As evidenced by Tab. 3, comparative analysis of row 2 vs. 3 and row 4 vs. 5 demonstrates the efficacy of CoT-M. Utilizing CoT-M-generated data enhances manufacturing process detection accuracy by 2.1% and 3.6% respectively, while maintaining comparable/superior performance in standard detection scenarios compared to non-CoT-M approaches. This finding substantiates that CoT-M empowers Triad to effectively leverage manufacturing process information for improved anomaly detection.

**Effect of EG-RoI.** Comparative analysis in Tab. 3 (row 3 vs. 5) confirms EG-RoI’s consistent performance gains

Table 3. Ablations on fine-tuning with different components: InstructIAD, CoT-M, and EG-RoI tokenizer. The base model is LLaVA-1.6-mistral-7B [22]

InstructIAD	CoT-M	EG-RoI	0-shot	+ MFG Proc.
✗	✗	✗	76.9%	75.9% (1.0%↓)
✓	✗	✗	78.8%	80.4% (1.6%↑)
✓	✓	✗	79.8%	82.5% (2.7%↑)
✓	✗	✓	85.4%	83.9% (1.5%↓)
✓	✓	✓	85.0%	87.5% (2.5%↑)

Table 4. Ablation on the robustness of the manufacturing process on MVTec-AD [2]. Three types of the manufacturing processes of 15 products on MVTec-AD are collected from the Internet (Web), generated by llama3 [9] (LLM), and generated by GPT4 [1] (GPT).

Model	Params	0-shot	+ MFG Proc. (Web)	+ MFG Proc. (LLM)	+ MFG Proc. (GPT)
Qwen2-VL [31]	2B	77.0%	51.7%(25.3%↓)	46.7%(30.3%↓)	46.8%(30.2%↓)
LLava-1.6 [22]	7B	76.9%	77.6% (0.7%↑)	75.9% (1.0%↓)	77.3% (0.4%↑)
LLaVA-OneVision-si [16]	7B	77.7%	69.4% (8.3%↓)	60.6%(17.1%↓)	67.2%(10.5%↓)
LLaVA-OneVision-ov [16]	7B	91.0%	88.8% (2.2%↓)	80.8%(10.2%↓)	87.1% (3.9%↓)
Qwen2-VL [31]	7B	84.4%	70.9%(13.5%↓)	61.1%(23.3%↓)	67.1%(17.3%↓)
Qwen2-VL [31]	72B	87.1%	85.2% (1.9%↓)	79.5% (7.6%↓)	82.6% (4.5%↓)
Myriad [19]	7B	79.3%	78.5% (0.8%↓)	81.5% (2.2%↑)	81.5% (2.2%↑)
Triad-llava-1.6	7B	85.0%	85.7% (0.7%↑)	87.5% (2.5%↑)	86.4% (1.4%↑)
Triad-ov	7B	91.2%	91.8% (0.6%↑)	92.6% (1.4%↑)	92.2% (1.0%↑)

across both manufacturing-aware and conventional detection scenarios. We further evaluate vision expert effectiveness through size-based defect categorization on MVTec-AD, adopting MS-COCO’s partitioning protocol [21]: small ( $< 0.01$  image area), medium ( $0.01 \sim 0.1$ ), and large ( $> 0.1$ ). As shown in Tab. 6, EG-RoI significantly outperforms LLaVA-1.6’s AnyRes module on small/medium defects. The baseline exhibits severe classification bias with 100% defect detection rate but 16.1% false positives on normal samples. Table 5 reveals two critical findings: (1) Anomaly map quality directly impacts detection accuracy, validating our expert integration strategy; (2) Triad achieves competitive performance even without annotated regions (AnyRes mode), surpassing conventional methods. Notably, EG-RoI maintains manufacturing-aware improvements despite random box noise injection (Null mode).

**Ablations on the different manufacturing process.** For systematic evaluation, we collect three distinct manufacturing process variants for MVTec AD [2] through multi-source construction: (1) Internet: Domain-specific processes collected from the internet for all 15 product categories; (2) LLM Generation: Automated process synthesis using general-purpose LLMs (LLaMA-3 [9] and GPT-4 [1]). As evidenced in Tab. 4, Triad demonstrates consistent performance gains across all process variants (0.7%-2.5% accuracy improved), empirically validating its dual capability in manufacturing process comprehension and generalization to unseen industrial workflows. Extended process examples are provided in Sec. B.

Table 5. Comparison between different vision experts with the proposed EG-RoI. The model is based on LLaVA-1.6-mistral-7B [22]. The quality of anomaly maps by experts on zero-shot IAD tasks is measured by Expert P-AUROC.

Vision Expert	Expert P-AUROC	base	+ MFG Proc.
Null	-	83.3%	83.4% (0.1%↑)
AnyRes	-	84.0%	85.1% (1.1%↑)
April-GAN [7]	87.6%	83.0%	85.0% (2.0%↑)
AnomalyClip [40]	91.1%	84.6%	86.1% (1.5%↑)
MuSc [18]	97.3%	85.0%	87.5% (2.5%↑)

Table 6. Ablation on how Expert-Guided RoI module affects the performance. Defects are divided into small, medium, and large according to their size. The model is based on LLaVA-1.6-mistral-7B. All results are tested without context.

Module	small defects	medium defects	large defects	normal	Accuracy
baseline (llava-1.6) [22]	100.0%	100.0%	100.0%	16.1%	76.9%
AnyRes (Finetune) [22]	90.9%	81.0%	65.7%	82.9%	79.8%
EG-RoI (Finetune)	95.5%	94.1%	81.8%	72.8%	85.0%

## 5. Conclusion

In this paper, we introduced **Triad**, a novel large multi-modal model (LMM) tailored for industrial anomaly detection. By integrating an expert-guided region-of-interest tokenizer (EG-RoI) to highlight suspicious regions identified by existing IAD methods, Triad improves its ability to pinpoint subtle defects in complex industrial scenarios. In addition, we proposed a manufacturing-driven IAD paradigm that embeds causal knowledge of manufacturing processes into the model’s reasoning for anomaly detection. Specifically, we contribute an instruction-tuning dataset, *InstructIAD*, and a Chain-of-Thought with manufacturing strategy, CoT-M, enabling Triad to reason about defect formation in relation to manufacturing steps. Experimental results on standard IAD benchmarks demonstrate that Triad achieves superior performance in 0-/1-shot settings compared to both general-purpose and domain-specific LMMs. Extensive evaluations reveal Triad’s novel ability to leverage manufacturing processes to achieve improved anomaly detection. Qualitative results show Triad’s significant reasoning ability based on attribute recognition and manufacturing comprehension. These findings confirm the significance of combining expert-guided region-of-interest tokenizer with manufacturing-aware reasoning for robust and interpretable anomaly detection.

We publicly release InstructIAD’s dataset and CoT-M data organization to facilitate future research, bridging the critical gap between general-purpose LMMs and domain-specific industrial inspection needs. We believe this work lays a starting point for modern quality control systems that synergize human knowledge with multimodal AI reasoning.

## Acknowledgement

This work was supported by the National Key R&D Program of China under Grant No. 2023YFA1008500.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8
- [2] 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, 5, 6, 7, 8
- [3] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022. 3
- [4] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2025. 2
- [5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 5
- [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, 5
- [7] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 2, 4, 8
- [8] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pages 17–33. Springer, 2024. 2
- [9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 8
- [10] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. 3, 5, 6, 2
- [11] 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. 5, 6
- [12] 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
- [13] 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
- [14] Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramid-flow: High-resolution defect contrastive localization using pyramid normalizing flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14143–14152, 2023. 2
- [15] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5
- [16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 4, 5, 6, 8
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3
- [18] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *The Twelfth International Conference on Learning Representations*, 2024. 4, 7, 8, 3, 5
- [19] Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070*, 2023. 3, 5, 6, 8, 1
- [20] Yufei Liang, Jiangning Zhang, Shiwei Zhao, Runze Wu, Yong Liu, and Shuwen Pan. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*, 2023. 2
- [21] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 5, 8
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 4, 5, 6, 8, 3
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

- [24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 5
- [25] 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
- [26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 3
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4
- [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
- [29] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24511–24520, 2023. 2
- [30] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-ia-d: 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. 3, 1, 2
- [31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 5, 6, 8, 1
- [32] Xiaohao Xu, Yunkang Cao, Yongqi Chen, Weiming Shen, and Xiaonan Huang. Customizing visual-language foundation models for multi-modal anomaly detection and reasoning. *arXiv preprint arXiv:2403.11083*, 2024. 2
- [33] Hang Yao, Ming Liu, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. *European Conference on Computer Vision*, pages 1–17, 2024. 2, 5, 7, 1, 3
- [34] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2, 5, 6, 1
- [35] 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 Wenhao Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 5
- [36] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem: a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2
- [37] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, 2023. 4
- [38] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 2
- [39] 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
- [40] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2023. 4, 8, 5
- [41] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 3, 1, 2