

OmniAD: Detect and Understand Industrial Anomaly via Multimodal Reasoning

Shifang Zhao^{1,2}, Yiheng Lin^{1,2}, Lu Han^{3,4}, Yao Zhao^{1,2}, Yunchao Wei^{1,2*}

¹ Institute of Information Science, Beijing Jiaotong University

² Visual Intelligence + X International Joint Laboratory of the Ministry of Education

³Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences

⁴University of Chinese Academy of Sciences

shifang.zhao@bjtu.edu.cn

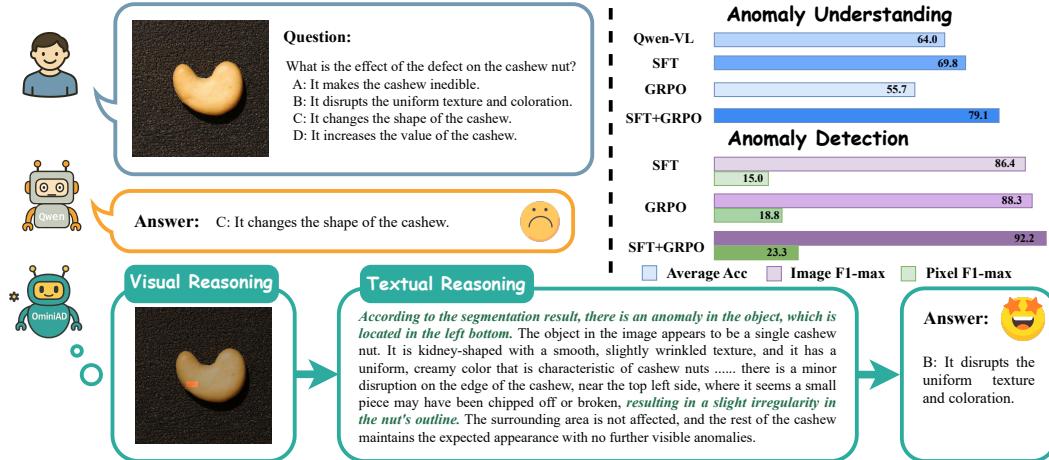


Figure 1: OmniAD unifies anomaly detection and understanding through Multimodal Reasoning. It leverages visual reasoning for anomaly classification and localization, followed by textual reasoning for a comprehensive analysis. The integrated supervised fine-tuning (SFT) and reinforcement learning (GRPO) training strategy ensures superior generalization with few-shot sample.

Abstract

While anomaly detection has made significant progress, generating detailed analyses that incorporate industrial knowledge remains a challenge. To address this gap, we introduce OmniAD, a novel framework that unifies anomaly detection and understanding for fine-grained analysis. OmniAD is a multimodal reasoner that combines visual and textual reasoning processes. The visual reasoning provides detailed inspection by leveraging Text-as-Mask Encoding to perform anomaly detection through text generation without manually selected thresholds. Following this, Visual Guided Textual Reasoning conducts comprehensive analysis by integrating visual perception. To enhance few-shot generalization, we employ an integrated training strategy that combines supervised fine-tuning (SFT) with reinforcement learning (GRPO), incorporating three sophisticated reward functions.

*Corresponding author.

Experimental results demonstrate that OmniAD achieves a performance of 79.1 on the MMAD benchmark, surpassing models such as Qwen2.5-VL-7B and GPT-4o. It also shows strong results across multiple anomaly detection benchmarks. These results highlight the importance of enhancing visual perception for effective reasoning in anomaly understanding. All codes and models will be publicly available.

1 Introduction

"Accuracy of observation is the equivalent of accuracy of thinking."

—Wallace Stevens

Anomaly detection is an important research area in computer vision due to its crucial role in the development of highly autonomous factories. Traditional anomaly detection models typically focus on pixel-level or image-level label prediction. While these approaches can identify the presence of anomalies, they do not offer detailed insights into the anomalous conditions, especially those requiring a high-level understanding. To address this limitation, the task of anomaly understanding has been introduced. This task involves developing models that can provide a deeper understanding of the sample image and the specific factors associated with the detected anomaly. The development of Multimodal Large Language Models (MLLMs) has demonstrated their potential [7, 58, 33], but a gap remains in accurately inspecting anomalies.

What makes anomaly understanding difficult for MLLMs? We argue that there are two key challenges: **1) Anomaly Unawareness.** Perceptual anomalies are difficult to identify because they require detailed inspection of low-level observations, which differs from tasks with natural images. Anomaly understanding is a hierarchical process, where the model must first identify the anomalous region based on perceptual results, and then apply relevant domain knowledge to conduct a thorough analysis. **2) Data Limitation.** Since MLLMs are trained on general perceptual and cognitive tasks, they require domain-specific data to inject specialized knowledge. However, the scarcity of anomalous samples and the extreme class imbalance make it impractical to construct large training datasets that capture comprehensive industrial knowledge. This highlights the importance of effectively and efficiently leveraging limited data.

To address the challenge of Anomaly Unawareness, OmniAD employs a proposed Multimodal Reasoning process that unifies detection and understanding. It firstly conducts visual reasoning by performing anomaly detection with Text-as Mask Encoding through text generation. It then carries out textual reasoning for high-level analysis, integrating the results from the visual reasoning. This method addresses a bottleneck associated with integrating detection results, a common challenge in expert-based systems [34, 12, 27]. By directly generating textual format localization of anomalies, the approach avoids reliance on manually selected detection thresholds, a significant shortcoming of many traditional anomaly detection models. Additionally, the proposed visual and textual reasoning framework offers transparency into the anomaly understanding process, facilitating both validation and adoption. To tackle the challenge of Data Limitation, we employ an integrated training strategy that combines supervised fine-tuning (SFT) and reinforcement learning, specifically Group Relative Policy Optimization(GRPO), inspired by recent advances in reasoning models [38, 37, 44]. This approach enhances few-shot generalization for both anomaly detection and understanding tasks.

Experimental results show that OmniAD demonstrates strong capabilities in both anomaly detection and understanding. We evaluate its anomaly understanding performance on the comprehensive benchmark MMAD [25], where OmniAD achieves an accuracy of 79.1, surpassing the base model Qwen2.5-VL-7B [3] and even GPT-4o by 4.2. In four anomaly detection datasets [5, 4, 59, 53], OmniAD outperforms many traditional anomaly detection models without relying on manually selected thresholds. This not only highlights the robust performance of OmniAD in anomaly detection but also underscores the crucial role of reinforcement learning in enhancing visual perception and reasoning.

We summarize our contributions as follows:

- We propose OmniAD, a novel framework designed to enhance anomaly understanding through anomaly perceptive capabilities. It represents the unified method for both anomaly detection and understanding.

- We develop Multimodal Reasoning for fine-grained anomaly inspection by performing visual reasoning to detect anomalies using Text-as-Mask Encoding, followed by Visual-Guided Textual Reasoning for a thorough analysis.
- Extensive experiments demonstrate the effectiveness of our few-shot training strategy based on SFT and GRPO with three sophisticated rewards, achieving improved performance in both anomaly detection and understanding.

2 Related Work

2.1 Industrial Anomaly Detection

Traditional anomaly detection methods intend to discriminate and localize defects using only normal samples during training, and can generally be categorized into two types. Embedding-based methods [42, 22, 26, 13, 19] typically extract embeddings from a pre-trained model into a compact subspace or a memory bank. Anomalies are then separated from normal samples by comparing embeddings. Reconstruction-based methods [20, 14, 51, 54, 52, 15] train a reconstruction network using only normal samples. By failing to accurately reconstruct anomalous regions, reconstruction errors serve as anomaly scores.

Recently, few-shot anomaly detection methods have focused on modeling anomalies using a limited number of training samples. Approaches such as WinCLIP [24], AnomalyCLIP [57], AdaCLIP [8], CLIP-AD [10], and AA-CLIP [39] leverage the generalization capability of CLIP [40] for anomaly detection through prompt learning and multi-scale feature aggregation. AnomalyGPT [16] first introduced large language models to anomaly detection by employing an image decoder to provide fine-grained semantics and designing a prompt learner for fine-tuning. SAA [6] and UniVAD [17] adapt SAM [29] using patch-level feature matching with language prompts or stored normal features.

However, many of these methods rely considerably on manually selected thresholds, which can reduce performance robustness. In contrast, our OmniAD treats segmentation as a text generation task, thereby avoiding reliance on such thresholds. This offers greater applicability across various anomaly detection tasks.

2.2 Industrial Anomaly Understanding

Recent years have witnessed significant advancements in MLLMs, with many studies now leveraging their robust capabilities in analysis and explanation for industrial anomaly understanding. Myriad [33] first used an existing expert model to localize anomaly regions and an MLLM for its instruction-following capabilities. AnomalyGPT [16] explores a new paradigm that generates masks and textual descriptions in two parallel branches, which are trained on synthetic anomaly samples. This approach eliminates the threshold for image-level detection through direct text output but still depends on such a threshold for pixel-level detection. LogicAD [28], LogiCode [55], and LAD-Reasoner [32] focus on logical anomaly detection and reasoning. Specifically, LogicAD employs a one-shot algorithm with text feature memory banks and multimodal models. LogiCode uses Large Language Models to generate executable Python code for interpreting logical relationships. LAD-Reasoner introduces a customized tiny multimodal language model with a two-stage training approach for improved reasoning in anomaly detection. FabGPT [27] conducts a three-stage model for integrating textual and visual knowledge. AnomalyR1 [9] further improves reasoning capabilities using reinforcement learning. Echo [12] employs four expert modules for stepwise reasoning. Additionally, MMAD [25] serves as the first comprehensive benchmark for anomaly understanding.

Although MLLMs possess strong visual analysis and reasoning capabilities, it is still difficult to localize anomalies accurately and to seamlessly integrate their operations with the outputs of expert models. Our OmniAD addresses these current limitations with a new unified paradigm for detection and understanding.

2.3 Visual Reinforcement Learning

With the emergence of reasoning models such as OpenAI’s o1 [23] and DeepSeek-R1 [18], visual reasoning has been introduced to better perform vision tasks. These reasoning capabilities primarily result from reinforcement learning, especially Group Relative Policy Optimization (GRPO) [43] with

verifiable reward. VLM-R1 [44] first proposed a framework that showed improvement compared to SFT counterparts, especially on complex, real-world, out-of-domain benchmarks. Visual-RFT [38] further extends efficient learning under limited data conditions, demonstrating strong generality. Seg-ZERO [37] verified the generality of GRPO in segmentation tasks involving reasoning.

Some works have attempted to apply GRPO for anomaly understanding because its reasoning process can provide more interpretable information and exhibits strong generality under limited supervision. LAD-Reasoner [32] was proposed for logical anomaly detection, offering a readable reasoning process; however, it was not directly capable of anomaly understanding tasks. AnomalyR1 [9] designed a reasoned outcome alignment metric for an end-to-end framework for reasoning and localization, but it also treated the two tasks independently, and its bounding box format is not suitable for the variable shapes of anomaly regions. To fill this gap, our OmniAD is designed to unify detection and understanding, employing GRPO and enabling patch-level accurate localization.

3 Method

In this paper, we propose OmniAD, a method designed to adapt MLLMs to both anomaly detection and understanding. Unlike previous approaches that treat detection and understanding independently, we introduce a Multimodal Reasoning process that efficiently integrates detection results into the reasoning workflow, using visual reasoning to enhance understanding capabilities. This unified task paradigm is based on Text-as-Mask Encoding, which transforms the segmentation task into a text generation problem. To train this reasoning capability, we employ a combined SFT and GRPO integrated training strategy, utilizing multimodal reasoning format rewards, detection accuracy rewards, and answer rewards.

3.1 Multimodal Reasoning

To enhance fine-grained anomaly perception and analysis capabilities, we integrate the anomaly detection results with Multimodal Reasoning. Directly employing an expert model typically leads to performance degradation [25], so we introduce Text-as-Mask Encoding to transform anomaly detection into text generation, eliminating the need for additional modules or rearchitecture. Then the Visual Guided Textual Reasoning combines the results of visual reasoning to provide a thorough analysis of industrial objects or potential anomalies.

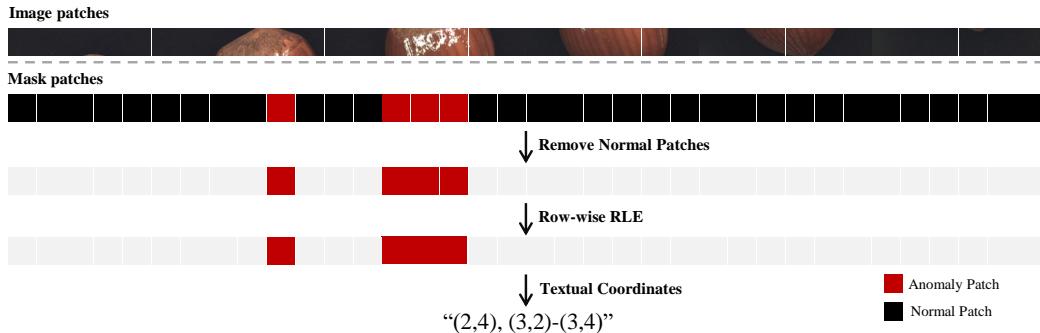


Figure 2: An illustration of Text-as-Mask encoding process.

Text-as-Mask Encoding. Building on recent advancements in reasoning segmentation [30], this approach utilizes the next-token prediction paradigm of MLLM to perform image segmentation by encoding masks as text. We extend this approach to anomaly detection. As illustrated by the example in Fig. 2, the process begins by dividing the image into fixed-size patches (6x6 in this example) and flattening them. Each patch is then labeled with a pixel label ("normal" or "anomaly") corresponding to the content of the patch. However, the masks used in anomaly detection are typically more sparse and irregular in shape compared to those in natural image semantic segmentation. This discrepancy leads to an overwhelming number of negative samples if anomaly patches are treated as positive, complicating the optimization process. To address this issue, we modify the encoding format. First, following the one-class approach commonly used in traditional anomaly detection methods, which focuses on learning representations for a single class, we eliminate normal patches from the sequence

and retain the anomaly patches only. This eliminates the need for labeling each patch individually, allowing them to be efficiently recorded using only the spatial information. Next, we apply run-length encoding (RLE) to compress continuous patches in each row, separating them with commas. This step further shortens the patch sequence, reducing the computational burden. Finally, we convert the patch sequence into a string of coordinates. The resulting encoding format is shorter and has a higher information density.

Visual Guided Textual Reasoning. The Text-as-Mask method can directly perform anomaly detection within text generation, providing valuable information for a better understanding of anomalies by offering both anomaly classification and location. Inspired by this, we introduce a Visual Guided Textual Thinking process that begins with anomaly detection and leverages the detection results in textual reasoning to generate a comprehensive, well-considered answer. The response format is structured as `<seg>...<\seg><think>...<\think><answer>...<\answer>`, which separates different processing steps using special tokens, with the system prompt outlined in Table 1. This reasoning process not only enables simultaneous detection and understanding in a single text generation step but also efficiently combines these two tasks for hierarchical analysis.

Table 1: The system prompt for leading model output Visual Guided Textual Thinking.

"You are a special assistant for analysis image. The user asks a question, and you answer with the option's letter from the given choices directly. Firstly, you should detect anomalies in patches within `<seg>/<\seg>` tags. Then think about the reasoning process in the mind and then provide the user with the answer. The reasoning process and answer are enclosed within `<think> /<\think>` and `<answer> /<\answer>` tags, respectively, i.e., `<seg>`segmentation results here`</seg>``<think>` reasoning process here`</think>``<answer>` answer letter here`</answer>`"

3.2 Reward Functions

After DeepSeekR1 [18] first introduced GRPO to enhance model reasoning capabilities, it has demonstrated potential across many tasks due to its adaptability. The core idea behind GRPO is the effectiveness and efficiency of verifiable rewards. GRPO is particularly well-suited for anomaly detection and analysis, as it often involves verifiable results, making relative optimization across multiple solutions a natural fit. In our task, the verifiable reward function plays a crucial role in improving model performance by providing clear and objective feedback during training. Therefore, we propose three reward functions to enhance unified anomaly detection and understanding capabilities.

Multimodal Reasoning Format Reward. This reward is designed to encourage the model to engage in a structured thinking process. It guides the model to output its visual reasoning within `<seg>` and `</seg>`, its textual reasoning within `<think>` and `</think>`, and the final answer between `<answer>` and `</answer>`. A reward of 1 is given if the format is followed; otherwise, the reward is 0.

Detection Accuracy Reward. In contrast to previous reward designs [37, 38] for segmentation tasks that evaluate predictions using Intersection-over-Union (IoU), we use the F1-score as the evaluation metric because it better focuses on precision and recall in class-imbalanced conditions. The process first converts the Text-as-Mask Encoding format prediction into a mask, and then computes the F1-score between the predicted mask and the ground truth. The reward is defined as:

$$R_{\text{Acc}} = \begin{cases} 1 & \text{if } \mathcal{F}_G = \emptyset \text{ and } \mathcal{F}_P = \emptyset \\ 0 & \text{if } \mathcal{F}_G = \emptyset \text{ and } \mathcal{F}_P \neq \emptyset \\ \alpha \cdot \text{F1-score} & \text{if } \mathcal{F}_G \neq \emptyset \end{cases} \quad (1)$$

where \mathcal{F}_G and \mathcal{F}_P represent the anomaly regions in the ground truth and prediction, respectively. α is a scale factor used to balance with other rewards, as the F1-score is typically a relatively small number. This reward function incentivizes higher rewards for more accurate detections and provides no reward for detections in normal images.

Answer Accuracy Reward. This reward assesses the accuracy of the answer compared to the ground truth. A reward of 1 is given for a correct answer, while a reward of 0.1 is assigned for an incorrect answer.

Table 2: The performance comparison of both proprietary and open-source MLLMs in MMAD is presented. All methods are evaluated in a 1-shot setting unless otherwise noted. Anomaly Discrimination uses the average accuracy across the normal and abnormal categories.

Model	Scale	Anomaly		Defect			Object		Average
		Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Random Chance	-	50.0	25.0	25.0	25.0	25.0	25.0	25.0	28.6
Human (expert) [25]	-	95.2	75.0	92.3	83.3	94.2	86.1	80.4	86.7
Human (ordinary) [25]	-	86.9	66.3	85.6	71.3	81.5	89.6	69.7	78.7
Claude-3.5-sonnet	-	60.1	60.1	48.8	67.1	79.1	85.2	79.8	68.4
Gemini-1.5-flash [46]	-	58.6	54.7	49.1	66.5	82.2	91.5	79.7	68.9
Gemini-1.5-pro [46]	-	68.6	60.1	58.6	70.4	82.5	89.2	82.3	73.1
GPT-4o-mini	-	64.3	48.6	38.8	63.7	80.4	88.6	79.7	66.3
GPT-4o	-	68.6	65.8	55.6	73.2	83.4	95.0	82.8	74.9
AnomalyGPT [16]	7B	65.6	27.5	28.0	36.9	32.1	29.8	35.8	36.5
LLaVA-1.5 [35]	7B	51.3	37.0	36.6	50.6	69.8	68.3	69.5	54.7
LLaVA-OneVision [31]	7B	51.8	46.1	41.9	62.2	69.7	90.3	80.9	63.3
MiniCPM-V2.6 [50]	8B	57.3	49.2	43.3	65.9	75.2	92.0	80.8	66.3
InternVL2 [11]	8B	60.0	43.9	47.9	57.6	78.1	74.2	80.4	63.1
LLaVA-1.5 [35]	13B	50.0	38.8	46.2	58.2	73.1	73.6	71.0	58.7
Qwen-2.5-VL [3]	7B	55.9	39.8	45.8	54.8	73.7	93.3	84.7	64.0
LLaVA-NeXT [36]	34B	57.9	48.8	52.9	71.3	80.3	81.1	77.8	67.2
InternVL2 [11]	76B	68.3	54.2	56.7	66.3	80.5	86.4	82.9	70.8
AnomalyRI [9]	3B	60.2	63.5	70.1	80.4	85.2	82.4	86.1	76.9
OmniAD(0-shot)	7B	61.4	67.8	73.5	81.7	87.5	95.2	86.9	79.1
OmniAD(1-shot)	7B	68.8	78.8	75.5	67.2	86.4	96.0	86.4	79.9

3.3 Training

SFT & GRPO Training. To enhance multimodal reasoning capabilities and relevant knowledge, we employ an integrated training strategy combining SFT and GRPO. SFT, based on dense supervision and a robust knowledge distillation process, is a common and highly effective step in post-training for foundation models. We utilize SFT to quickly familiarize the model with industry-relevant object knowledge and multimodal reasoning processes through our Text-as-Mask Encoding format and Visual Guided Textual Thinking. Following SFT, we implement GRPO, a reinforcement learning algorithm designed to improve the model’s reasoning ability without requiring precise reasoning annotations in the training data, relying solely on a verifiable reward function. Our findings show that this training strategy significantly boosts the model’s reasoning performance, ultimately leading to improved detection and understanding.

Data preparation. To train OmniAD for unified detection and understanding capabilities, we need to construct a multimodal training dataset. For the SFT training stage, dense annotated text is required for reasoning and answering. Unlike fields such as mathematics or science, where abundant high-quality reasoning data is available for training, collecting analysis data for anomaly processing is challenging due to the need for relevant information and specialized skills. To address this, we implement a data generation pipeline for high-quality visual question answering (VQA) following MMAD [25]. We select one example per category from several industrial anomaly detection datasets [48, 56, 2, 49, 21, 47]. Following the approach in [9], we add one example per category from MMAD to avoid domain gaps between datasets. For the GRPO training stage, reasoning annotations are not necessary, so we also use one example per category from MMAD. In total, there are 1.6K images with 6.4k QA pairs for SFT and 293 images with 1.1k QA pairs for GRPO. To enable more flexible inference, we extend our method to the one-shot setting by modifying the question with the prefix `<image> This is an image of a normal object.`, where `<image>` refers to a randomly selected normal image from the same category as the query image.

4 Experiments

4.1 Experimental Settings

Implementation Details. OmniAD is built on Qwen2.5-VL-7B[3], one of the best open-source MLLMs, offering an optimal balance between performance and latency. It also includes several implementations with RL post-training, showcasing its potential. To mitigate the impact of input size, all images are resized to 512 during both training and testing. The patch number of Text-as-Mask Encoding is set to 24x24. The scale factor in Detection Accuracy Reward is set to 2. OmniAD is trained on 4xH800 GPUs using the DeepSpeed library [41]. During training, we use a total batch size



Question:
There is a defect in the object. What is the appearance of the defect?
A: A series of pulled threads. B: A discoloration on the metal part.
C: A large hole. **D: A small area that looks like a smudge or stain.**



Answer:
<seg>(14, 8)-(14, 9), (15, 8)-(15, 10), (16, 8)-(16, 10)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the left. The object in the image is a black zipper, prominently displaying its teeth and fabric tape on either side. Upon inspection of the zipper using the provided normal image for comparison, there is evidence of a defect present. The defect appears to be a small spot of dirt or a similar substance located near the bottom center of the zipper. This anomaly is characterized by a slightly darker color compared to the surrounding area of the zipper, indicating that it may be a foreign material adhered to the surface or a residue from manufacturing. The appearance of the defect is relatively discrete but noticeable upon close examination, differing in texture and color from the rest of the zipper's surface.</think>
<answer>D</answer>





Question:
There is a defect in the object. Where is the defect region?
A: Top left B: Bottom right C: Center D: Scattered throughout



Answer:
<seg>(4, 3)-(4, 4), (5, 3)-(5, 4)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the left top. The image shows a group of green capsules lying on a textured gray background. These capsules appear to be uniform in size and shape, with a glossy finish. Upon inspection, there is a visible defect on one of the capsules positioned at the top left of the image. The type of defect present appears to be an anomaly on the surface of the capsule, which could be categorized as an inclusion or bubble within the capsule shell. The defect manifests as a small, irregularly shaped area that stands out from the otherwise smooth and consistent exterior of the capsule. The rest of the capsules in the image do not exhibit any apparent defects or anomalies.</think>
<answer>A</answer>



Figure 3: Qualitative results on MMAD [25]. The multimodal reasoning process facilitates accurate anomaly detection and question analysis. The ground truth mask and selected choice are indicated by the blue line.

of 16 for full parameter fine-tuning. For SFT training, the initial learning rate is set to 1e-4, with a weight decay of 0.03. For GRPO training, the initial learning rate is set to 1e-6, the weight decay is 0.1, and the generation number is set to 16 per training step.

Benchmarks. As OmniAD serves as a unified anomaly detection and understanding method, we conduct a comprehensive performance evaluation for both tasks. In the anomaly understanding evaluation, we use MMAD [25] as the benchmark, which evaluates seven subtasks: Anomaly Discrimination, Defect Classification, Defect Localization, Defect Description, Defect Analysis, Object Classification, and Object Analysis through multiple-choice questions. For anomaly detection, we utilize four datasets: MVTec-AD [5], VisA [59], MVTec-LOCO [4], and GoodsAD [53]. The first two are widely used, as they mainly contain textural anomalies, while the latter two were recently designed for evaluating structural and logical anomalies.

Evaluation Metrics. For anomaly detection, to evaluate the impact of the threshold on robust detection results, we calculate threshold-dependent metrics such as F1-score and accuracy at both the pixel and image levels. We follow the approach outlined in Anomalib [1], where the threshold is optimized to maximize the harmonic mean of pixel-level and image-level across the entire dataset. Both pixel-level and image-level metrics are then computed using the same threshold. Note that OmniAD does not require threshold selection and all experimental results are obtained without any cost for threshold selection. For anomaly understanding, we calculate the accuracy of correct selections.

Inference Setting. For all experiments, the few-shot inference settings follow previous works [16, 24]. A 0-shot setting means only a query image is provided, while a 1-shot setting includes an additional normal image from the same category as the query image for reference. We test the method in all settings it supports.

Table 3: Comparison of pixel-level anomaly detection performance in both 0-shot and 1-shot settings. Detection results are report as (F1-score, Acc).

Methods	MVTec-AD		VisA		MVTec-LOCO		GoodsAD		Average	
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
WinCLIP [24]	19.0, 90.7	19.8, 90.1	8.4, 98.1	12.5, 98.8	10.4, 18.5	10.4, 85.2	10.6, 94.4	10.1, 88.7	12.1, 75.4	13.2, 90.7
AnomalyGPT [16]	23.4, 92.9	42.7, 96.1	10.8, 98.2	32.0 , 98.8	13.6, 81.8	24.7, 86.5	7.2, 88.9	11.2, 79.1	13.8, 90.5	27.7, 90.1
AnomalyCLIP [57]	34.1, 95.9	-	17.8, 97.8	-	15.9 , 61.1	-	16.8, 96.9	-	21.2, 71.1	-
AdaCLIP [8]	30.2, 95.5	-	32.0 , 99.3	-	11.3, 6.0	-	7.9, 94.8	-	20.3, 73.9	-
AA-CLIP [39]	35.4, 95.4	-	4.7, 98.3	-	15.6, 79.9	-	12.5, 94.6	-	17.1, 92.1	-
UniVAD [17]	-	42.7, 96.0	-	27.2, 99.1	-	28.1 , 88.4	-	14.2, 94.1	-	28.1 , 94.4
OmniAD	45.6 , 96.9	52.1 , 97.6	18.6, 98.6	29.5, 99.1	14.4, 89.9	14.7, 91.7	20.4 , 96.4	26.7 , 97.4	23.3 , 95.9	27.9 , 96.8

Table 4: Comparison of image-level anomaly detection performance in both 0-shot and 1-shot settings. Detection results are report as (F1-score, Acc).

Methods	MVTec-AD		VisA		MVTec-LOCO		GoodsAD		Average	
	0-shot	1-shot								
WinCLIP [24]	74.5, 65.8	70.1, 62.7	64.7, 62.2	57.4, 61.3	77.3, 63.2	75.5, 62.0	57.9, 53.0	68.3, 55.2	68.6, 61.0	67.8, 60.3
AnomalyGPT [16]	83.2, 72.8	89.8, 84.9	66.6, 56.5	81.9, 77.5	77.2, 63.0	77.3, 63.2	71.1, 55.3	71.1, 55.4	74.5, 61.9	80.0, 70.2
AnomalyCLIP [57]	86.7, 79.9	-	75.4, 64.5	-	77.3, 63.2	-	71.1, 55.4	-	77.6, 65.7	-
AdaCLIP [8]	81.9, 73.8	-	78.7, 75.5	-	77.3, 63.2	-	71.1, 56.2	-	77.2, 67.1	-
AA-CLIP [39]	85.2, 77.6	-	66.4, 54.2	-	77.2, 63.1	-	71.5, 56.5	-	75.0, 62.8	-
UniVAD [17]	-	88.5, 82.7	-	79.4, 75.5	-	77.6, 64.1	-	71.3, 56.5	-	79.2, 69.7
OmniAD	97.6 , 96.5	97.2 , 96.0	89.2 , 87.4	87.5 , 86.6	92.6 , 90.3	93.2 , 91.1	90.0 , 87.9	90.1 , 88.4	92.2 , 90.1	92.0 , 91.1

4.2 Main Results

Anomaly Understanding. Table 2 presents the anomaly understanding results of OmniAD compared with several proprietary and open-source MLLMs. OmniAD achieves consistently strong performance across all question types, substantially outperforming the other methods. Compared to the baseline model Qwen2.5-VL-7B, OmniAD shows significant improvements in Defect Classification, Defect Localization, and Defect Description, demonstrating the effectiveness of enhanced anomaly perception capabilities with our Multimodal Reasoning. Although AnomalyGPT is a typical example of an MLLM tailored for anomaly detection, it performs poorly in understanding tasks, showing a significant performance gap even when compared to general-purpose models. This may be due to catastrophic forgetting when training the MLLM to leverage the results of expert models. Compared to another domain-specific method, AnomalyR1, which directly incorporates reinforcement learning for anomaly understanding, OmniAD outperforms it in Defect Localization. This suggests that our Multimodal Reasoning approach provides a more accurate perception of anomalies, leading to better image analysis. In terms of average accuracy, OmniAD performs comparably with ordinary human annotators and even surpasses them in Defect Analysis and Object Classification. Moreover, our paradigm demonstrates greater robustness in leveraging reference information to improve performance, as directly applying 1-shot learning to a general model typically causes performance degradation [25]. To provide more intuitive insight, we visualize OmniAD’s results in Figure 3.

Anomaly Detection. To evaluate the effectiveness and potential of our multimodal reasoning, not only for anomaly understanding but also for accurate detection, we compare the anomaly detection performance with six baseline models in both 0-shot and 1-shot settings. Table 3 presents pixel-level results, where OmniAD demonstrates performance on par with current state-of-the-art methods, even outperforming them in the 0-shot setting. This supports the effectiveness of our Text-as-Mask Encoding in achieving robust detection results. Table 4 provides image-level results, where OmniAD shows strong performance across all datasets, benefiting from its threshold-free ap-

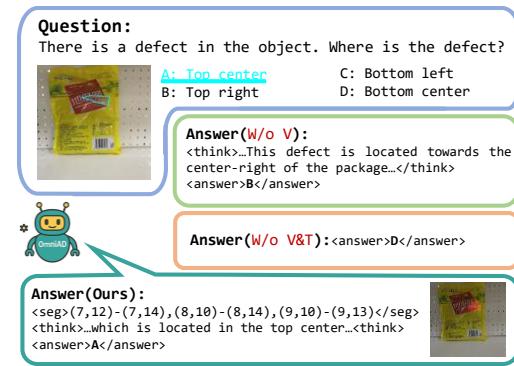


Figure 4: Case Study of Multimodal Reasoning.

Table 5: Ablations on different training strategies. Detection results are report as (F1-score, Acc).

Training Strategy	Understanding						Detection	
	Def Loc.	Def Ana.	Def Des.	Obj Ana.	Avg.	Pixel-level	Image-level	
SFT	64.6	77.7	69.1	76.6	69.8	15.0, 93.4	86.4, 81.7	
GRPO	47.6	64.8	48.9	71.2	55.7	18.8, 97.0	88.3, 85.9	
SFT+GRPO	73.5	87.5	81.7	86.9	79.1	23.3, 95.9	92.2, 90.1	

Table 6: Ablation study on Multimodal Reasoning. ‘Def. Loc.’ and ‘Obj. Ana.’ refer to Defect Localization and Object Analysis, respectively. ‘V’ and ‘T’ represent visual reasoning and textual reasoning.

Method	Understanding		
	Def. Loc.	Obj. Ana.	Avg.
w/o V&T	64.9	72.2	68.0
w/o V	71.2	83.6	77.4
All	73.5	86.9	79.1

Table 7: Ablation Study on Reward Function. ARR refers to Answer Accuracy Reward and DAR refers to Detection Accuracy Reward. P-level refers to pixel-level detection and I-level refer to image-level detection. Detection results are report as (F1-score, Acc).

Reward	Understanding		Detection	
	Avg.	P-level	I-level	
w/o AAR	70.7	17.4, 94.0	92.3, 90.2	
w/o DAR	78.3	21.4, 95.6	92.4, 90.5	
All	79.1	23.3, 95.9	92.2, 90.1	

proach. Notably, many baseline models perform poorly on MVTec-LOCO and GoodsAD, almost classifying all samples as normal. This is because these methods are mainly designed for detecting textural anomalies, whereas the two datasets primarily contain structural and logical anomalies. While textural anomalies can be detected by inspecting low-level features, structural and logical anomalies are closely related to understanding and analysis capabilities, which results in poor generalization for those models. In contrast, OmniAD performs detection and analysis together, overcoming this limitation and achieving outstanding performance on both datasets.

4.3 Ablation Study

Training Strategy. We first evaluate the effect of different training strategies for both understanding and detection, shown in Table 5. Our integrated SFT and GRPO strategy significantly outperforms training with solely SFT or GRPO, leading to a substantial improvement. It is important to note that training solely with GRPO results in considerable performance degradation, as it may introduce confusion in reasoning exploration and make it difficult for the model to learn new knowledge using only reinforcement learning. These results align with recent findings [45], which suggest that GRPO enhances sampling efficiency for obtaining correct results.

Multimodal Reasoning. We examine the effects of visual reasoning and textual reasoning in the multimodal reasoning process. Table 6 demonstrates that multimodal reasoning leads to a significant performance improvement. While textual reasoning plays a crucial role in providing basic reasoning capabilities, visual reasoning offers more detailed visual perception, which assists the model in anomaly localization and analysis. An example, illustrated in Figure 4, shows how, with the help of visual reasoning, the model successfully localizes the anomaly region that the base model and textual reasoning alone cannot.

Reward Function. In Table 7, we analyze the effect of the reward function. Both rewards enhance the detection and understanding of performance. We find that although the Detection Accuracy Reward improves detection performance, the Answer Accuracy Reward has a greater impact, not only on understanding but also on detection. This suggests that a correct answer is derived from accurate detection and increases the importance of accurate anomaly perception capabilities.

5 Conclusion

This paper introduces OmniAD, a novel framework that unifies anomaly detection and understanding. OmniAD utilizes a multimodal reasoning process to enhance the anomaly cognitive and analysis capabilities of MLLMs. The Multimodal Reasoning consists of visual reasoning, which performs anomaly detection through our proposed Text-as-Mask Encoding, and textual reasoning, which provides further analysis to ensure correct results. By leveraging an SFT and GRPO integrated

training strategy, OmniAD demonstrates excellent performance in both anomaly detection and understanding tasks with few-shot training data.

Limitations. Our Text-as-Mask Encoding can accurately localize the anomaly region, but the patch label still introduces errors at the edges compared to the mask. Additionally, the model is not efficient enough for practical applications, as MLLM inference requires significant computational resources.

References

- [1] Samet Akcay, Dick Amelin, 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.
- [2] Akshatha Arodi, Margaux Luck, Jean-Luc Bedwani, Aldo Zaimi, Ge Li, Nicolas Pouliot, Julien Beaudry, and Gaétan M Caron. Cableinspect-ad: An expert-annotated anomaly detection dataset. *Advances in Neural Information Processing Systems*, 37:64703–64716, 2024.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] 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.
- [5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtect 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.
- [6] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.
- [7] Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.
- [8] 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, 2024.
- [9] Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection. *arXiv preprint arXiv:2504.11914*, 2025.
- [10] 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.
- [11] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.
- [12] Zhiling Chen, Hanning Chen, Mohsen Imani, and Farhad Imani. Can multimodal large language models be guided to improve industrial anomaly detection? *arXiv preprint arXiv:2501.15795*, 2025.
- [13] 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.
- [14] Lei Fan, Junjie Huang, Donglin Di, Anyang Su, Maurice Pagnucco, and Yang Song. Revitalizing reconstruction models for multi-class anomaly detection via class-aware contrastive learning. *arXiv preprint arXiv:2412.04769*, 2024.
- [15] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023.
- [16] 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*, volume 38, pages 1932–1940, 2024.

- [17] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. *arXiv preprint arXiv:2412.03342*, 2024.
- [18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [19] 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. *arXiv preprint arXiv:2405.14325*, 2024.
- [20] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2404.06564*, 2024.
- [21] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020.
- [22] Jeeho Hyun, Sangyun Kim, Giyoung Jeon, Seung Hwan Kim, Kyunghoon Bae, and Byung Jun Kang. Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2052–2061, 2024.
- [23] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Hel-
yar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [24] 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.
- [25] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [26] Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022.
- [27] Yuqi Jiang, Xudong Lu, Qian Jin, Qi Sun, Hanming Wu, and Cheng Zhuo. Fabgpt: An efficient large multimodal model for complex wafer defect knowledge queries. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pages 1–8, 2024.
- [28] Er Jin, Qihui Feng, Yongli Mou, Gerhard Lakemeyer, Stefan Decker, Oliver Simons, and Johannes Stegmaier. Logicad: Explainable anomaly detection via vlm-based text feature extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4129–4137, 2025.
- [29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- [30] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024.
- [31] 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.
- [32] Weijia Li, Guanglei Chu, Jiong Chen, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. Lad-reasoner: Tiny multimodal models are good reasoners for logical anomaly detection. *arXiv preprint arXiv:2504.12749*, 2025.
- [33] 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.
- [34] Yuanze Li, Shihao Yuan, Haolin Wang, Qizhang Li, Ming Liu, Chen Xu, Guangming Shi, and Wangmeng Zuo. Triad: Empowering lmm-based anomaly detection with vision expert-guided visual tokenizer and manufacturing process. *arXiv preprint arXiv:2503.13184*, 2025.

- [35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
- [37] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.
- [38] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [39] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. *arXiv preprint arXiv:2503.06661*, 2025.
- [40] 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.
- [41] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.
- [42] 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.
- [43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [44] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- [45] Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.
- [46] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [47] 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.
- [48] Enquan Yang, Peng Xing, Hanyang Sun, Wenbo Guo, Yuanwei Ma, Zechao Li, and Dan Zeng. 3cad: A large-scale real-world 3c product dataset for unsupervised anomaly. *arXiv preprint arXiv:2502.05761*, 2025.
- [49] Shuai Yang, Zhifei Chen, Pengguang Chen, Xi Fang, Yixun Liang, Shu Liu, and Yingcong Chen. Defect spectrum: a granular look of large-scale defect datasets with rich semantics. In *European Conference on Computer Vision*, pages 187–203. Springer, 2024.
- [50] 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.
- [51] 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.
- [52] Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *arXiv preprint arXiv:2303.08730*, 2023.

- [53] 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.
- [54] 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.
- [55] Yiheng Zhang, Yunkang Cao, Xiaohao Xu, and Weiming Shen. Logicode: an llm-driven framework for logical anomaly detection. *IEEE Transactions on Automation Science and Engineering*, 2024.
- [56] Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry*, 151:103990, 2023.
- [57] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.
- [58] Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Junran Wu. Do llms understand visual anomalies? uncovering llm’s capabilities in zero-shot anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 48–57, 2024.
- [59] 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.

A Effects of Expert Model.

Table 8: Performance comparison of various visualization types using the expert model and ground truth (GT). The baseline model is Qwen2.5-VL-7B [3], while AnomalyCLIP [57] is used as the expert model.

Type	Anomaly		Defect				Object		Average
	Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis		
Expert	-	55.9	39.8	45.8	54.8	73.7	93.3	84.7	64.0
	bbox	-1.5	+2.2	+6.6	+2.3	-2.5	+1.7	-3.1	+0.8
	contour	-1.2	-1.1	+6.1	+0.7	-3.2	+1.6	-3.2	0.0
	heatmap	-8.0	-5.5	+7.6	-2.7	+1.9	-2.4	-5.7	-2.1
	highlight	-6.7	-14.7	+6.5	-11.0	-3.3	+1.6	-3.4	-4.4
	mask	-6.0	-2.4	+3.8	-2.7	+0.1	+1.6	-4.3	-1.4
GT	bbox	+14.7	+8.5	+26.2	+9.4	-1.1	+2.0	-2.7	+8.2
	contour	+16.4	+7.7	+23.9	+9.7	-2.1	+1.8	-2.4	+7.9
	highlight	+17.3	-3.6	+17.3	+0.3	-2.3	+1.4	-2.5	+4.0
	mask	+9.8	+9.2	+5.7	+4.4	-2.4	+1.1	-2.4	+3.6
Ours		+5.5	+28.0	+27.7	+26.9	+13.8	+1.9	+2.2	+15.1

With recent advances in anomaly detection models, the classification and localization of anomalies have been significantly improved. Therefore, we employ an anomaly detection model as an expert to provide additional information about anomalies, as MLLMs tend to struggle with anomaly detection. Table 8 presents the results of using an expert model or being directly guided by ground truth (GT). Considering the different types of visualizations, the expert model only improves Anomaly Localization but leads to a decrease in Anomaly Discrimination, Defect Classification, and Object Analysis, resulting in a drop in average accuracy. In contrast, using GT significantly improves the average accuracy, with only a slight decrease in Defect and Object Analysis, likely due to occlusions in the visualizations. This also suggests that MLLMs have varying responses to different visualization types, with the bounding box approach proving to be the most effective overall. By leveraging Multimodal Reasoning, OmniAD not only enhances anomaly perception but also mitigates performance degradation, resulting in an improvement in comprehensive capabilities.

B An analysis of threshold selection.

Table 9: Comparison of the optimal threshold, with the threshold range spanning from 0 to 100.

Methods	MVTec-AD		VisA		MVTec-LOCO		GoodsAD	
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
WinCLIP [24]	28.8	45.4	41.5	67.4	8.7	47.4	33.9	48.2
AnomalyGPT [16]	0.7	47.0	41.4	44.4	0.7	32.3	3.0	56.4
AnomalyCLIP [57]	63.6	-	95.6	-	1.0	-	42.4	-
AdaCLIP [8]	1.0	-	12.1	-	0.2	-	2.0	-
AA-CLIP [39]	37.3	-	10.1	-	21.2	-	35.3	-
UniVAD [17]	-	58.5	-	58.7	-	48.5	-	66.1

Table 9 presents the optimal threshold used to calculate the metrics. There are significant variations across different datasets, which requires manual selection of anomaly samples—an impractical approach. As shown in Table 8, the bounding box (bbox) is the most suitable visualization type. However, this unstable threshold can make it difficult for MLLMs to interpret the expert output, resulting in poor performance. In contrast, OmniAD employs a visual reasoning process that does not require a threshold, leading to more robust and reliable results.

C More Quantitative Results for Each Category on The MVTec-AD Dataset.

Table 10 and Table 11 respectively present the results of image-level anomaly detection and pixel-level anomaly localization quantitative outcomes across all categories within the MVTec-AD dataset. The results further demonstrate the superiority of our method over various SoTA approaches.

Table 10: Comparison with SoTA methods on **MVTec-AD** dataset for pixel-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Bottle	32.15/94.82	21.00/94.70	42.5/93.6	73.9 /97.0	44.5/90.7	-	57.9/96.1	-	37.0/95.4	-	-	-	73.9 / 97.1	54.6/93.4	52.6/ 95.2
Cable	5.70/70.83	8.57/86.22	14.5/88.7	51.1/96.6	10.6/96.7	-	26.2/95.7	-	22.4/91.7	-	-	-	49.8/96.3	47.1/96.3	53.9 / 97.1
Capsule	6.18/98.07	4.27/98.45	12.6/95.9	14.7/97.3	30.7/98.6	-	17.4/ 99.1	-	34.8 /98.8	-	-	-	32.6/ 99.1	29.5/98.4	31.9/98.8
Hazelnut	32.95/92.85	27.82/89.17	22.0/92.0	50.3/95.6	43.8/96.1	-	27.3/89.4	-	37.6/93.9	-	-	-	57.3 / 97.1	54.9/97.6	51.2/ 98.2
Metal Nut	6.70/77.09	22.21/66.51	50.6 /87.4	31.1/89.1	16.4/87.4	-	15.8/83.4	-	16.9/88.8	-	-	-	24.2/89.3	50.1/96.8	50.5/ 97.3
Pill	26.98/92.17	41.03/94.52	17.7/94.3	48.1 / 97.2	34.2/96.0	-	23.1/96.9	-	30.4/96.2	-	-	-	40.4/97.1	40.3/97.8	43.9/ 98.5
Screw	7.19/79.09	5.13/95.92	21.5/98.9	28.4/99.4	-	13.3/97.3	-	6.7/93.5	-	-	-	-	32.7 / 99.6	14.0/99.0	17.6/99.4
Toothbrush	11.99/84.90	9.89/72.36	16.5/85.3	45.5/96.2	25.3/97.6	-	24.9/94.6	-	21.2/90.9	-	-	-	46.2 / 97.6	23.0/95.9	25.1/ 98.0
Transistor	17.02/79.21	20.98/84.44	12.2/79.6	19.8 / 95.5	14.9/94.3	-	12.1/94.2	-	14.8/93.0	-	-	-	33.1/94.2	19.3/91.8	46.2 / 94.2
Zipper	8.93/97.09	10.56/97.89	17.8/96.6	39.0/97.9	35.7/98.2	-	14.4/98.1	-	42.9 / 98.4	-	-	-	42.7/97.7	35.8/97.2	36.4/97.4
Carpet	33.11/96.81	32.40/98.66	11.3/98.3	62.4/98.0	54.7/97.8	-	52.5/98.9	-	72.4 / 99.1	-	-	-	44.2/96.0	42.2/97.1	54.8/98.5
Grip	6.21/93.98	6.53/96.51	10.9/99.1	33.3/97.1	31.6/98.1	-	28.8 / 99.3	-	34.9 /99.2	-	-	-	28.2/96.8	26.1/98.3	26.4/98.8
Leather	30.90/98.07	35.52/99.06	49.6 / 99.4	32.1/97.0	26.0/96.5	-	45.4/98.8	-	46.9/99.1	-	-	-	21.9/95.4	42.3/99.0	53.9 / 99.2
Tile	21.25/90.09	4.52/92.84	31.1/93.6	68.1/94.2	61.5/95.9	-	40.7/94.6	-	59.2/95.7	-	-	-	62.3/92.9	68.9/95.4	70.3 / 95.8
Wood	38.12/95.36	41.93/95.98	33.5/96.5	47.6/92.1	54.8/95.4	-	52.8/96.9	-	-	-	-	-	50.3/93.8	52.7/95.8	55.2 / 97.0
Mean	19.03/90.67	19.49/90.88	23.5/92.9	42.6/96.0	34.2/95.8	-	30.2/95.5	-	35.4/95.4	-	-	-	42.7/96.0	45.6/96.9	52.1 / 97.7

Table 11: Comparison with SoTA methods on **MVTec-AD** dataset for image-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Bottle	97.6/96.4	90.4/86.8	86.3/75.9	98.4 /97.6	86.3/75.9	-	90.0/83.1	-	90.4/86.8	-	-	-	98.4 / 97.6	98.2/98.2	97.3/ 98.7
Cable	76.0/61.3	76.0/61.3	76.0/61.3	81.1/73.3	80.0/72.7	-	76.0/61.3	-	75.5/60.7	-	-	-	78.0/65.3	95.0 / 91.3	95.8/ 93.4
Capsule	89.5/81.1	88.1/78.8	90.5/82.6	87.3/79.6	93.3/89.4	-	62.0/54.6	-	89.4/81.8	-	-	-	83.9/76.5	97.3 / 95.4	96.8/ 99.2
Hazelnut	77.8/63.6	77.8/63.6	86.4/80.0	80.5/69.1	-	77.8/63.6	-	78.2/64.6	-	-	-	80.5/69.1	96.6/ 93.3	96.8 / 93.8	
Metal Nut	98.4/80.9	89.4/80.9	89.4/80.9	97.9/96.5	90.5/83.5	-	87.8/78.3	-	82.1/73.9	-	-	-	96.8/94.8	100.0 / 100.0	100.0 / 100.0
Pill	91.6/84.4	91.6/84.4	91.6/84.4	94.2/90.4	93.0/87.4	-	85.6/77.3	-	89.7/81.4	-	-	-	96.4/94.0	99.1 / 98.2	98.6/ 99.4
Screw	85.3/74.4	85.3/74.4	85.3/74.4	82.8/75.0	87.9/81.9	-	85.3/74.4	-	85.3/74.4	-	-	-	81.9/75.6	95.2 / 92.0	90.0/90.9
Toothbrush	83.3/71.4	83.3/71.4	83.3/71.4	87.9/81.0	81.7/69.1	-	83.3/71.4	-	83.3/71.4	-	-	-	87.0/78.6	98.8 / 100.0	96.2/ 100.0
Transistor	57.1/40.0	57.1/40.0	57.1/40.0	80.0/84.0	59.3/45.0	-	57.1/40.0	-	55.1/38.0	-	-	-	62.5/52.0	89.3/81.8	97.4 / 94.9
Zipper	90.0/82.8	93.2/88.7	88.2/77.8	85.0/79.5	77.3/70.9	-	66.3/60.3	-	88.2/82.8	-	-	-	90.4/83.4	96.4/95.5	98.1 / 97.5
Carpet	86.4/76.1	91.2/86.3	86.4/76.1	98.3/97.4	97.3/95.7	-	92.1/88.9	-	97.1/95.7	-	-	-	94.2/90.6	99.6 / 99.1	99.1/ 100.0
Grid	84.4/73.1	81.6/70.5	84.4/73.1	93.4/89.7	96.4/94.9	-	80.4/75.6	-	81.3/76.9	-	-	-	99.1 / 98.7	99.3 / 100.0	96.4/ 100.0
Leather	85.2/74.2	89.3/82.3	85.2/74.2	98.9/98.4	91.5/86.3	-	99.5 / 99.2	-	99.5 / 99.2	-	-	-	98.9/98.4	100.0 / 100.0	100.0 / 100.0
Tile	83.6/71.8	89.4/84.6	83.6/71.8	98.3/97.4	95.3/93.2	-	91.8/88.9	-	88.3/84.6	-	-	-	88.4/81.2	100.0 / 100.0	99.5/ 100.0
Wood	86.3/76.0	88.9/81.0	86.3/76.0	92.3/87.3	90.9/84.8	-	93.3/89.9	-	94.0/91.1	-	-	-	90.9/84.8	100.0 / 100.0	98.7/ 100.0
Mean	84.2/73.8	84.9/75.7	83.4/72.3	90.8/87.2	86.7/80.0	-	81.9/73.8	-	85.2/77.6	-	-	-	88.5/82.7	97.7 / 96.6	97.2/96.0

D More Quantitative Results for Each Category on The MVTec-LOCO Dataset.

Table 12 and Table 13 respectively present the results of image-level anomaly detection and pixel-level anomaly localization quantitative outcomes across all categories within the MVTec-LOCO dataset. The results further demonstrate the superiority of our method over various SoTA approaches.

Table 12: Comparison with SoTA methods on **MVTec-LOCO** dataset for pixel-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Breakfast box	11.9/9.2	12.6/83.5	8.6/74.3	34.7/83.0	16.3/43.5	-	12.5/6.7	-	19.2/70.8	-	-	-	39.8 / 93.0	18.1/76.0	22.4/ 83.1
Juice bottle	16.8/50.6	22.7/84.5	23.5/79.9	41.1 / 92.0	25.4/74.6	-	10.6/5.6	-	21.6/81.6	-	-	-	30.4 / 93.3	17.0/92.7	19.8/93.1
Pushpins	1.6/2.4	1.1/80.8	3.0/92.7	10.5/93.6	3.3/59.0	-	5.5/2.8	-	4.5/91.4	-	-	-	33.4 / 94.3	6.5/95.7	5.5 / 96.6
Screw bag	7.8/16.6	2.0 / 95.2	17.2/79.5	18.0 / 76.9	17.3/52.7	-	13.8/7.4	-	16.6/73.1	-	-	-	17.3/77.0	13.9/90.7	7.8/91.2
Splicing connectors	13.8/13.9	13.7/81.8	15.8/82.9	19.3/87.5	17.4/76.1	-	14.0/7.5	-	16.0/82.7	-	-	-	19.6 / 84.3	5.2/92.2	4.5 / 92.6
Mean	10.4/18.5	10.4/85.2	13.6/81.9	24.7/86.6	15.9/61.2	-	11.3/6.0	-	15.6/79.9	-	-	-	28.1 / 88.4	14.4/89.9	14.8/ 91.7

Table 13: Comparison with SoTA methods on **MVTec-LOCO** dataset for image-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Breakfast box	77.2 / 62.9	77.2 / 62.9	77.2 / 62.9	77.2 / 62.9	-	77.2 / 62.9	-	77.2 / 62.9	-	-	-	77.8/64.0	93.7/88.4	94.3 / 98.9	
Juice bottle	83.4/71.5	83.4/71.5	83.4/71.5	83.4/71.5	83.4/71.5	-	83.4/71.5	-	83.4/71.5	-	-	-	84.7/75.8	94.9 / 91.2	94.6/ 98.0</b

Table 14: Comparison with SoTA methods on **VisA** dataset for pixel-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Candle	5.4/99.7	4.4/99.8	21.3/99.8	32.7/99.7	18.8/99.1	-	43.4/99.9	-	0.8/98.9	-	-	-	31.9/99.7	14.6/99.3	19.4/99.6
Capsules	0.2/99.4	1.9/99.6	10.4/99.2	49.4/99.5	21.0/98.6	-	51.1/99.7	-	6.3/99.0	-	-	-	42.2/99.5	21.2/99.0	31.4/99.5
Cashew	6.6/95.5	8.5/98.4	9.4/98.8	30.2/98.7	19.4/98.8	-	19.0/98.9	-	5.7/96.0	-	-	-	28.8/99.0	23.7/98.8	44.7/98.9
Chewinggum	29.3/98.9	41.8/99.5	19.0/98.2	44.6/98.3	46.5/98.8	-	67.2/99.5	-	27.7/99.4	-	-	-	45.5/98.3	43.1/99.0	52.4/99.2
Fryum	14.1/94.4	22.3/95.8	10.3/97.1	31.7/97.3	24.5/97.6	-	17.6/98.1	-	3.3/98.2	-	-	-	23.7/98.1	16.1/96.3	18.2/97.3
Macaroni1	0.4/98.6	1.2/98.1	6.1/99.9	25.3/99.9	3.1/98.4	-	28.9/99.9	-	0.2/96.6	-	-	-	13.1/99.9	3.7/99.6	4.5/99.8
Macaroni2	0.1/99.1	0.5/99.9	0.4/99.8	14.7/99.7	0.5/89.9	-	4.2/99.2	-	0.3/99.3	-	-	-	4.9/99.7	1.6/99.5	1.7/99.7
Pcb1	0.0/98.8	1.7/98.9	2.7/98.5	25.4/97.0	12.6/99.2	-	21.5/99.4	-	1.8/97.8	-	-	-	44.3/99.2	3.6/98.2	4.5/98.7
Pcb2	0.4/97.6	3.5/99.3	6.5/98.3	23.0/99.5	8.2/98.4	-	31.6/99.7	-	0.0/98.8	-	-	-	16.3/99.5	3.8/98.0	9.7/98.8
Pcb3	1.1/98.6	2.5/99.2	8.8/98.3	38.0/99.4	9.1/99.2	-	27.2/99.4	-	1.4/98.6	-	-	-	23.4/99.6	12.1/99.0	12.2/99.3
Pcb4	21.2/98.7	22.5/98.1	17.5/96.1	20.2/98.3	24.5/97.4	-	42.7/99.3	-	0.2/97.7	-	-	-	17.5/97.5	27.5/98.2	33.6/98.7
Pipe fryum	21.3/97.8	39.3/98.5	17.3/97.7	49.2/98.3	25.7/98.6	-	29.0/98.7	-	8.1/98.6	-	-	-	34.8/98.5	36.7/98.6	64.9/99.0
Mean	8.4/98.1	12.5/98.8	10.8/98.3	32.0/98.8	17.8/97.8	-	32.0/99.3	-	4.7/98.3	-	-	-	27.2/99.1	18.6/98.7	29.5/99.1

Table 15: Comparison with SoTA methods on **VisA** dataset for image-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Candle	47.8/65.0	23.0/56.5	48.7/60.0	85.3/86.5	67.6/52.5	-	71.8/78.0	-	66.7/50.0	-	-	-	90.8/91.0	95.3/93.3	92.2/95.5
Capsules	32.6/45.6	26.4/44.4	75.0/60.0	83.6/79.4	77.2/63.1	-	83.2/80.6	-	76.9/62.5	-	-	-	87.4/85.5	91.3/91.6	84.2/91.6
Cashew	80.8/68.7	77.0/69.3	56.0/54.0	87.9/84.0	85.6/82.7	-	70.1/69.3	-	80.0/66.7	-	-	-	78.8/76.7	84.1/88.2	84.1/93.8
Chewinggum	70.5/69.3	55.1/58.7	78.5/64.7	95.1/93.3	94.5/92.7	-	95.4/94.0	-	80.2/67.3	-	-	-	94.4/92.7	96.4/95.3	94.6/94.2
Fryum	80.9/71.3	82.5/75.0	77.1/64.3	93.5/91.3	88.8/84.3	-	86.6/82.3	-	80.0/66.7	-	-	-	89.4/86.3	92.6/92.6	88.5/95.7
Macaroni1	61.3/54.0	66.7/52.0	60.3/60.5	82.7/84.5	64.9/48.0	-	73.5/78.0	-	66.7/50.0	-	-	-	66.3/68.5	80.8/74.3	77.8/79.8
Macaroni2	58.1/56.0	8.8/48.0	56.7/46.5	66.9/58.0	66.2/49.5	-	67.1/52.0	-	66.7/50.0	-	-	-	64.5/55.5	78.4/73.1	69.3/70.0
Pcb1	59.3/56.0	74.6/72.5	66.7/50.0	71.9/61.0	67.9/55.5	-	70.9/63.0	-	66.7/50.0	-	-	-	77.2/71.0	90.7/84.1	89.4/85.5
Pcb2	68.1/57.0	54.5/53.5	66.7/50.0	71.1/63.5	68.7/54.5	-	76.1/74.5	-	66.7/50.0	-	-	-	73.8/67.0	87.8/78.3	90.7/85.3
Pcb3	59.7/53.7	58.4/53.2	66.4/49.7	71.5/63.2	63.8/49.7	-	65.7/52.7	-	66.5/49.8	-	-	-	68.4/65.7	88.0/78.5	91.2/84.3
Pcb4	74.5/73.1	75.4/69.1	66.4/49.7	78.7/73.1	67.3/51.7	-	91.7/91.0	-	66.5/49.8	-	-	-	67.4/53.7	91.0/83.4	96.4/94.6
Pipe fryum	82.6/76.7	86.0/83.3	81.3/69.3	94.7/92.7	92.6/90.7	-	92.5/90.0	-	80.0/66.7	-	-	-	94.3/92.0	96.3/98.1	85.9/98.8
Mean	64.7/62.2	57.3/61.3	66.7/56.6	81.9/77.5	75.4/64.6	-	78.7/75.5	-	71.9/56.6	-	-	-	79.4/75.5	89.3/87.5	87.5/86.6

F More Quantitative Results for Each Category on The GoodsAD Dataset.

Table 16 and Table 17 respectively present the results of image-level anomaly detection and pixel-level anomaly localization quantitative outcomes across all categories within the GoodsAD dataset. The results further demonstrate the superiority of our method over various SoTA approaches.

Table 16: Comparison with SoTA methods on **GoodsAD** dataset for pixel-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Cigarette box	18.8/86.6	15.7/72.4	10.5/83.5	24.7/82.7	20.0/94.5	-	4.3/93.9	-	10.5/93.7	-	-	-	23.6/94.6	24.8/93.2	31.1/94.9
Drink bottle	2.8/98.8	4.5/96.9	6.0/90.0	7.8/81.9	15.8/97.9	-	11.2/96.6	-	13.7/96.4	-	-	-	15.0/96.9	14.9/97.7	23.7/98.3
Drink can	11.0/96.5	14.8/91.0	8.4/90.7	12.2/77.1	14.4/96.7	-	10.4/94.6	-	16.1/94.9	-	-	-	17.9/93.3	20.8/96.8	26.2/97.3
Food bottle	19.2/97.8	16.7/95.5	8.2/90.4	11.3/83.4	23.0/97.7	-	10.3/95.9	-	17.7/94.3	-	-	-	15.9/94.6	29.0/96.9	35.5/97.8
Food box	5.7/97.3	3.9/94.3	6.1/91.8	7.1/73.7	14.4/97.5	-	4.4/95.6	-	7.8/94.8	-	-	-	7.2/93.0	12.0/96.5	13.7/97.4
Food package	5.4/92.2	4.6/82.1	4.5/87.4	4.4/76.5	13.3/97.4	-	6.9/92.1	-	9.4/93.6	-	-	-	5.4/92.5	10.9/96.9	19.4/97.9
Mean	10.5/94.9	10.1/88.7	7.3/89.0	11.2/79.2	16.8/96.9	-	7.9/94.8	-	12.5/94.6	-	-	-	14.2/94.1	20.4/96.5	26.7/97.4

Table 17: Comparison with SoTA methods on **GoodsAD** dataset for image-level anomaly detection with F1-score/ACC.

Method → Category ↓	WinCLIP CVPR'23		AnomalyGPT AAAI'24		AnomalyCLIP ICLR'24		AdaCLIP ECCV'24		AA-CLIP CVPR'25		UniVAD CVPR'25		OmniAD (ours)		
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	
Cigarette box	72.9/57.3	72.9/57.3	72.9/57.3	73.0/57.6	72.9/57.3	-	73.0/58.7	-	72.4/57.3	-	-	-	74.2/60.4	94.4/89.7	95.8/94.1
Drink bottle	70.5/54.6	70.5/54.6	70.5/54.4	70.5/54.6	70.5/54.4	-	70.7/55.3	-	71.0/58.1	-	-	-	70.4/54.4	89.0/80.5	87.4/78.1
Drink can	66.5/49.8	66.5/49.8	66.5/49.8	66.7/50.2	66.5/49.8	-	67.0/50.9	-	69.3/56.7	-	-	-	66.7/50.5	88.2/81.4	90.1/87.9
Food bottle	74.8/59.8	74.8/59.8	74.8/59.8	74.9/59.9	74.8/59.8	-	74.7/59.8	-	75.0/60.6	-	-	-	75.1/60.4	91.0/84.3	90.3/85.8
Food box	77.5/63.2	77.5/63.2	77.5/63.2	77.5/63.2	77.5/63.2	-	77.7/63.7	-	78.0/64.5	-	-	-	77.6/63.7	92.9/88.2	93.5/90.6
Food package	64.5/47.6	64.5/47.6	64.5/47.6	64.5/47.6	64.5/47.6	-	64.5/47.6	-	65.3/49.3	-	-	-	64.4/47.6	84.9/73.9	86.1/77.2
Mean	71.1/55.4	71.1/55.4	71.1/55.4	71.2/55.5	71.1/55.4	-	71.3/56.0	-	71.8/57.8	-	-	-	71.4/56.2	90.0/87.9	90.1/88.4

G More Visualization

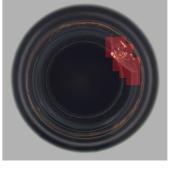


Question:
Is there any defect in the object?

A: No. **B: Yes.**

Answer:

```
<seg>(5, 15)-(5, 17), (6, 15)-(6, 18), (7, 15)-(7, 19), (8, 16)-(8, 19), (9, 16)-(9, 19), (10, 17)-(10, 19), (11, 18)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the right top. The object in the image appears to be a circular section of a bottle, typically the opening at the neck, with a smooth black interior and a copper-colored ring around the edge indicating the bottleneck's rim. Comparing this image to the standard normal sample, there is an anomaly present. The defect is a small break or chip located at the center-right edge of the bottle's rim. The blemish breaks the smooth line of the bottleneck's edge and shows irregular, jagged textures different from the rest of the smooth metallic rim.</think>
<answer>B</answer>
```





Question:
There is a defect in the object. What is the type of the defect?

A: Color fading. B: Mold.
C: Crack. **D: Irregular marking.**

Answer:

```
<seg>(6, 10)-(6, 12), (7, 9)-(7, 13), (8, 8)-(8, 13), (9, 7)-(9, 13), (10, 7)-(10, 12), (11, 7)-(11, 11), (12, 7)-(12, 10), (13, 7)-(13, 8)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the center. The object in the image appears to be a hazelnut with a smooth, shiny surface and a characteristic pointed oval shape. The hazelnut's color is a rich, warm brown, and it has the typical textured cap at the wider end where it attaches to the tree. Upon inspecting the hazelnut, there is an anomaly present. This defect is in the form of what looks like white markings or print near the center of the nut's surface. The markings contrast sharply with the natural brown hue of the hazelnut, indicating it's not a natural part of the shell. The defect is irregular in shape and is mainly concentrated in the center region, slightly towards the top of the hazelnut.</think>
<answer>D</answer>
```





Question:
There is a defect in the object. What is the type of the defect?

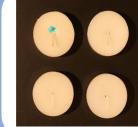
A: Tear. B: Hole.
C: Wrinkle. **D: Discoloration.**

Answer:

```
<seg>(8, 16)-(8, 17), (9, 15)-(9, 17), (10, 15)-(10, 16)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the right. The image appears to show a close-up of a textured leather surface with a rich brown color and a pattern that resembles the natural grain found in animal hides. The leather features a consistent pebbled texture throughout. Upon inspection, there is a noticeable defect present on the leather surface. The defect appears to be related to color. It is located towards the center right area of the image. This anomaly manifests as a distinct reddish mark that contrasts with the uniform brown color of the leather. The affected area stands out due to its abnormal coloration, disrupting the otherwise consistent pebbled texture of the leather's surface.</think>
<answer>D</answer>
```



Figure 5: An illustration of results in MVTec-AD dataset.



Question:

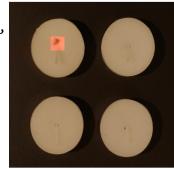
There is a defect in the object. Where is the defect?

- A: Bottom left candle. B: Bottom right candle.
[C: Top left candle.](#) D: Top right candle.



Answer:

```
<seg>(5, 6)-(5, 7), (6, 6)-(6, 7), (7, 6)-(7, 7), (8, 6)-(8, 7), (9, 7)</seg>
<think> According to the segmentation result, there is an anomaly in the object,
which is located in the left top. The query image depicts four round, white
candles arranged in a two-by-two grid pattern against a dark background. Each
candle has a wick at its center. Upon inspection, there appears to be an
anomaly on one of the candles. The top left candle shows a defect where the
wick is not properly centered or straight. It seems to be leaning or crooked
compared to the other wicks which appear to be straight and centrally
positioned. This type of defect could potentially affect the burning quality of
the candle.</think>
<answer>C</answer>
```



Question:

There is a defect in the object. What is the appearance of the defect?

- A: At the corner. B: On the edge.
[C: In the center.](#) D: Near the bottom.



Answer:

```
<seg>(11, 10)-(11, 11), (12, 10)-(12, 13), (13, 10)-(13, 13), (14, 10)-(14,
12)</seg>
<think> According to the segmentation result, there is an anomaly in the object,
which is located in the center. The object in the image appears to be a
rectangular piece placed on a textured black surface that could be seen as
resembling the texture of a leather or woven fabric. The object is uniformly
shaped with rounded edges and has a smooth, somewhat shiny surface. It seems to
be a single solid color, which is a pale, off-white or light beige. Upon
closer inspection, there is a noticeable defect on the object. This defect is
characterized by an irregularity in the center of the object. It resembles a
small, elongated indentation or mark that disrupts the uniformity of the
object's surface. The anomaly is subtle but is discernible due to its
contrasting appearance against the otherwise smooth surface of the
object.</think>
<answer>C</answer>
```



Question:

There is a defect in the object. What is the potential effect of the defect?

- A: Affects taste. [B: Affects structural integrity.](#)
C: Affects cooking time. D: Affects shelf life.



Answer:

```
<seg>(13, 14), (14, 14)-(14, 15)</seg>
<think> According to the segmentation result, there is an anomaly in the object,
which is located in the center. The object in the image appears to be a
circular item with a wheel-like design, featuring spokes connecting a central
hub to an outer rim with cut-out sections. It is placed against a green
textured background. Upon inspection, there is a noticeable defect on the
object. The defect is located at the bottom center of the object. It looks like
a small chip or breakage in one of the spokes, disrupting the uniformity of the
wheel's structure. This anomaly could potentially affect the structural
integrity of the object.</think>
<answer>B</answer>
```



Figure 6: An illustration of results in VisA dataset.

**Question:****There is a defect in the object. Where is the defect?**

- A: Top left. B: Bottom right.
C: Bottom center. D: **Top center.**

**Answer:**

```
<seg>(3, 13)-(3, 14), (4, 14)-(4, 15), (5, 15)-(5, 16), (6, 16)-(6, 17)</seg>
<think> According to the segmentation result, there is an anomaly in the object, which is located in the top center. The image depicts a food container divided into two sections. On the left side, there are three whole oranges and one peach. On the right side, there is a mix of granola, dried fruits, and nuts. Upon inspection, there appears to be an anomaly in the form of a metal object. This object is located at the top center of the image, within the granola section. It is a long, thin, metallic item that contrasts with the surrounding granola and other food items. The appearance of this object suggests it could be a piece of metal, which is not typically found in food products and could potentially pose a safety hazard.</think>
<answer>D</answer>
```

**Question:****There is a defect in the object. What is the appearance of the defect?**

- A: Uniform and consistent. B: Large and circular.
C: Faded and blurred. D: **Small and irregular in shape.**

**Answer:**

```
<seg>(4, 11)-(4, 13), (5, 11)-(5, 13)</seg>
<think>The query image depicts a clear glass bottle containing a yellow liquid, presumably juice, with a label that features an orange fruit and the text '100% Juice'. The bottle has a screw cap and appears to be sealed. Upon inspection, there is a noticeable defect on the bottle. The defect is located at the top center of the bottle, where it seems like the seal or cap might be compromised. It appears as a small, irregularly shaped area that looks different from the rest of the surface, suggesting that the seal may not be intact or properly sealed. This could potentially indicate a leak or an issue with the contents of the bottle.</think>
<answer>D</answer>
```

**Question:****There is a defect in the object. What is the appearance of the defect?**

- A: **Coiled, metallic, and darker.** B: Cylindrical body with round top.
C: Uniform with other pushpins. D: Smaller in size.

**Answer:**

```
<seg>(10, 6)-(10, 7), (11, 6)-(11, 7), (12, 6)-(12, 7)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the left. The query image depicts a clear plastic case containing various yellow pushpins, which are typically used to attach papers or notes to boards. Each pushpin consists of a yellow cylindrical head and a metal pin with a sharp point at one end. The pushpins are arranged in a grid pattern within the case. Upon inspection, there appears to be an anomaly present in the object. Specifically, there is a pushpin that looks different from the others. This pushpin is located in the bottom left section of the case. It has a coiled, metallic, and darker appearance compared to the uniform yellow color of the other pushpins. Additionally, it seems to have a cylindrical body with a round top, which is not characteristic of the standard pushpins seen in the rest of the image. </think>
<answer>A</answer>
```

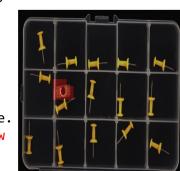
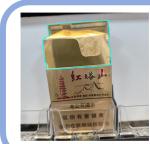


Figure 7: An illustration of results in MVTec-LOCO dataset.

**Question:**

Is there any defect in the object?

- A: Yes. B: No.

**Answer:**

```
<seg>(1, 5)-(1, 16), (2, 4)-(2, 17), (3, 4)-(3, 17), (4, 4)-(4, 17), (5, 4)-(5, 17), (6, 4)-(6, 17), (7, 4)-(7, 17), (8, 4)-(8, 17), (9, 4)-(9, 17), (10, 4)-(10, 17), (11, 4)-(11, 17), (12, 4)-(12, 17), (13, 4)-(13, 17), (14, 4)-(14, 17), (15, 4)-(15, 17), (16, 4)-(16, 17), (17, 4)-(17, 17)</seg>
<think> According to the segmentation result, there is an anomaly in the object, which is located in the top. The query image depicts a pack of cigarettes with Chinese characters on it, which suggests that this is a product from a Chinese-speaking region. The packaging features a prominent logo and text, along with an illustration of a mountain range, possibly indicating a brand or design theme. Upon inspection, there appears to be a defect/anomaly present on the cigarette pack. The defect is located at the top center of the pack. It looks like a type of damage or deformation, possibly a crease or a tear, affecting the integrity of the packaging. This kind of defect could potentially compromise the contents inside the pack.</think>
```

**Question:**

There is a defect in the object. What is the appearance of the defect?

- A: The bottle is cracked. B: The cap is not fully screwed down.
C: The label is peeling off. D: The liquid is leaking out.

**Answer:**

```
<seg>(7, 12)-(7, 13), (8, 11)-(8, 14), (9, 10)-(9, 14), (10, 11)-(10, 14), (11, 12)-(11, 14)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the center. The query image depicts a beverage bottle with a label that includes nutritional information and other details. The bottle appears to be made of clear plastic with a pinkish tint, and it has a white cap. The label features text and icons, presumably indicating the product's contents and serving suggestions. Upon inspection, there is an anomaly present on the bottle. The defect is located at the top center of the bottle, where the cap meets the neck of the bottle. It appears as though the cap is not fully screwed down, which could potentially compromise the seal of the bottle. This type of defect could lead to leakage or contamination of the contents inside.</think>
```

**Question:**

There is a defect in the object. What is the type of the defect?

- A: Structural deformation. B: Discoloration.
C: Surface damage. D: Contamination.

**Answer:**

```
<seg>(16, 12), (17, 11)-(17, 13), (18, 11)-(18, 13), (19, 11)-(19, 13), (20, 11)-(20, 13)</seg>
<think>According to the segmentation result, there is an anomaly in the object, which is located in the bottom. The object in the image is a cylindrical food container with a plastic body and a foil seal covered by a plastic lid. The container's label features non-English text and an image of the contents, which appears to be a type of tea beverage given its brown, opaque nature and association with food grading. There does appear to be a defect in the query image. The defect can be described as a surface damage located at the bottom center of the container. This anomaly manifests as a notable deformation or indentation on the surface of the plastic that is visible even without direct indication. The integrity of the container at the damaged spot may be compromised, potentially affecting the product's seal and the overall package quality.</think>
```



Figure 8: An illustration of results in GoodsAD dataset.