JUDO: A JUXTAPOSED DOMAIN-ORIENTED MULTI-MODAL REASONER FOR INDUSTRIAL ANOMALY QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial anomaly detection has been significantly advanced by large multimodal models (LMMs), enabling diverse human instructions beyond detection, particularly through visual-grounded reasoning for better image understanding. However, the lack of domain-specific knowledge of LMMs limits the accurate generation of responses in complex industrial scenarios. In this work, we present JUDO, Juxtaposed Domain-Oriented Multi-modal Reasoner, a framework that efficiently incorporates domain knowledge and context in visual and text reasoning. The visual reasoning provides detailed inspection by segmenting the defect region in the query image by juxtaposing it with the normal image as visual domain context, enabling a fine-grained visual comparative analysis. Furthermore, we inject domain knowledge through supervised fine-tuning (SFT) to enhance context understanding and subsequently guide domain reasoning through reinforcement learning (GRPO) with three tailored rewards, opting for a domain-oriented thought process. Experimental results demonstrate that JUDO achieves superior performance on the MMAD benchmark, surpassing models such as Qwen2.5-VL-7B and GPT4o. These results highlight the importance of enhancing domain knowledge and context for effective reasoning in anomaly understanding. The implementation of JUDO can be found in `https://anonymous.4open.science/r/JUDO-9C8B`.

1 INTRODUCTION

Visual anomaly detection (Jiang et al., 2025; Xu et al., 2025), which aims to identify unique anomaly patterns from images in industrial environments, has evolved significantly with the recent emergence of Large Multimodal Models (LMMs) (Hurst et al., 2024; Bai et al., 2025; Wang et al., 2025). While most anomaly patterns are specialized for domain-specific settings, current LMMs have demonstrated strong generalization capabilities in complex industrial environments, as shown by the MMAD benchmark (Jiang et al., 2025). Building upon this foundation, recent models such as AnomalyR1 (Chao et al., 2025) and OmniAD (Zhao et al., 2025) have employed Group Relative Policy Optimization (GRPO) (Shao et al., 2024) training to advance multimodal anomaly detection capabilities further. Specifically, AnomalyR1 is the first to introduce GRPO training to anomaly detection tasks, while OmniAD further advances this approach by integrating anomaly segmentation to enable visual-grounded reasoning.

Despite these advances, current GRPO-based models primarily optimize instruction-response matching, overlooking the incorporation of domain knowledge essential for robust and generalizable defect reasoning. Such domain knowledge in text generally encompasses characteristics of normal and defect samples—including their definitions, causes, and consequences, which serve as the prior knowledge necessary for reliable defect analysis. In addition, normal images play the role of visual context, offering reference information for distinguishing between normality and anomaly. However, this knowledge and visual context are often specialized and rarely encountered during the pretraining of LMMs, making reasoning based solely on the inherent generalized knowledge of LMMs insufficient. Recent studies (Jiang et al., 2025) have demonstrated that providing external knowledge (Zhao et al., 2024) or normal samples in context at inference can alleviate this limitation of LMMs, bridging this knowledge gap. However, providing knowledge and context only at inference time has limitations when LMMs lack sufficient internal knowledge, as LMMs become overly dependent on external context, leading to misaligned responses that prioritize contextual plausibility over accuracy. This limitation still stems from insufficient internalized domain knowledge and context in the model, ultimately hindering reliable and accurate domain-oriented reasoning.

To address this challenge, we propose **JUDO**, **Ju**xtaposed **D**omain-**O**riented Multi-modal Reasoner, the first approach that systematically internalizes domain knowledge for industrial anomaly detection through learning. Unlike prior GRPO-based models that apply post-training without domain alignment, JUDO unifies domain understanding across visual grounding and textual reasoning. Our framework begins in Stage 1 by establishing domain-aware visual reasoning through juxtaposed segmentation learning, where comparing normal and defect images internalizes visual context. This approach addresses the underexplored potential of normal samples, which typically serve only as optional context in inference, while JUDO incorporates it as a core reasoning context during training. In Stage 2, we enhance domain-oriented textual reasoning of LMMs by injecting domain knowledge into model parameters, unlike prior work (Jiang et al., 2025) that injects textual domain knowledge externally via prompts. Our approach builds foundational domain knowledge for industrial anomaly detection reasoning, yielding more reliable domain-aligned reasoning. Finally, Stage 3 unifies visual grounding and domain semantics through reinforcement learning (GRPO) (Shao et al., 2024) with rewards composed of domain reasoning, segmentation, choice, and structural alignment rewards, ensuring that the model produces reliable and domain-aware reasoning for anomaly understanding.

Extensive experiments on industrial anomaly detection benchmarks MMAD (Jiang et al., 2025) demonstrate the effectiveness of our approach. JUDO achieves superior performance, highlighting that internalizing domain knowledge and context during training time fundamentally matters in industrial anomaly understanding. Moreover, JUDO enhances both reliability and explainability by aligning its reasoning with the learned domain knowledge and grounding anomaly regions in visual evidence, through unified training to align domain knowledge and context.

We summarize our contributions as follows:

- We propose JUDO, the first approach to systematically internalize domain knowledge and context into both visual and textual reasoning for industrial anomaly understanding.
- We introduce a novel framework that builds domain understanding through visual segmentation and textual knowledge internalization, and unifies these capabilities via reinforcement learning with domain-aligned reward designs.
- Through comprehensive experiments, we demonstrate that JUDO achieves superior performance while enhancing explainability through segmentation-based visualizations and ensuring reliability via domain-aligned reasoning, addressing real-world industrial needs.

2 Related Work

2.1 Large Multimodal Models (LMMs)

Recent Large Multimodal Models (LMMs) have advanced visual understanding through high-quality instruction tuning (Hurst et al., 2024; Bai et al., 2025; Chen et al., 2024; Team et al., 2025; Comanici et al., 2025), stronger cross-modal architectures (Wang et al., 2025; Agrawal et al., 2024b), and more sophisticated training pipelines (Xiaomi et al., 2025; Deitke et al., 2025). Most multimodal models now support multi-image inputs (Xiaomi et al., 2025), with GPT-4o, GPT-5 (Hurst et al., 2024), the Gemini-2.5 model (Comanici et al., 2025), and InternVL3.5 (Wang et al., 2025) demonstrating improved reasoning performance through cross-image comparison. However, the majority of these models (Hurst et al., 2024; Bai et al., 2025; Wang et al., 2025) remain optimized for common question–answering tasks grounded in general knowledge. When applied to out-of-distribution scenarios such as industrial anomaly detection—where domain knowledge is crucial—their reasoning performance tends to be noticeably less accurate.

2.2 Incorporating Domain Knowledge into LMMs

To address the demand for specialized knowledge for LMMs (Song et al., 2025), specifically in fields such as finance (Qian et al., 2025), biomedicine (Liu et al., 2025), education (Agrawal et al., 2024a), and materials (Prabhakar et al., 2025), researchers have actively studied to integrate domain-specific expertise into LMMs. Most approaches fall into two categories: dynamic injection and learning-based integration through pretraining or fine-tuning. Dynamic injection-based methods provide the external knowledge in a similar way to RAG (Zhao et al., 2024) at inference time without additional training; its effectiveness is highly dependent on retrieval quality. In contrast, learning-based methods encode the domain knowledge into the model's parameters. However, this line of work remains underexplored for industrial anomaly understanding.

2.3 Industrial Anomaly Detection and LMMs

Visual anomaly detection in industrial settings has evolved significantly with deep learning. Early and still widely used methods are often unsupervised, focusing on learning a model of normal data and identifying deviations (Salehi et al., 2021; Bergmann et al., 2021). These include reconstruction-based methods using autoencoders or GANs, where high reconstruction error signals an anomaly (An & Cho, 2015; Schlegl et al., 2017), and embedding-based methods, which map normal samples to a tight cluster in a feature space (Roth et al., 2022; Defard et al., 2021). While effective for detection and localization on benchmarks like MVTec AD (Bergmann et al., 2019), these approaches typically do not provide defect analysis for their predictions.

The advent of LMMs has introduced a new paradigm that enables comprehensive defect analysis. Rather than simply detecting anomalies, LMMs can now respond to diverse analytical queries about defect characteristics—such as identifying defect types, describing their visual appearance, and analyzing their potential consequences. To facilitate research in this area, a MMAD benchmark (Jiang et al., 2025) has been presented, providing a suite for evaluating an LMM's reasoning abilities on industrial anomaly problems. Meanwhile, several works have recently applied LMMs to this challenge. AnomalyGPT (Gu et al., 2024) was an early method using LMMs for zero-shot anomaly detection and generating descriptive reports. More recently, a few methods employ reinforcement learning to improve reasoning quality. AnomalyR1 (Chao et al., 2025) incorporates Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to refine its reasoning. Furthermore, OmniAD (Zhao et al., 2025) unifies the anomaly segmentation and anomaly reasoning problem for fine-grained defect understanding.

3 Method

Our proposed model, **JUDO**, is a domain-oriented multimodal reasoner for industrial anomaly understanding. The main claim of JUDO is that domain knowledge and contextual information fundamentally matter in industrial anomaly understanding, yet this approach remains underexplored and non-trivial. In this direction, our core contribution is a unified learning objective in which comparative visual reasoning and domain semantics fuse together into the domain-aligned reasoning process, described in Figure 1. The first stage introduces a comparative reasoning ability between query and normal images for patch-level segmentation, then the second stage internalizes textual domain knowledge into model parameters, and the final stage unifies the visual grounding and domain semantics using GRPO-based optimization tailored reward designs. The following subsections detail each training stage, with the corresponding dataset construction process described in Appendix A.

3.1 Stage 1: Learning Anomaly Segmentation-Based Juxtaposed Reasoning

For reliable and fine-grained visual reasoning, we employ anomaly segmentation to pinpoint defect regions and integrate this capability into the reasoning process. To encompass proper domain visual context, we propose the juxtaposed reasoning paradigm that explicitly compares defective images against their normal samples during training. While most existing segmentation methods achieve reasonable performance by directly predicting the target regions, the lack of clear reasoning criteria for anomaly judgment limits segmentation performance in defect detection. Additionally, although using normal images as contextual reference has become common practice during inference (Jiang et al., 2025), how to efficiently harness this comparative context to enable explicit reasoning of defects during training has yet to be addressed. Our comparative training paradigm shifts the objective from a simple pattern memorization to a fine-grained juxtaposed reasoning, enabling the model

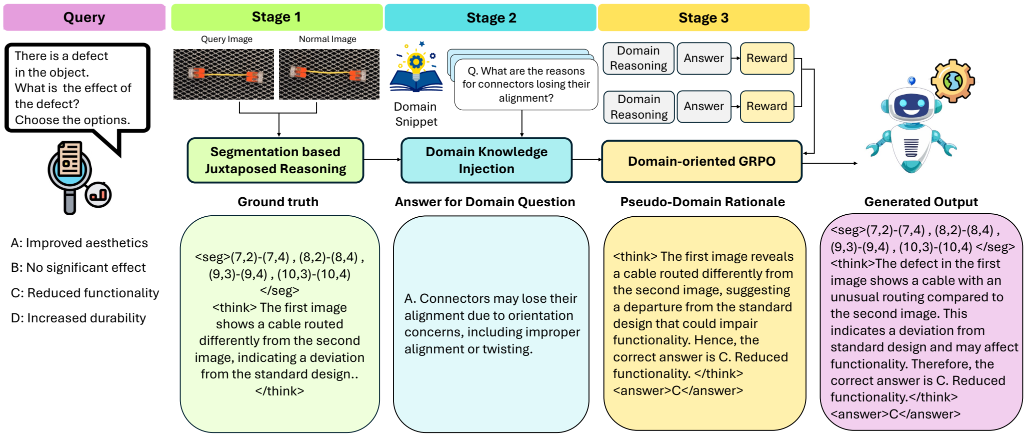

Query
Stage 1
Stage 2
Stage 3
There is a defect in the object. What is the effect of the defect? Choose the options.
Query Image
Normal Image
Domain Snippet
Q. What are the reasons for connectors losing their alignment?
Domain Reasoning
Answer
Reward
Domain Reasoning
Answer
Reward
Segmentation based Juxtaposed Reasoning
Domain Knowledge Injection
Domain-oriented GRPO
A: Improved aesthetics
B: No significant effect
C: Reduced functionality
D: Increased durability
Ground truth
<seg>(7,2)-(7,4) , (8,2)-(8,4) , (9,3)-(9,4) , (10,3)-(10,4) </seg> <think> The first image shows a cable routed differently from the second image, indicating a deviation from the standard design.. </think>
Answer for Domain Question
A. Connectors may lose their alignment due to orientation concerns, including improper alignment or twisting.
Pseudo-Domain Rationale
<think> The first image reveals a cable routed differently from the second image, suggesting a departure from the standard design that could impair functionality. Hence, the correct answer is C. Reduced functionality. </think> <answer>C</answer>
Generated Output
<seg>(7,2)-(7,4) , (8,2)-(8,4) , (9,3)-(9,4) , (10,3)-(10,4) </seg> <think>The defect in the first image shows a cable with an unusual routing compared to the second image. This indicates a deviation from standard design and may affect functionality. Therefore, the correct answer is C. Reduced functionality.</think> <answer>C</answer>

Figure 1: Overview of JUDO (Juxtaposed Domain-Oriented Multi-modal Reasoner). The framework progresses in three stages: (1) Stage 1: Learning Anomaly Segmentation-Based Juxtaposed Reasoning, (2) Stage 2: Domain-knowledge Injection, and (3) Stage 3: Domain-Oriented Group Relative Policy Optimization. Through the progressive stages, we incorporate the domain knowledge and context into LMMs for reliable and robust anomaly analysis.

to internalize a more generalized and fundamental notion of *normality* as a baseline for anomaly understanding.

Inspired by text-based segmentation approaches, such as Text4Seg (Lan et al., 2025; Zhao et al., 2025), the model is trained to output the coordinates of anomalous patches within a 16x16 grid through SFT (Schlegl et al., 2017). The instruction is given to describe the anomaly region and explain the visual evidence by comparing query images with normal samples as juxtaposed reasoning. For instance, a defect region is represented as a textual sequence such as (11,12)-(11,14), (12,11), pinpointing specific patches that deviate from the norm, including juxtaposed explanations. This format compels the model to perform a direct, patch-level juxtaposition, moving beyond a general comparison to a fine-grained recognition of differences against the normal template. This spatial grounding is crucial, as it forces the reasoning process to be tied to specific visual evidence, making the subsequent textual explanation more accurate and reliable.

Dataset Construction. We construct a training dataset for juxtaposed fine-grained anomaly segmentation. We utilize only one image from each defect category in the MMAD dataset, and additionally, incorporate the REAL-IAD (Wang et al., 2024) dataset for more generalized segmentation capability. As illustrated in Figure 5 of the Appendix, each anomalous query image is paired with a randomly sampled normal template from the same object category. The model is then trained to generate a dual-part response from this pairing: first, a textual sequence of patch coordinates between `<seg></seg>` tags that serves to ground the explanation by identifying the anomalous region, and second, a corresponding comparative explanation within `<think></think>` tags, which is synthetically generated to describe the defect by contrasting the two images, given the anomalous region bounded by a red line and the defect type.

3.2 Stage 2: Domain-knowledge Injection

While Stage 1 equips the model with comparative reasoning ability using normal images as the domain context, it primarily focuses on visual reasoning processes. However, the model still lacks sufficient domain knowledge for industrial anomaly problems. As a result, the model struggles to deliver accurate text reasoning and often fails to derive correct solutions without proper foundational knowledge. To address this limitation, we inject the domain knowledge (Mecklenburg et al., 2024) by constructing domain QA datasets and employing supervised fine-tuning (SFT). This approach differs from conventional supervised learning that directly trains on correct answers to queries in the MMAD dataset, as it focuses on supervised learning of domain knowledge that may not directly correspond to specific instructions but provides foundational knowledge essential for solving related problems. Through supervised fine-tuning on the domain QA dataset, the model acquires a more generalizable understanding of how domain knowledge applies across different object categories and defect types. This enhanced understanding forms the basis for Stage 3, where reinforcement alignment further refines the precision and reliability of domain reasoning.

Dataset Construction. We generate question–answer pairs using the raw domain snippets provided in MMAD (Jiang et al., 2025). Each snippet contains an unstructured textual description of the characteristics of an object category and its associated defect types, and serves as textual knowledge source for constructing the Stage 2 dataset (see Table 4 in the Appendix). Specifically, we prompt GPT 4o to reflect inspection knowledge from the unstructured domain snippet into structured QA pairs —for example, questions related to defect criteria, functional implications (e.g., "What criteria indicate that the fabric border is defective?" or "Why is detecting loose threads important for product reliability?"). These questions are derived purely from the textual information in the snippet and are not tied to any specific anomalous image sample. To improve robustness, each QA is further paraphrased into multiple semantically consistent but lexically diverse variants. We show a more detailed data construction process in Section A. Finally, every QA instance is paired with a normal image from the corresponding object category, grounding the textual domain knowledge in its object-level visual context and naturally prompting the model to recall relevant domain knowledge during reasoning. We provide examples of the generated QA pairs in Section B.2.

3.3 Stage 3: Domain-Oriented Group Relative Policy Optimization

The final stage of our framework, Stage 3, is designed to elevate the model's capabilities in the visual and textual domain-oriented reasoning, mainly obtained from prior stages. While Stage 1 builds a foundation in visual juxtaposed reasoning and Stage 2 internalizes textual domain knowledge, these skills are not yet integrated. Thus, we employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which is a reinforcement learning method that optimizes policies by comparing relative performance within groups of trajectories for better generalization. We design GRPO policies to align the model's behavior, ensuring that the final output is a seamless integration of accurate visual grounding and deep domain understanding. This alignment is guided by multi-faceted reward functions that provide a comprehensive feedback signal, composed of three primary components. We explain each policy with respect to reward functions as follows.

Domain Reasoning Reward. The primary objective of the Domain Reasoning Reward is to guide the reasoning process presented within the <think></think> tags toward aligning with the domain-oriented reasoning process, specifically queries targeted of the MMAD benchmark. Since there is no ground-truth for each instruction encompassing domain knowledge and context for MMAD datasets, we generate the *pseudo-domain rationale* using GPT-4o, providing full contexts such as the query, correct answer, images including normal reference, juxtaposed reasoning from Stage 1 pipeline, and relevant domain knowledge. We treat this generated pseudo-domain rationale as the semantic target and design the policy to guide domain-oriented direction. Specifically, we define the reward, R_{domain}, using the cosine similarity between the embedding vectors of the model's generated reasoning as E_{gen} and a pseudo-domain rationale denoted as $E_{pdomain}$ using the representation space of **all-MiniLM-L6-v2** SentenceTransformer (Reimers & Gurevych, 2019) model, which is effectively encode the semantic meaning of one or multiple sentences, denoted as $\phi(\cdot)$. This reward mechanism is formalized as:

$$R_{domain} = \frac{\phi(E_{gen}) \cdot \phi(E_{pdomain})}{\|\phi(E_{gen})\| \|\phi(E_{pdomain})\|}, \quad \text{if } R_{domain} \geq 0.5, \text{ otherwise } 0. \tag{1}$$

The domain reasoning reward leverages semantic similarity to align the model's reasoning with a reference rationale that is grounded strictly in the provided evidences. The pseudo-domain rationale reorganizes the inputs without introducing new knowledge, ensuring that GPT-4o serves only as an evidence structurer. This reward differs from typical GRPO settings that emphasize answer correctness or output format, since it directly encourages consistent domain-oriented reasoning patterns while remaining flexible to the phrasing of the generated explanation.

Segmentation Reward. This reward evaluates the spatial precision of the anomaly patch coordinates generated within the <seg></seg> tags through F1 score calculation (Zhao et al., 2025), reinforcing the skill developed in Stage 1. A reward of 0 is immediately assigned for any improperly formatted coordinate strings. For validly formatted outputs, the reward function compares the set of predicted grid cells (P) against the ground-truth set (P_G) using the following piecewise function:

$$R_{seg} = \begin{cases} 1.0 & \text{if } P = \emptyset \text{ and } P_G = \emptyset, \\ 0.2 + 0.8 \cdot \mathrm{F1}(P, P_G) & \text{if } P \neq \emptyset \text{ and } P_G \neq \emptyset, \\ 0.0 & \text{otherwise.} \end{cases} \quad (2)$$

This formulation incentivizes not only high-overlap localization but also correct format adherence and accurate identification of anomaly-free instances, penalizing cases where only one set is empty.

Choice and Structural Alignment Reward. This composite reward ensures correct, well-structured, and logically sound output through three components (Shao et al., 2024): 1) **Choice Reward** rewards correct multiple-choice selection within `<answer></answer>` tags to promote accurate decision-making; 2) **Format Reward** ensures adherence to the required format as `<seg>...<think>...<answer>` structure, ensuring the structure of responses is consistent and parsable; and 3) **Reasoning Structure Reward** encourages both the conclusive answer with in the reasoning and final answer to be correct while penalizing answer choices mentioned in the first half of reasoning text, preventing premature commitment.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. We evaluate our approach on the MMAD benchmarks (Jiang et al., 2025), which integrates four datasets, such as MVTec-AD (Bergmann et al., 2019), MVTec-LOCO (Bergmann et al., 2022), VisA (Zou et al., 2022), and GoodsAD (Zhang et al., 2024) datasets, with multiple choice QA covering seven key subtasks. We report averaged accuracy as the evaluation metric, computed as the ratio of correct predictions to the total number of predictions. For Stage 1, we sample 10 instances from each category in the Real-IAD dataset (Wang et al., 2024) together with one instance per category from MMAD, yielding a total of 1.4k and 293 images, respectively. In Stage 2, we construct a domain-specific QA corpus by leveraging the MMAD domain knowledge JSON files. For each category, we generate 30 unique questions and augment them with two paraphrased variants, resulting in approximately 13k QA pairs. Finally, Stage 3 applies GRPO using a sparse sampling strategy, where only one training instance per category from MMAD—identical to the sampled used in Stage 1—is utilized, yielding 1.4k QA pairs for reinforcement alignment.

Implementation Details. Our framework is built on PyTorch 2.5.1 with HuggingFace Transformers and the TRL GRPO trainer. We use Qwen2.5-VL-7B as the base model, initialized from a vision–language pre-trained checkpoint. Training runs on a single node with 4×NVIDIA H200 GPUs using `torchrun` and DeepSpeed ZeRO-3. For Stage 1 and Stage 2, we train for 8 and 2 epochs, respectively, using learning rates of 1×10^{-6} and 5×10^{-7} in `bf16` precision. For GRPO, we use 16 generations per prompt, batch size of 8, and train for 14 epochs, also in `bf16`.

Baselines and Inference Settings. We benchmark JUDO against leading general-purpose models such as Qwen2.5-VL, InternVL3.5, and various commercial models. Our primary industrial specialized baseline is AnomalyR1 (Chao et al., 2025); another recent method, OmniAD (Zhao et al., 2025), was not included due to the unavailability of its public codebase. For consistency with JUDO's architecture, we re-implemented the AnomalyR1 approach on a Qwen2.5-VL 7B model using the author's provided code and our sampled dataset. The result is referred as the AnomalyR1 in Table 1 and Base GRPO model in our ablation study in Section 4.3. All models are evaluated under a strict 1-shot inference protocol in the equivalent way of MMAD, using both a query image and a normal template as input to ensure a fair comparison.

4.2 Experimental results

Table 1 presents the comparison against both general-purpose LMMs and the anomaly-focused baseline AnomalyR1. JUDO achieves strong performance across the MMAD seven subtasks, reaching the highest overall averaged accuracy of 80.73% among the open-source models. Advantages of JUDO are particularly evident across the four defect-related subtasks: classification, localization, description, and analysis. Since these tasks assume that the domain-specific knowledge plays a crucial role, the performance gains are a direct result of the effective incorporation of domain knowledge

Table 1: Performance comparison of both commercial and open-source LMMs in MMAD with the standard 1-shot setting. Anomaly Discrimination uses the average accuracy across the normal and abnormal categories. The best scores are highlighted in bold while the second best are underlined. * indicates the best performance among the open-source models.

Model	Scale	Anomaly	Defect				Object		Average
		Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Random Chance	-	50.00	25.00	25.00	25.00	25.00	25.00	25.00	28.57
Human (expert)	-	95.24	75.00	92.31	83.33	94.20	86.11	80.37	86.65
Human (ordinary)	-	86.90	66.25	85.58	71.25	81.52	89.58	69.72	78.69
Claude-3.5-sonnet	-	60.14	60.14	48.81	67.13	79.11	85.19	79.83	68.36
Gemini-1.5-pro	-	68.63	60.12	58.56	70.38	82.46	89.20	82.25	73.09
Gemini-2.5-pro	-	83.07	**73.86**	67.20	79.97	86.27	94.88	83.08	**81.19**
Gemini-2.5-flash	-	**93.38**	69.88	63.30	76.41	81.57	94.04	82.00	80.08
GPT-4o	-	68.63	65.80	55.62	73.21	83.41	**94.98**	82.80	74.92
GPT-5-mini	-	64.10	67.35	69.07	79.02	86.72	93.96	83.37	77.65
Qwen2.5-VL	7B	71.39	54.35	61.17	65.81	79.32	91.44	84.43	72.56
LLaVA-OneVision	7B	51.77	46.13	41.85	62.19	69.73	90.31	80.93	63.27
InternVL3.5	8B	67.50	49.37	57.9	58.07	77.66	72.01	81.11	66.30
Kimi-VL-A3B	16B	72.93*	53.49	59.66	72.39	81.74	91.91	85.89	74.00
MiMo-VL	8B	54.45	59.56	60.77	71.50	78.48	90.56	81.60	70.99
AnomalyR1	7B	60.93	64.81	70.72	79.06	85.52	93.12	**86.91***	77.29
JUDO	7B	64.51	72.17*	**75.95***	**84.38***	**87.76***	94.24*	86.07	80.73*

and reasoning through our progressive stages. Specifically, Stage 1 establishes a strong foundation by using juxtaposed segmentation training to foster fine-grained comparative analysis, which directly boosts defect localization to 75.59% (from 61.17% for the base Qwen). This visual grounding is then enriched in Stage 2, where the integration of domain-specific knowledge provides the conceptual basis needed for accurate explanations, as seen in the 84.38% and 87.76% accuracy on the defect description and analysis, respectively. Finally, Stage 3 is critical for merging these capabilities. The multi-reward GRPO training ensures the model's final output is a coherent synthesis of juxtaposed visual evidence and textual expertise. This integrated process allows JUDO to not only locate defects precisely but also to classify, describe, and analyze them with a domain-aligned thinking process, leading to more robust and reliable reasoning.

However, our JUDO does not achieve superior performance in anomaly discrimination tasks, even with rich domain-specific foundational knowledge. Similarly, other models trained with a GRPO framework, such as AnomalyR1, exhibit a trade-off in raw binary detection accuracy. This is evident in the results: AnomalyR1 achieves 60.93% accuracy, while JUDO shows only a modest improvement at 64.51%. Both remain below the 72.93% and 71.39% reached by recent LLMs such as Kimi-VL, as well as the base model Qwen2.5-VL, respectively. We attribute Kimi-VL's higher performance to its more advanced visual encoder, although it still lacks domain-specialized knowledge, as reflected in its relatively low performance on defect-related tasks. In Section 4.4, we further discuss the impact of our multi-staged learning framework under anomaly discrimination.

Commercial multimodal models such as Gemini-2.5-Pro and Gemini-2.5-Flash show strong performance on anomaly discrimination, likely due to their highly capable vision encoders. However, their performance on domain-level defect reasoning remains overall lower than JUDO's. This contrast highlights a key distinction: while large commercial LMMs excel at broad visual pattern recognition, they do not have sufficient defect semantics or comparative industrial cues in a way that generalizes across classification, localization, description, and analysis. In contrast, despite being built on a smaller open-source backbone, JUDO consistently achieves higher defect-reasoning accuracy because its training explicitly aligns comparative visual grounding with structured domain knowledge. These results demonstrate that domain-aligned training can surpass much larger commercial systems when the task requires specialized industrial reasoning rather than generic visual understanding.

4.3 Ablation Study

We conduct an ablation study to systematically evaluate the contribution of each component in the JUDO framework, with results summarized in Table 2. The study begins with the baseline Qwen2.5-VL-7B model, which achieves an average accuracy of 72.56%. We note that this result is obtained without using the Chain-of-Thought process, not requiring the reasoning process gener-

Table 2: Ablation study of different methods. We denote Stage 1 as SegJux, Stage 2 as DomInj, and the full Stage 3 as GRPO^{dom}, where the domain reasoning reward is used. The segmentation reward is only used in methods that go through Stage 1 training. Average accuracy (%) is reported.

Method	Average
Qwen2.5-VL-7B	72.56
+ GRPO	77.29
+ GRPO + RAG	76.29
+ GRPO + DomInj	79.82
+ GRPO + SegJux + DomInj	80.35
+ GRPO^{dom} + SegJux + DomInj (JUDO)	**80.73**

ally in <think> tags, but outputting the answer directly, since it is not trained to learn the proper reasoning, yet. Applying a standard GRPO training stage improves general instruction-following and raises the accuracy to 77.30% in the equivalent of AnomalyR1. To evaluate the effectiveness of domain knowledge injection, we employ RAG, which dynamically provides external information upon the model trained on vanilla GRPO, referring to the model as (+ GRPO + RAG). The way of employing RAG is followed in the MMAD (Jiang et al., 2025). While RAG-based approaches have been claimed to often improve the performance, the external context attached to our trained reasoning model acted in the negative direction rather than bringing in positive gain. In contrast, the impact of domain injection was remarkably effective, resulting in an improvement of almost 5% gain.

The effectiveness of JUDO's learning-based approach becomes evident in the subsequent steps as shown in the result. The most significant performance leap comes from Stage 2's Domain Injection (+ GRPO + DomInj), which internalizes domain knowledge through supervised fine-tuning and increases accuracy to 79.82%. This result is substantially higher than the RAG-based method, demonstrating the superiority of integrating textual knowledge directly into the model's parameters. Building on this, the addition of Stage 1's juxtaposed segmentation training (+ GRPO + SegJux + DomInj) brings the accuracy to 80.35%, highlighting the value of grounding textual reasoning in fine-grained juxtaposed visual analysis. The final refinement, which uses the complete Stage 3 GRPO with a specialized domain reasoning reward (+ GRPO$^{\text{dom}}$), achieves the peak performance of 80.75%. This confirms that each stage of the JUDO framework provides a meaningful and incremental benefit, leading to its final state-of-the-art performance.

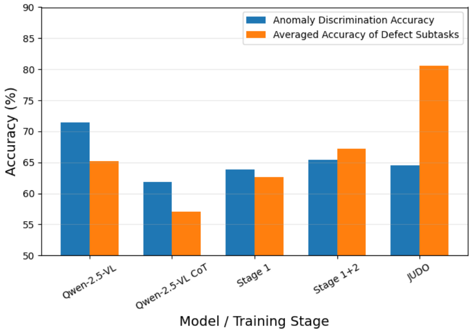
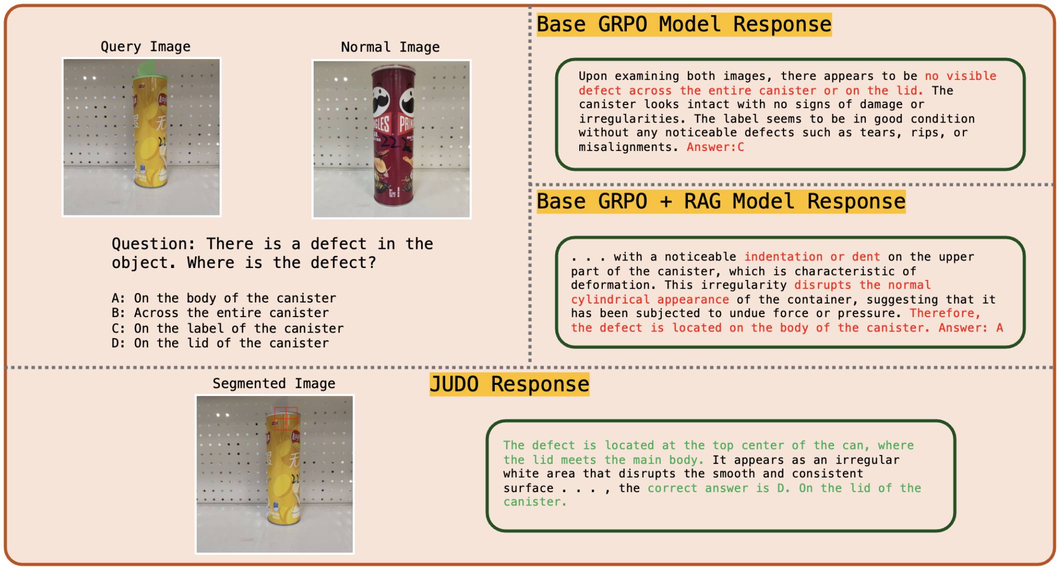


Figure 2: Performance across different stages.

4.4 Discussion of Multi-staged Learning Framework

Figure 2 presents the progression of anomaly discrimination accuracy and averaged accuracy of defect subtasks (classification, localization, description, and analysis) across the key checkpoints of our pipeline: the Qwen2.5-VL baseline, its Chain-of-Thought (CoT) variant, and each JUDO stage. The figure shows that the most significant performance shift occurs before any of JUDO's optimization takes place. When the model switches from direct answering to CoT reasoning, anomaly discrimination accuracy drops sharply (71.39% → 61.90%), and the averaged accuracy over the four defect subtasks decreases from 65.15% to 57.00%. This confirms that the initial degradation stems from a reasoning-mode behavioral change rather than from the design of our multi-stage training. The drop in anomaly discrimination in particular is consistent with a recent finding (Liu et al., 2024) that explicit verbal reasoning can impair performance on perception-heavy tasks by interfering with rapid, pattern-based visual judgments, a phenomenon analogous to verbal overshadowing. After entering



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

A: On the body of the canister
B: Across the entire canister
C: On the label of the canister
D: On the lid of the canister

Upon examining both images, there appears to be no visible defect across the entire canister or on the lid. The canister looks intact with no signs of damage or irregularities. The label seems to be in good condition without any noticeable defects such as tears, rips, or misalignments. Answer:C

. . . with a noticeable indentation or dent on the upper part of the canister, which is characteristic of deformation. This irregularity disrupts the normal cylindrical appearance of the container, suggesting that it has been subjected to undue force or pressure. Therefore, the defect is located on the body of the canister. Answer: A

The defect is located at the top center of the can, where the lid meets the main body. It appears as an irregular white area that disrupts the smooth and consistent surface . . . , the correct answer is D. On the lid of the canister.

Figure 3: Response comparison between Base GRPO, Base GRPO + RAG and JUDO. The anomalous region in the query image is highlighted in green, while the segmentation output from JUDO is represented as red patches in the Segmented Image.

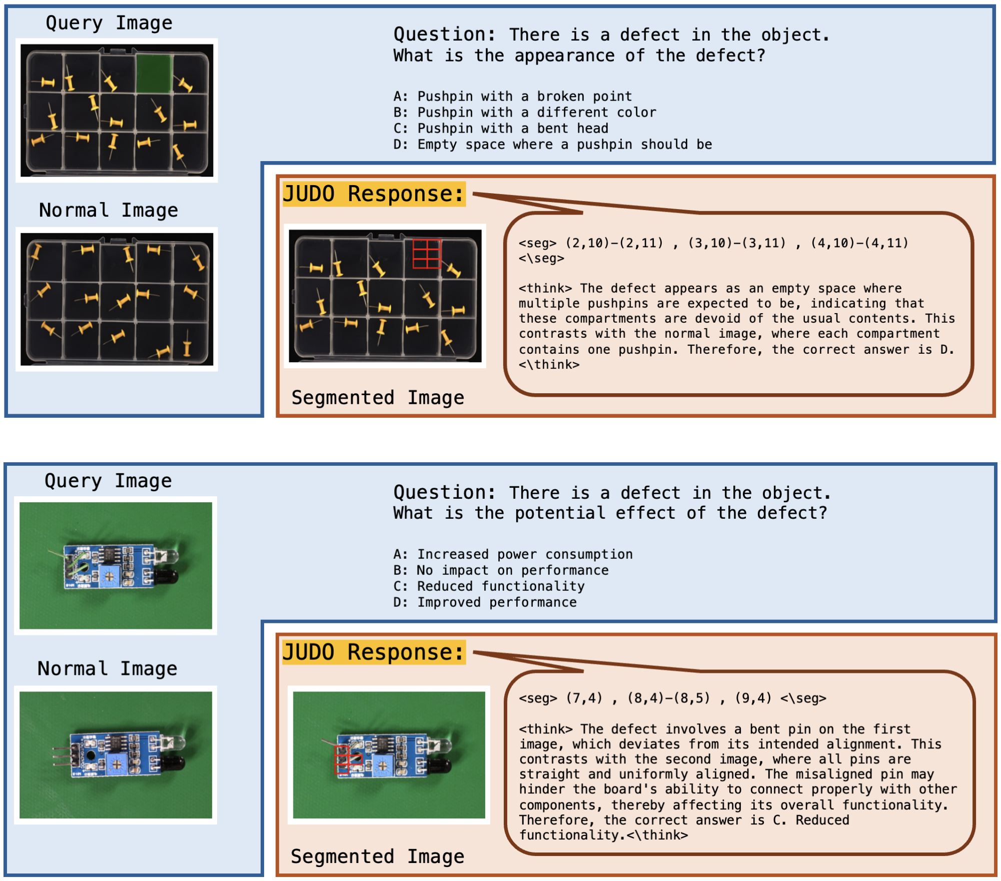
the pipeline, anomaly-detection performance remains largely stable, with only a small fluctuation from 65.42% (Stage 1+2) to 64.51% in the final model, indicating that catastrophic forgetting is present but limited.

In contrast, the progressive optimization produces substantial and consistent gains across the four defect subtasks. As shown in Figure 2, accuracy increases from 62.58% after Stage 1 to 67.18% after Stage 1+2, and ultimately reaches 80.57% in the full model. This pattern demonstrates that the GRPO-based domain alignment strengthens domain-aware classification, localization, description, and analysis, even as low-level anomaly discrimination remains stable. These results highlight that JUDO's multi-stage design primarily enhances domain-oriented reasoning capabilities while preserving binary anomaly detection performance.

4.5 Qualitative Analysis

Figure 3 provides a qualitative comparison that highlights the limitations of baseline models and the effectiveness of JUDO's integrated reasoning. The Base GRPO model exhibits a critical failure, as it is unable to detect the defect and states there is "no visible defect," while paradoxically offering an incorrect answer ("C"). The Base GRPO + RAG model demonstrates a different failure mode rooted in contextual distraction. Misled by the retrieved domain snippet, which contains descriptions of multiple potential defects, it latches onto a plausible but incorrect description of a "noticeable indentation or dent"—likely related to deformation—instead of the correct "opened" defect type. This failure to ground the textual context in the visual evidence leads to a mislocalization of the "body of the canister" and the wrong answer ("A"). This negative impact is also confirmed by the quantitative results, where the RAG model's accuracy (76.29%) is lower than the Base GRPO model's (77.29%).

In stark contrast, JUDO leverages its internalized knowledge for a successful and coherent analysis. It correctly identifies and locates the defect on the "lid of the canister," provides a visually-grounded explanation of an "irregular white area," and supports its reasoning with an accurate segmentation map to arrive at the correct answer ("D"). This direct comparison clearly illustrates that JUDO's integrated, learning-based approach is essential for achieving reliable and fine-grained reasoning, overcoming the weaknesses of models that either lack domain knowledge or are distracted by it.



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

A: Pushpin with a broken point
B: Pushpin with a different color
C: Pushpin with a bent head
D: Empty space where a pushpin should be

```
<seg> (2,10)-(2,11) , (3,10)-(3,11) , (4,10)-(4,11)
<\seg>

<think> The defect appears as an empty space where
multiple pushpins are expected to be, indicating that
these compartments are devoid of the usual contents. This
contrasts with the normal image, where each compartment
contains one pushpin. Therefore, the correct answer is D.
<\think>
```

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

A: Increased power consumption
B: No impact on performance
C: Reduced functionality
D: Improved performance

```
<seg> (7,4) , (8,4)-(8,5) , (9,4) <\seg>

<think> The defect involves a bent pin on the first
image, which deviates from its intended alignment. This
contrasts with the second image, where all pins are
straight and uniformly aligned. The misaligned pin may
hinder the board's ability to connect properly with other
components, thereby affecting its overall functionality.
Therefore, the correct answer is C. Reduced
functionality.<\think>
```

Figure 4: Examples of JUDO's output on MMAD dataset. The anomalous region in the query image is highlighted in green, while the segmentation output from JUDO is represented as red patches in the Segmented Image.

5 Conclusion

In this work, we presented JUDO, a novel framework that addresses the core challenge of aligning domain knowledge in industrial anomaly understanding. By unifying comparative visual reasoning and domain semantics into a domain-aligned learning objective, our approach significantly enhances anomaly understanding capabilities. Experimental results on the MMAD benchmark demonstrate superior performance compared to state-of-the-art models in defect-reasoning tasks. These findings suggest that internalizing domain knowledge and context during training is not only highly effective but also significantly more beneficial than providing context solely at inference time. This approach, therefore, enables more reliable and accurate multimodal reasoning systems in complex industrial scenarios.

Ethics Statement

The authors have read and adhered to the ICLR Code of Ethics. This research is based on publicly available datasets for industrial anomaly detection (MMAD, REAL-IAD), which contain images of inanimate industrial objects. No human subjects were involved in this study, and no personally identifiable information was used, thereby minimizing privacy concerns. Our work uses GPT-4o for the programmatic generation of structured training data, as detailed in Appendix A.1, to ensure a consistent and replicable data construction pipeline and grammar correction. The goal of this research is to advance industrial quality control and automation, and we do not foresee any direct negative societal impacts or ethical concerns arising from the proposed method.

Reproducibility Statement

We are committed to ensuring the reproducibility of our work. The complete source code for our framework, including data preprocessing scripts, implementation of all three training stages, and evaluation protocols, is provided in the supplementary materials, accessible via an anonymous repository at `https://anonymous.4open.science/r/JUDO-9C8B`. The core methodology of our three-stage training process is detailed in Section 3. Our experimental setup, including the specific datasets, baselines, and key implementation details such as the base model, libraries, and hyperparameters, is described in Section 4.1. Furthermore, a comprehensive description of our programmatic dataset construction pipeline for all training stages can be found in Appendix A.1. We believe these resources provide the necessary details for the research community to reproduce our results.

References

Garima Agrawal, Kuntal Pal, Yuli Deng, Huan Liu, and Ying-Chih Chen. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23164–23172, 2024a.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024b.

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2:1–18, 2015.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Pengfei Wang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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*, pp. 9592–9600, 2019.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

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.

Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection. *arXiv e-prints*, pp. arXiv–2504, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Thomas Defard, Aleksandr Setkov, Angelika Loesch, and Romann Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

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, pp. 1932–1940, 2024.

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.

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.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

Zhiqiang Liu, Chengtao Gan, Junjie Wang, Yichi Zhang, Zhongpu Bo, Mengshu Sun, Huajun Chen, and Wen Zhang. Ontotune: Ontology-driven self-training for aligning large language models. In *Proceedings of the ACM on Web Conference 2025*, pp. 119–133, 2025.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, et al. Injecting new knowledge into large language models via supervised fine-tuning. *arXiv preprint arXiv:2404.00213*, 2024.

Vignesh Prabhakar, Md Amirul Islam, Adam Atanas, Yao-Ting Wang, Joah Han, Aastha Jhunjhunwala, Rucha Apte, Robert Clark, Kang Xu, Zihan Wang, et al. Omniscience: A domain-specialized llm for scientific reasoning and discovery. *arXiv preprint arXiv:2503.17604*, 2025.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv e-prints*, pp. arXiv–2502, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Zeynep Akata, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19551–19561, 2022.

Mohammadhassan Salehi, Niousha Sadjadi, Mehran Vaini, Erfan Entezari, and S Mehran Razavi. A survey of deep learning for industrial visual anomaly detection. *arXiv preprint arXiv:2112.06246*, 2021.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pp. 146–157. Springer, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. Injecting domain-specific knowledge into large language models: a comprehensive survey. *arXiv preprint arXiv:2502.10708*, 2025.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

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*, pp. 22883–22892, 2024.

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

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.

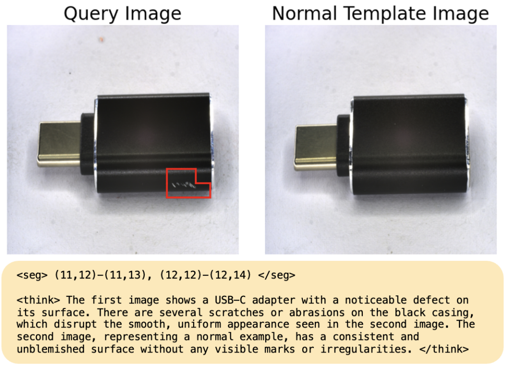
Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20370–20382, 2025.

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.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.

Shifang Zhao, Yiheng Lin, Lu Han, Yao Zhao, and Yunchao Wei. Omniad: Detect and understand industrial anomaly via multimodal reasoning. *arXiv preprint arXiv:2505.22039*, 2025.

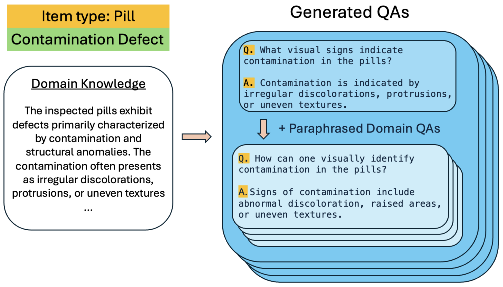
Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.



```
<seg> (11,12)-(11,13), (12,12)-(12,14) </seg>

<think> The first image shows a USB-C adapter with a noticeable defect on
its surface. There are several scratches or abrasions on the black casing,
which disrupt the smooth, uniform appearance seen in the second image. The
second image, representing a normal example, has a consistent and
unblemished surface without any visible marks or irregularities. </think>
```

Figure 5: Stage 1 Training Data Example



The inspected pills exhibit defects primarily characterized by contamination and structural anomalies. The contamination often presents as irregular discolorations, protrusions, or uneven textures
...

Q. What visual signs indicate contamination in the pills?

A. Contamination is indicated by irregular discolorations, protrusions, or uneven textures.

Q. How can one visually identify contamination in the pills?

A. Signs of contamination include abnormal discoloration, raised areas, or uneven textures.

Figure 6: Stage 2 Training Data Example

A Detailed Dataset Construction Process

To construct the datasets for our three-stage framework, we developed a programmatic pipeline that utilizes GPT-4o. This approach was chosen to ensure high quality, structural consistency, and scalability across all data. The pipeline consisted of three stages: constructing comparative explanations (Stage 1), constructing domain Q&As datasets (Stage 2), and generating pseudo-domain reference reasoning (Stage 3). The specific process for each dataset construction is detailed below.

Stage 1: Comparative Explanation. The primary goal of this stage is to create a dataset that teaches the model to perform fine-grained, juxtaposed reasoning by comparing anomalous and normal images. The construction process for each data instance is as follows:

1. **Input:** The process begins with an anomalous query image, its corresponding binary anomaly mask, and a randomly selected normal template image from the same object category.
2. **Anomaly Patch Generation:** The provided anomaly mask is used to identify the precise location of the defect. We overlay a 16x16 grid on the image, and any grid cell that overlaps with the mask is marked as anomalous. The coordinates of these anomalous patches are then converted into a textual sequence. This sequence uses `(row, column)` notation for individual patches and `(row, col_start)-(row, col_end)` for contiguous horizontal patches. This text string is then encapsulated within `<seg>` tags (e.g., `<seg>(11,12)-(11,14), (12,11)</seg>`).
3. **Comparative Explanation Synthesis:** To generate the reasoning text, we first annotate the anomalous query image by using its mask to draw a red outline around the defect region. This visually grounded image, along with the normal template image, the defect type, and any relevant domain notes, is provided as input to GPT-4o. The model is prompted to act as an expert and generate a concise comparative explanation, describing the visual differences between the two images with a focus on the defective region's characteristics. This synthetically generated text forms the content for the `<think>` tags.
4. **Final Data Assembly:** The final training instance combines the visual and textual components. It consists of the original anomalous image and the normal template image as inputs, paired with the generated dual-part response: the `<seg>` tag containing the patch coordinates, followed by the `<think>` tag containing the comparative explanation. Figure 5 shows an example of a training data for Stage 1.

Stage 2: Domain Q&A. This stage aims to inject essential, category-specific domain knowledge into the model through supervised fine-tuning. The dataset is constructed by converting the unstructured textual knowledge from MMAD into a structured question-answer format.

1. **Input:** We utilize the "domain snippets" from the MMAD benchmark as the foundational knowledge source. Each snippet provides a textual description of the characteristics of a specific object category and its associated defect types. Table 4 shows an example of a domain snippet for "Squeezed Teeth" defect of the object "Zipper".

Table 3: System prompt for data construction in Stage 2

Generate unique QA pairs grounded in the snippet with enforced categories.
You are tasked with generating high-quality Q&A pairs grounded strictly in the given snippet. Rules: - DO NOT mention ”snippet” or ”according to the snippet”.
- All answers must be strictly grounded in the text, no outside knowledge.
- Produce exactly count unique question–answer pairs.
- Use a variety of question styles, including:
1. Criteria-based (e.g., ”What criteria indicate a defect in ...?”)
2. Defect understanding (e.g., ”How does this affect. . . ?”)
3. Comparative reasoning (e.g., ”What distinguishes X from Y?”)
4. Functional impact (e.g., ”Why does this defect matter?”)
5. Recognition (e.g., ”What is. . . ?”)
6. Quality control reasoning (e.g., ”Why is it important to detect. . . ?”)
7. Aesthetic/structural concerns
- Ensure balance: include multiple styles, not just one.
- Keep answers factual and strictly tied to the snippet.
Return as a JSON array: [{{”question”: ”...”, ”answer”: ”...”}}]
Snippet:{snippet}

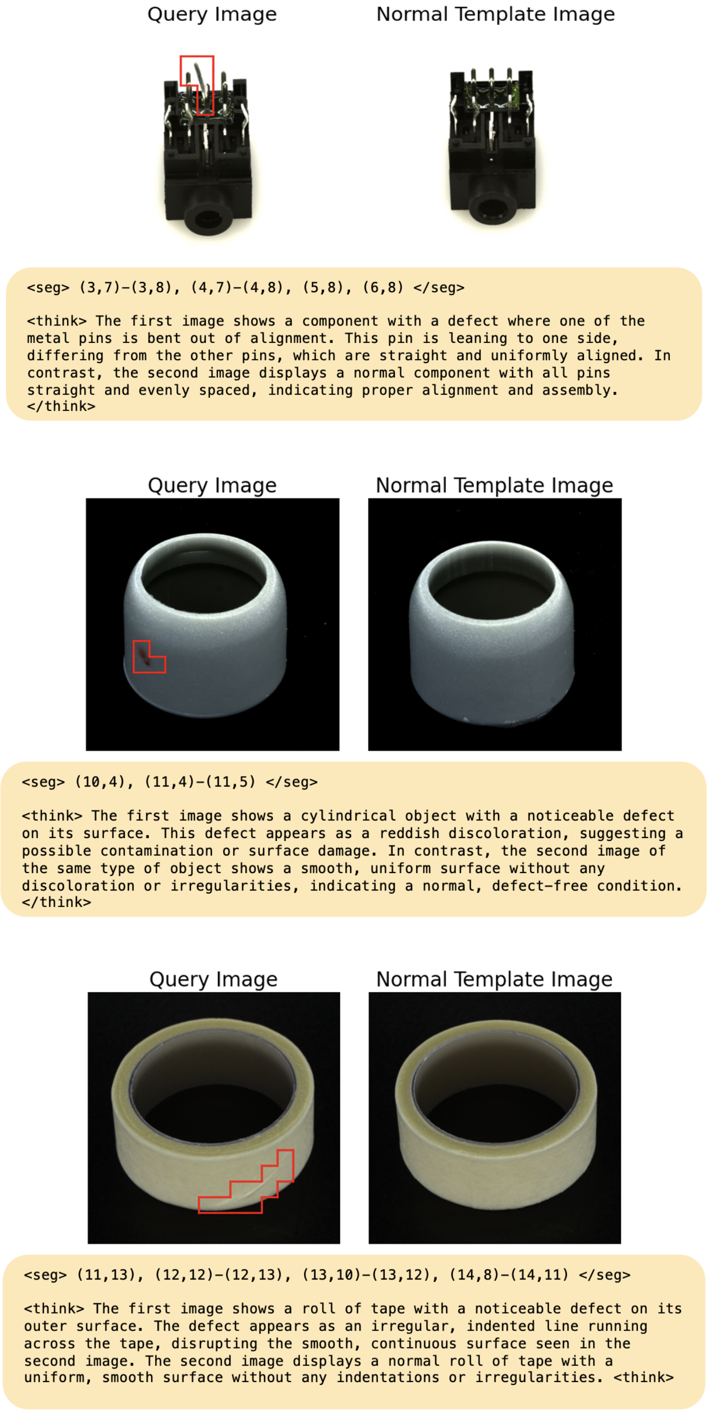
Table 4: Example of a domain snippet for “Squeezed Teeth” defect of the object “Zipper”.

The defect manifests as teeth of the zipper that appear squeezed or compressed together, disrupting the otherwise uniform alignment and spacing of the zipper teeth. This irregularity can occur at various locations along the zipper, often concentrated in the central or top/bottom areas. Characteristics of the defect include noticeable distortion in the shape of the teeth, which appear pinched and may vary in spacing compared to the adjacent, properly aligned teeth. The presence of squeezed teeth could potentially hinder the functionality of the zipper, leading to difficulties in smooth opening and closing, or causing the zipper slider to get stuck during use. The affected area is identifiable by its misshapen and irregular appearance, contrasting with the uniform and symmetrical pattern of the healthy zipper teeth surrounding it.

2. **Initial Q&A Generation:** For each domain snippet, we employ GPT-4o to generate an initial set of 30 unique question-answer pairs. The generation is guided by a prompt that instructs the model to formulate questions reflecting industrial inspection knowledge, such as inquiries about defect criteria, visual appearance, or functional impact. A crucial constraint is that all answers must be strictly grounded in the provided text snippet, with no external knowledge introduced.
3. **Paraphrasing for Robustness:** To enhance the model's robustness and prevent overfitting on specific phrasing, each of the initial 30 Q&A pairs is further processed. We use GPT-4o again to generate two semantically identical but lexically diverse paraphrases for each pair. This step expands the dataset with varied phrasings of the same core knowledge.
4. **Final Data Assembly:** The final dataset consists of the original and paraphrased Q&A pairs. To anchor this textual knowledge to a visual context, each Q&A instance is paired with a randomly sampled normal image from the corresponding object category. This encourages the model to associate the domain knowledge with the visual appearance of the object, facilitating better knowledge recall during multimodal reasoning tasks. Figure 8 shows examples of the final Q&A pairs in Stage 2 generated from the domain snippets.

B Dataset Examples

B.1 Stage 1



<seg> (3,7)–(3,8), (4,7)–(4,8), (5,8), (6,8) </seg>

<think> The first image shows a component with a defect where one of the metal pins is bent out of alignment. This pin is leaning to one side, differing from the other pins, which are straight and uniformly aligned. In contrast, the second image displays a normal component with all pins straight and evenly spaced, indicating proper alignment and assembly. </think>

<seg> (10,4), (11,4)–(11,5) </seg>

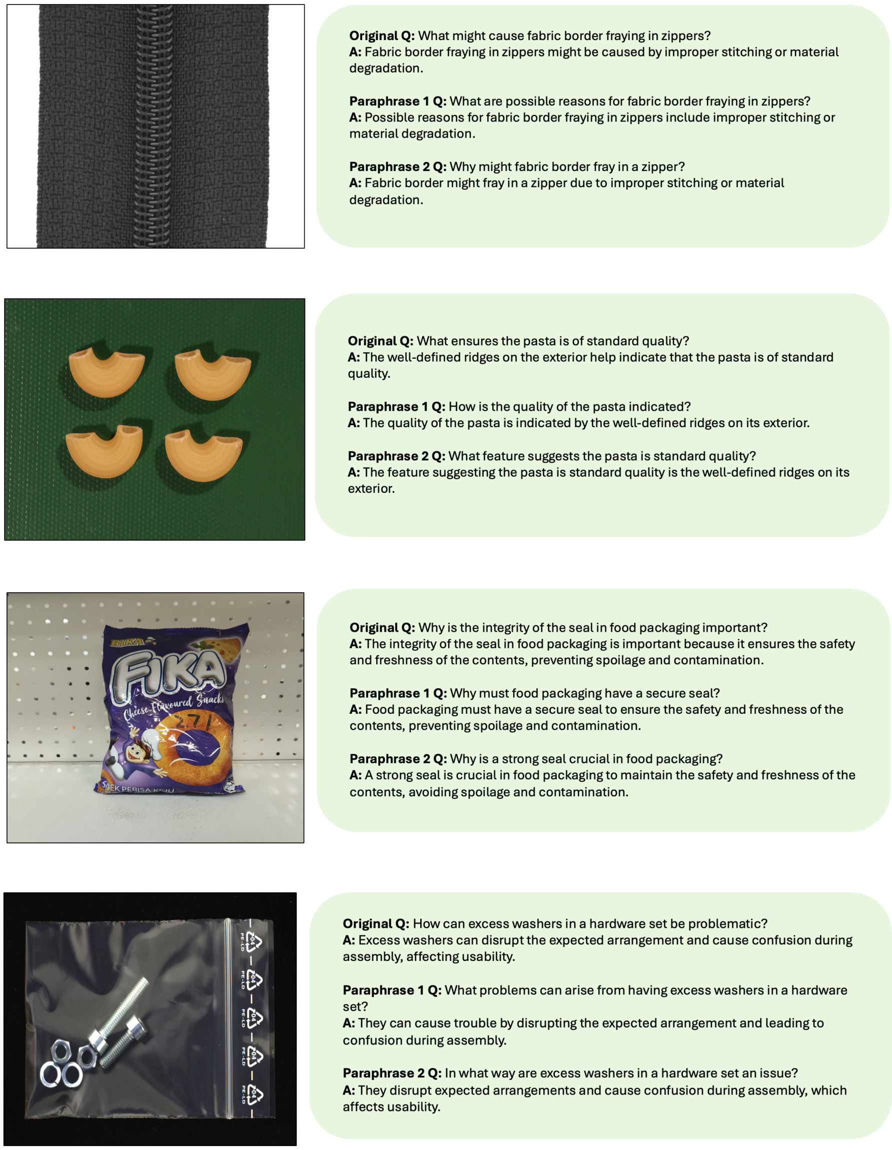
<think> The first image shows a cylindrical object with a noticeable defect on its surface. This defect appears as a reddish discoloration, suggesting a possible contamination or surface damage. In contrast, the second image of the same type of object shows a smooth, uniform surface without any discoloration or irregularities, indicating a normal, defect-free condition. </think>

<seg> (11,13), (12,12)–(12,13), (13,10)–(13,12), (14,8)–(14,11) </seg>

<think> The first image shows a roll of tape with a noticeable defect on its outer surface. The defect appears as an irregular, indented line running across the tape, disrupting the smooth, continuous surface seen in the second image. The second image displays a normal roll of tape with a uniform, smooth surface without any indentations or irregularities. <think>

Figure 7: Stage 1 Training Dataset Examples

B.2 STAGE 2



Original Q: What might cause fabric border fraying in zippers?
A: Fabric border fraying in zippers might be caused by improper stitching or material degradation.

Paraphrase 1 Q: What are possible reasons for fabric border fraying in zippers?
A: Possible reasons for fabric border fraying in zippers include improper stitching or material degradation.

Paraphrase 2 Q: Why might fabric border fray in a zipper?
A: Fabric border might fray in a zipper due to improper stitching or material degradation.

Original Q: What ensures the pasta is of standard quality?
A: The well-defined ridges on the exterior help indicate that the pasta is of standard quality.

Paraphrase 1 Q: How is the quality of the pasta indicated?
A: The quality of the pasta is indicated by the well-defined ridges on its exterior.

Paraphrase 2 Q: What feature suggests the pasta is standard quality?
A: The feature suggesting the pasta is standard quality is the well-defined ridges on its exterior.

Original Q: Why is the integrity of the seal in food packaging important?
A: The integrity of the seal in food packaging is important because it ensures the safety and freshness of the contents, preventing spoilage and contamination.

Paraphrase 1 Q: Why must food packaging have a secure seal?
A: Food packaging must have a secure seal to ensure the safety and freshness of the contents, preventing spoilage and contamination.

Paraphrase 2 Q: Why is a strong seal crucial in food packaging?
A: A strong seal is crucial in food packaging to maintain the safety and freshness of the contents, avoiding spoilage and contamination.

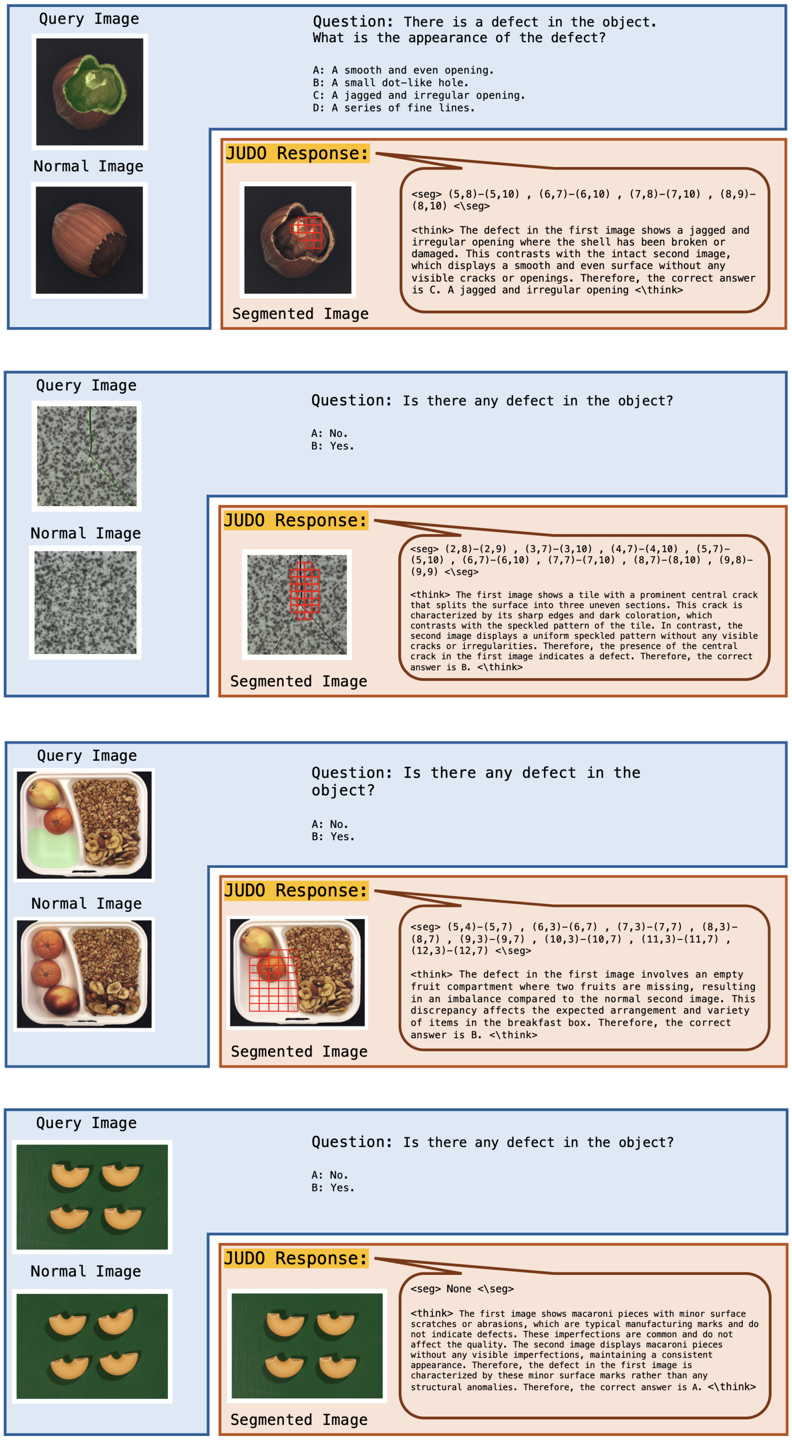
Original Q: How can excess washers in a hardware set be problematic?
A: Excess washers can disrupt the expected arrangement and cause confusion during assembly, affecting usability.

Paraphrase 1 Q: What problems can arise from having excess washers in a hardware set?
A: They can cause trouble by disrupting the expected arrangement and leading to confusion during assembly.

Paraphrase 2 Q: In what way are excess washers in a hardware set an issue?
A: They disrupt expected arrangements and cause confusion during assembly, which affects usability.

Figure 8: Stage 2 Training Dataset Examples

C ADDITIONAL JUDO RESPONSES



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

A: A smooth and even opening.
B: A small dot-like hole.
C: A jagged and irregular opening.
D: A series of fine lines.

JUDO Response:

<seg> (5,8)-(5,10) , (6,7)-(6,10) , (7,8)-(7,10) , (8,9)-(8,10) <\seg>

<think> The defect in the first image shows a jagged and irregular opening where the shell has been broken or damaged. This contrasts with the intact second image, which displays a smooth and even surface without any visible cracks or openings. Therefore, the correct answer is C. A jagged and irregular opening <\think>

Question: Is there any defect in the object?

A: No.
B: Yes.

JUDO Response:

<seg> (2,8)-(2,9) , (3,7)-(3,10) , (4,7)-(4,10) , (5,7)-(5,10) , (6,7)-(6,10) , (7,7)-(7,10) , (8,7)-(8,10) , (9,8)-(9,9) <\seg>

<think> The first image shows a tile with a prominent central crack that splits the surface into three uneven sections. This crack is characterized by its sharp edges and dark coloration, which contrasts with the speckled pattern of the tile. In contrast, the second image displays a uniform speckled pattern without any visible cracks or irregularities. Therefore, the presence of the central crack in the first image indicates a defect. Therefore, the correct answer is B. <\think>

Question: Is there any defect in the object?

A: No.
B: Yes.

JUDO Response:

<seg> (5,4)-(5,7) , (6,3)-(6,7) , (7,3)-(7,7) , (8,3)-(8,7) , (9,3)-(9,7) , (10,3)-(10,7) , (11,3)-(11,7) , (12,3)-(12,7) <\seg>

<think> The defect in the first image involves an empty fruit compartment where two fruits are missing, resulting in an imbalance compared to the normal second image. This discrepancy affects the expected arrangement and variety of items in the breakfast box. Therefore, the correct answer is B. <\think>

Question: Is there any defect in the object?

A: No.
B: Yes.

JUDO Response:

<seg> None <\seg>

<think> The first image shows macaroni pieces with minor surface scratches or abrasions, which are typical manufacturing marks and do not indicate defects. These imperfections are common and do not affect the quality. The second image displays macaroni pieces without any visible imperfections, maintaining a consistent appearance. Therefore, the defect in the first image is characterized by these minor surface marks rather than any structural anomalies. Therefore, the correct answer is A. <\think>

Figure 9: Examples of JUDO's output on MMAD dataset. The anomalous region in the query image is highlighted in green while the segmentation output from JUDO is represented as red patches in the Segmented Image