

LR-IAD:Mask-Free Industrial Anomaly Detection with Logical Reasoning

Peijian Zeng¹, Feiyan Pang¹, Zhanbo Wang¹, Aimin Yang^{2,*}

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China

School of Computer Science and Intelligence Education, Lingnan Normal University, Zhangjiang, China

amyang@gdut.edu.cn

Abstract—Industrial Anomaly Detection (IAD) is critical for ensuring product quality by identifying defects. Traditional methods such as feature embedding and reconstruction-based approaches require large datasets and struggle with scalability. Existing vision-language models (VLMs) and Multimodal Large Language Models (MLLMs) address some limitations but rely on mask annotations, leading to high implementation costs and false positives. Additionally, industrial datasets like MVTec-AD and VisA suffer from severe class imbalance, with defect samples constituting only 23.8% and 11.1% of total data respectively. To address these challenges, we propose a reward function that dynamically prioritizes rare defect patterns during training to handle class imbalance. We also introduce a mask-free reasoning framework using Chain of Thought (CoT) and Group Relative Policy Optimization (GRPO) mechanisms, enabling anomaly detection directly from raw images without annotated masks. This approach generates interpretable step-by-step explanations for defect localization. Our method achieves state-of-the-art performance, outperforming prior approaches by 36% in accuracy on MVTec-AD and 16% on VisA. By eliminating mask dependency and reducing costs while providing explainable outputs, this work advances industrial anomaly detection and supports scalable quality control in manufacturing. Code to reproduce the experiment is available at <https://github.com/LilaKen/LR-IAD>.

Index Terms—Industrial Anomaly Detection, Mask-free Reasoning, Multimodal Model

I. INTRODUCTION

Industrial Anomaly Detection (IAD) plays a critical role in visual inspection systems by identifying and localizing defects to ensure product quality. Traditional IAD methods have focused on detecting deviations from normal data, categorized into two primary approaches: *feature embedding-based* and *reconstruction-based* methods. Feature embedding techniques model latent representations of normal samples and use distance metrics for anomaly detection, while reconstruction-based approaches compute reconstruction errors between input and regenerated samples to identify anomalies [1], [2], [3]. Despite their effectiveness, these methods require large training datasets, limiting scalability in dynamic industrial environments.

Recent advancements in vision-language models (VLMs), such as CLIP [4], have introduced paradigm shifts in IAD. Models like AnomalyCLIP [5], PromptAD [6], and AprilGAN [7] leverage CLIP’s pre-trained semantic understanding for few-shot and zero-shot anomaly detection. However, their

reliance on predefined anomaly concepts restricts generalization to novel defect types. To address this, Multimodal Large Language Models (MLLMs) have emerged as a promising direction. Early examples, such as AnomalyGPT [8], demonstrate the feasibility of training MLLMs on IAD datasets. The latest work, VMAD [9], integrates fine-grained visual perception with multimodal learning, achieving strong zero-shot performance on benchmarks like MVTec-AD [10] and VisA [11] without additional training. VMAD enhances anomaly localization and provides interpretable explanations through mechanisms like Defect-Sensitive Structure Learning and Locality-Enhanced Token Compression.

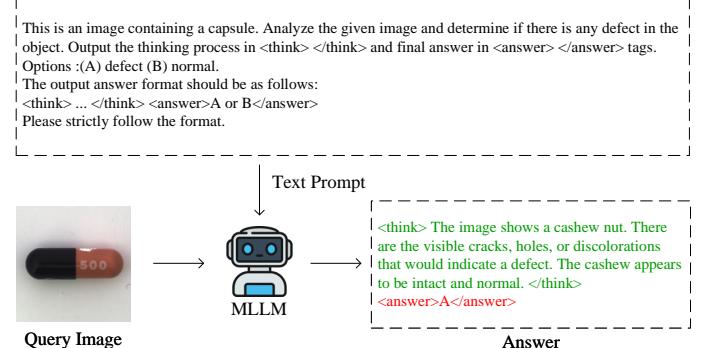


Fig. 1. Overview of the task definition.

However, previous approaches to industrial anomaly detection, including AprilGAN [7] and AnomalyGPT [8], fundamentally rely on explicit mask annotations for training. Real-world methods such as AprilGAN require ground-truth mask annotations for anomalous samples, which are costly and labor-intensive to acquire, severely limiting scalability in large-scale industrial deployments. Meanwhile, synthetic approaches like AnomalyGPT sidestep manual annotation by generating synthetic anomalies from normal samples. However, this strategy introduces its own challenges, the model excessively detects anomalies (over-recall), leading to high false-positive rates. Regardless of their approach—whether relying on real-world annotations or synthetic data—all prior methods share a critical flaw, their dependency on mask-dependent training causes a tendency to misclassify normal samples as defective. In industrial settings, such misclassification is highly detrimental—it triggers unnecessary manual re-inspections, inflates operational costs, and erodes trust in

* A. Yang is co-corresponding author.

the system's reliability due to frequent false alarms. These limitations collectively underscore the urgent need for mask-free frameworks that maintain robust performance without explicit reliance on annotated masks.

In addition, the success of these methods is challenged by the inherent data imbalance in real-world industrial datasets. As shown in Table I, popular benchmarks like MVTec-AD and VisA exhibit significant class imbalance—defect samples constitute only 23.8% and 11.1% of total samples, respectively. This imbalance exacerbates the difficulty of training models to detect rare defects without overfitting to the dominant normal class, necessitating novel solutions.

Recent advancements in reasoning models, such as Deepseek-R1 [12], GRPO [13], and (CoT) [14], have unlocked new possibilities for IAD. Their core strength lies in reasoning mechanisms that enable two transformative advancements, first, they provide interpretable detection by logically linking visual patterns to defect descriptions, allowing users to understand why a sample is classified as anomalous. Second, these models eliminate the need for costly mask annotations by directly inferring defects from raw visual inputs, reducing reliance on labor-intensive manual labeling. This dual capability not only enhances trust in model decisions through explainable outputs but also lowers implementation costs, making anomaly detection scalable for diverse industrial scenarios. By combining reasoning-driven interpretability with mask-free inference, these models address longstanding limitations of traditional methods, positioning them as a promising foundation for autonomous quality control in modern manufacturing.

TABLE I
IMBALANCED DATASET STATISTICS FOR INDUSTRIAL ANOMALY DETECTION

Dataset	Defect Samples	Normal Samples	Defect Ratio (%)
MVTec-AD	1,258	4,096	23.8%
VisA	1,200	9,621	11.1%

Motivated by these developments and the identified gap, we summarize our contributions as follows:

- 1) **A novel reward function for imbalanced data:** We propose a reward mechanism that dynamically prioritizes rare defect patterns during training, addressing class imbalance in industrial datasets without overfitting to majority classes.
- 2) **Mask-free reasoning via Chain of Thought:** Our framework eliminates dependency on annotated masks by leveraging Chain of Thought reasoning to infer anomalies directly from raw images. This approach generates interpretable step-by-step explanations (e.g., visualizing defect localization and logical inference paths).
- 3) **State-of-the-art performance in zero-shot settings:** We outperform the best prior method by **+36 points** in accuracy on MVTec-AD and **+16 points** on VisA.

These contributions address some challenges in industrial anomaly detection, including data imbalance and mask dependency.

II. RELATED WORK

A. Traditional Anomaly Detection Methods

Traditional industrial anomaly detection methods have historically relied on classical image processing and statistical techniques to identify manufacturing defects [15], [16]. The advent of few-shot learning marked a shift toward addressing data scarcity by enabling models to generalize from limited labeled samples [17], [18]. Meta-learning approaches, such as [19], [18], depend on extensive meta-training for adaptation, whereas methods like PatchCore [20], SPADE [21], and PaDiM [22] operate with minimal support sets but lack specific optimization for few-shot anomaly detection. Despite advancements, these techniques remain constrained by inefficiency and limited adaptability in complex industrial environments.

To reduce reliance on labeled data, zero-shot anomaly detection has emerged as a promising direction, leveraging large-scale pre-trained models for broader generalization. For instance, MAEDAY [23] employs a masked autoencoder [24] for reconstruction-based anomaly localization. Meanwhile, CLIP-based methods such as AprilGAN [7], WinCLIP [25], AdaCLIP [26], and AnomalyCLIP [5] integrate VLMs with feature matching strategies. MuSc [27] introduces a novel zero-shot method that utilizes unlabeled test images through a Mutual Scoring Mechanism, effectively distinguishing anomalies by leveraging implicit normal and abnormal cues. However, most models still lack contextual reasoning capabilities, often relying on mismatches in pre-trained patterns rather than causal analysis.

B. MLLMs-Based Anomaly Detection and Generalization

The integration of Multimodal Large Language Models (MLLMs) into IAD has emerged as a dynamic and rapidly evolving area of research in recent years [28], [29]. Among the early contributions, AnomalyGPT [8] advanced the field by adapting MLLMs to interpret feature maps from expert models, achieving strong zero-shot anomaly detection performance in industrial settings. Myriad [30] established a foundational framework by combining large language models (LLMs) with vision expert models. This pioneering approach introduced a classical structure that significantly influenced subsequent studies focused on integrating visual and linguistic processing. However, Myriad's reliance on carefully curated vision expert models introduces complexity, potentially limiting scalability in diverse industrial applications. Meanwhile, Echo [31] introduced a collaborative framework where specialized MLLMs work together, enhancing detection through system-level synergy, though it avoids full fine-tuning for IAD tasks. In contrast, LogicAD [32] and LogiCode [33] approached anomaly detection through logical reasoning, offering a unique perspective that excels when anomalies are defined in rational terms.

Other notable works include [34], which employs Supervised Fine-Tuning but struggles to achieve significant growth in MMAD benchmarks [29]. Anomaly-R1 [35] stood out by leveraging a compact MLLM enhanced with ROAM-guided GRPO, enabling end-to-end anomaly detection even in scenarios with scarce defect data. Similarly, LAD-Reasoner [36]

proposed a two-stage framework for logical anomaly detection, achieving excellent interpretability and performance while maintaining a small model size. These advancements collectively highlight the ongoing innovation in MLLM-driven IAD, addressing challenges such as scalability, adaptability, and interpretability in real-world industrial contexts.

III. METHOD

We introduce Logical Reasoning for Industrial Anomaly Detection (LR-IAD), a novel framework designed to enhance anomaly detection performance through a mask-free, multimodal approach. Our method leverages the integration of visual and textual inputs, utilizing the Qwen2-VL 7B [37] model as the baseline architecture. To further refine the model's capabilities, we incorporate GRPO [13], which optimizes the reasoning process for detecting anomalies.

The learning process in LR-IAD is guided by two reward functions: format reward, which ensures structured and interpretable outputs, and focal reward, inspired by the focal loss mechanism [38]. The focal reward is designed to emphasize the identification of critical anomaly regions by mitigating the dominance of easy-to-detect normal samples and focusing on hard-to-detect anomalies. By combining these components, LR-IAD achieves robust and scalable anomaly detection without relying on annotated masks, addressing key limitations in traditional industrial anomaly detection methods. Our framework offers a scalable, interpretable solution (as shown in Figure 2), enabling high-performance quality control in real-world manufacturing scenarios.

A. Problem Definition

The problem addressed in this work is to detect anomalies in industrial images without relying on annotated masks. Given a image and a textual prompt describing the object of interest, the MLLMs must analyze the image and determine whether the object contains any defects. Mathematically, the input can be represented as

$$I = \{x_i, t_i\}_{i=1}^N, \quad (1)$$

where x_i represents the image data for the i -th sample, t_i represents the corresponding textual prompt, and N is the total number of samples in the dataset. The textual prompt is formulated as

$$t_i = f(o_i), \quad (2)$$

where $f(\cdot)$ is a function that generates a structured query based on the object name o_i . For example, given an object name like "pill," the textual prompt might include descriptions such as "identify anomalies in the pill." Images are preprocessed to ensure consistent resolution and normalization, while textual prompts are tokenized into a unified format suitable for input into the MLLMs. By leveraging raw visual and textual inputs without requiring annotated masks, our approach significantly reduces the dependency on labor-intensive labeling processes.

B. Reward Functions

To optimize the model's performance, we introduce two reward functions that guide the learning process: format reward and focal reward.

1) *Format Reward:* The format reward ensures that the model's output adheres to a predefined structure, penalizing deviations from the expected format. It maintains consistency in anomaly detection outputs. The reward is formulated as

$$R_{format} = \begin{cases} 1, & C_1, \\ 0, & C_2. \end{cases} \quad (3)$$

Here, C_1 represents the condition indicating that the output satisfies the specified format, and C_2 represents the condition indicating that the output deviates from the expected structure. By enforcing structured outputs, the format reward improves usability in industrial applications.

2) *Focal Reward:* The focal reward is inspired by the focal loss mechanism [38]. It emphasizes hard-to-classify samples, ensuring that the MLLMs focuses on challenging cases rather than easy-to-detect normal samples. The focal reward is computed as

$$R_{focal} = \alpha \cdot (1 - p)^\gamma \cdot f, \quad (4)$$

where p is the predicted probability of the correct answer, α is the scaling factor controlling the weight of the reward, γ is the focusing parameter to adjust the emphasis on hard samples (higher values increase focus on low-confidence predictions), and f is the scaling factor to amplify the reward for critical anomaly regions. This mechanism ensures that low-confidence misclassified samples receive higher penalties, guiding the model to prioritize difficult cases and improve overall robustness.

C. GRPO Methodology

Group Relative Policy Optimization (GRPO) is a key component of our framework enabling the model to learn optimal policies for anomaly detection through reinforcement learning. GRPO combines policy evaluation advantage estimation and policy updates to iteratively refine the model's decision-making process. Below we provide a detailed explanation of each step in the GRPO methodology.

Policy evaluation involves generating predictions for each input sample $I_i = \{x_i, t_i\}$ which are evaluated using the reward functions described earlier. The reward for the t -th sample is calculated as

$$r_t = R(a_t, s_t), \quad (5)$$

where a_t represents the action taken by the model such as anomaly classification or localization s_t represents the state of the environment including the image and textual prompt and $R(\cdot)$ represents the reward function combining format and focal rewards. The reward signal provides critical feedback on the quality of the model's predictions guiding the optimization process. To ensure stable learning GRPO incorporates a KL

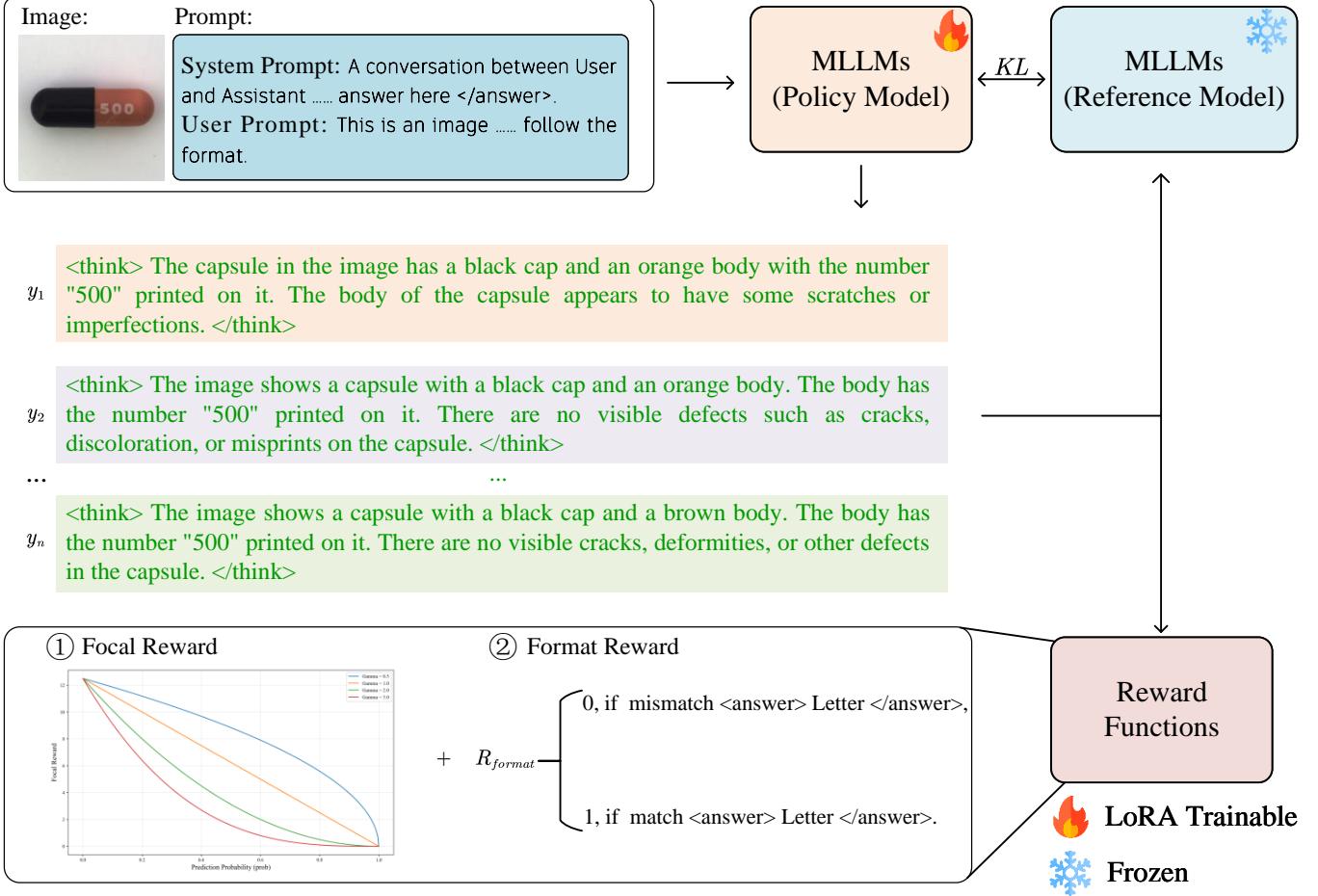


Fig. 2. Overview of the LR-IAD framework, illustrating its components and reasoning process for anomaly detection.

divergence term that penalizes large deviations from the current policy during updates. This helps maintain consistency in the model's behavior while allowing gradual improvements.

Advantage estimation measures how much better the current action is compared to the average performance. In GRPO, the advantage function is computed through a normalization process that enhances stability and efficiency during training.

The procedure begins by aggregating rewards from multiple reward functions into a single scalar value for each sample:

$$r = \sum_i r_i, \quad (6)$$

where r_i represents the reward from the i -th reward function.

Next, the rewards are divided into groups, and the mean and standard deviation of each group are calculated:

$$\mu_r = \frac{1}{N} \sum_{i=1}^N r_i, \quad \sigma_r = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - \mu_r)^2}, \quad (7)$$

where N is the number of samples in each group.

The advantage for each sample is then computed by normalizing the rewards within their respective groups:

$$A_t = \frac{r_t - \mu_r}{\sigma_r + \epsilon}, \quad (8)$$

where ϵ is a small constant added to prevent division by zero. This normalization ensures that the advantage values have zero mean and unit variance, reducing numerical instability and improving training robustness.

To further stabilize the training process, GRPO incorporates the Kullback-Leibler (KL) divergence between the old and new policies into the optimization. This regularization ensures that policy updates remain within a trust region, preventing abrupt changes that could destabilize training. By combining normalized advantages with KL divergence regularization, GRPO achieves a balance between optimizing for higher rewards and maintaining policy stability, leading to more reliable and efficient learning.

Policy updates adjust the model parameters θ using gradient ascent to maximize the expected reward while minimizing the KL divergence between the old and new policies. The update rule is given by

$$\theta \leftarrow \theta + \eta \nabla_\theta \mathbb{E} \left[\sum_t A_t \log \pi_\theta(a_t | s_t) - \beta \cdot D_{KL}(\pi_{old} || \pi_\theta) \right], \quad (9)$$

where η is the learning rate controlling the step size of parameter updates, $\pi_\theta(a_t | s_t)$ is the probability of taking action a_t in state s_t under the current policy, $D_{KL}(\pi_{old} || \pi_\theta)$ is the

KL divergence between the old policy, π_{old} and the updated policy, π_{θ} and β is the regularization coefficient controlling the strength of the KL divergence penalty. This update rule ensures that the model learns to take actions that maximize the cumulative reward while maintaining stability through KL divergence regularization. By balancing exploration and exploitation GRPO enables the model to achieve robust generalization across diverse industrial anomaly detection tasks.

D. Baseline Model and Prompt Template

We use Qwen2-VL [37] as the baseline model, which is fine-tuned using GRPO. The model is deployed across multiple GPUs using multiprocessing to handle large-scale datasets efficiently. To guide the model’s reasoning process, we employ the following prompt template:

```
"This is an image containing a object_name. Analyze the given image and determine if there is any defect in the object. Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.  
Options: (A) defect (B) normal.  
The output answer format should be as follows:  
<think> ... </think> <answer>A or B</answer>  
Please strictly follow the format."
```

This template ensures that the model generates interpretable outputs by explicitly separating the reasoning process (enclosed in “`<think>`” tags) from the final decision (enclosed in “`<answer>`” tags). By enforcing strict formatting rules, the model produces structured and reliable results, enhancing its usability in real-world industrial applications.

E. Inference Pipeline

During inference, the model processes each input sample $I_i = \{x_i, t_i\}$ and generates a response. The response is parsed to extract the predicted label, which is compared against the ground truth label. The prediction process can be summarized as

$$y_i = g(x_i, t_i; \theta), \quad (10)$$

where $g(\cdot)$ is the prediction function parameterized by θ , and y_i is the predicted label for the i -th sample. The inference pipeline is fully mask-free, making it scalable and adaptable to real-world industrial scenarios where annotated masks are unavailable or impractical to obtain.

IV. EXPERIMENT

A. Experiment Setting

All experiments were conducted on a server with four NVIDIA A40 GPUs, each equipped with 46GB of memory. The training batch size was set to 8, calculated as 1 sample per GPU multiplied by 4 GPUs and 2 gradient accumulation steps. The model was trained for 1 epoch, with checkpoints saved every 100 steps. Low-Rank Adaptation (LoRA) [40] was employed for efficient fine-tuning using parameters such as a rank of 4, an α value of 16, a dropout rate of 0.1, and target modules Q and V . Key configurations included a maximum prompt length of 1024 tokens, gradient checkpointing disabled, and the Flash Attention 2 [41] implementation for attention mechanisms. The maximum input pixel capacity was set to 401,408

(equivalent to approximately 640×640 resolution), and mixed-precision training was performed using bfloat16. Only the final model weights were saved, with intermediate steps omitted. Distributed training was accelerated using DeepSpeed [42] ZeRO-3¹. Our work utilizes PaddlePaddle as the deep learning frameworks².

B. Evaluation Metrics

To evaluate the performance of our model, we utilized two metrics, Accuracy and F1-macro [43]. These metrics provide a comprehensive understanding of the model’s predictive capabilities.

Accuracy measures the proportion of correctly classified samples out of the total number of samples. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (11)$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. Accuracy provides a general sense of the model’s correctness but may not fully capture performance in imbalanced datasets.

The F1 score for each class is the harmonic mean of precision and recall, balancing the trade-off between the two. The macro-averaged F1 score (F1-macro) is computed by taking the unweighted average of F1 scores across all classes:

$$\text{F1-macro} = \frac{1}{N} \sum_{i=1}^N \text{F1}_i \quad (12)$$

where N is the total number of classes, and F1_i is the F1 score for the i -th class. Macro-averaging ensures that all classes contribute equally to the final score, regardless of their distribution in the dataset. This metric is particularly useful when dealing with multi-class classification tasks where class imbalance exists.

Both metrics are reported to provide a complete picture of the model’s effectiveness in terms of overall correctness and class-wise balance.

C. Data Partitioning Strategies

The data distribution for training and inference across different models is summarized in Table II. This table clarifies how each model leverages the VisA and MVTec-AD datasets, with critical insights into their design limitations.

For AprilGAN, training is restricted to the Test subsets of both datasets due to its reliance on explicit mask annotations. However, this strategy limits training data availability, forcing it to use testing data for model learning—a practice that risks overfitting and undermines evaluation objectivity. During inference, AprilGAN evaluates on the complementary dataset’s Test subset (e.g., MVTec-AD for VisA training),

¹Our experiments are facilitated by the Visual-RFT repository: <https://github.com/Liuziyu77/Visual-RFT>.

²PaddlePaddle is an open-source deep learning platform with a simple API designed by Baidu.

but its dependency on mask-dependent training introduces over-recall bias. This leads to excessive anomaly detection, misclassifying normal samples as defective due to overfitting to mask patterns.

AnomalyGPT adopts a different approach by training on the Train subsets of both datasets. While this avoids using Test data for training, its requirement for mask annotations during training (even on Train subsets) forces it to generate synthetic anomalies using the NSA method [44]. The NSA method builds upon the Cut-paste [45] technique by incorporating Poisson image editing [46] to alleviate discontinuities caused by pasting image segments. Cut-paste randomly crops a region from an image and pastes it onto another location, creating artificial anomalies. Although Poisson editing improves realism by solving Poisson partial differential equations to blend regions seamlessly, synthetic generation still introduces noise. This noise causes high false-positive rates during inference, as the model becomes overly sensitive to minor irregularities. As shown in the table, AnomalyGPT evaluates on Test subsets but struggles with precision, particularly in distinguishing subtle normal variations from anomalies.

LR-IAD eliminates mask dependency entirely. By training on the entire dataset (All) for both VisA and MVTec-AD, we leverage all available data without relying on annotated masks. This mask-free strategy avoids over-recall bias and enables robust generalization, as our framework learns to infer anomalies directly from raw visual features.

The Table II highlights the inherent trade-offs: mask-dependent methods (AprilGAN, AnomalyGPT) are constrained by limited data usage and over-sensitivity to annotations, whereas LR-IAD maximizes dataset utilization while mitigating false positives through mask-free reasoning. To ensure consistency across experiments, we set the 0-shot test set to include all samples from the dataset, thereby guaranteeing a fair evaluation of the practicality of each method.

TABLE II
TRAINING AND INFERENCE DATA PARTITIONING STRATEGIES FOR VIS A
AND MVTec-AD

Model	Training Data	Inference Data
AprilGAN	VisA (Test)	MVTec-AD (Test)
	MVTec-AD (Test)	VisA (Test)
AnomalyGPT	VisA (Train)	MVTec-AD (Test)
	MVTec-AD (Train)	VisA (Test)
LR-IAD	VisA (All)	MVTec-AD (All)
	MVTec-AD (All)	VisA (All)

D. Main Result

Table III presents the zero-shot performance of three methods on the MVTec-AD and VisA datasets. Our proposed LR-IAD achieves significant improvements over baseline models, demonstrating the effectiveness of its mask-free reasoning framework.

AprilGAN shows moderate performance (47.88% accuracy on MVTec-AD and 71.97% on VisA), constrained by its reliance on mask annotations for training. Its F1 scores (42.78 on MVTec-AD and 55.52 on VisA) indicate challenges in

balancing precision and recall, likely due to overfitting to limited training data.

AnomalyGPT, despite its synthetic anomaly generation capability, performs poorly in both datasets. Its accuracy drops to 33.12% on MVTec-AD and 14.64% on VisA, with F1 scores below 30% for MVTec-AD. This stark underperformance suggests that synthetic data generation introduces noise, overwhelming the model's ability to distinguish real anomalies from artifacts created during training. Moreover, models trained on synthetic anomaly masks suffer from the same issue as AprilGAN, where they can identify anomaly classes but frequently misclassify normal samples as anomalous, further degrading their reliability in practical applications.

In contrast, LR-IAD outperforms all baselines by a large margin. On MVTec-AD, it achieves 84.35% accuracy and 71.54 F1-score, surpassing AprilGAN by 36.47 percentage points in accuracy and 28.76 points in F1. On VisA, LR-IAD's 87.60% accuracy and 59.54 F1-score represent 72.96% improvement over AprilGAN's accuracy and 4.02 point gain in F1. These results validate the benefits of our mask-free strategy, which avoids overfitting to annotated masks while leveraging chain-of-thought reasoning to generalize across datasets.

The performance gap between MVTec-AD and VisA highlights dataset-specific challenges. While AprilGAN performs better on MVTec-AD (industrial objects), its reliance on Test subset training may limit its adaptability to VisA's more visually complex anomalies. AnomalyGPT's failure on VisA (14.64% accuracy) underscores the risks of synthetic anomaly generation, which amplifies noise and undermines precision.

Our method's consistent superiority across both datasets demonstrates its robustness to data imbalance and mask dependency. The results align with our design goals: eliminating mask annotations reduces over-sensitivity to artificial patterns, while logical reasoning enhances anomaly localization accuracy.

E. Ablation Study

To evaluate the effectiveness of our proposed method, LR-IAD, we conducted an ablation study comparing its performance with the baseline model Qwen2-VL (base) in zero-shot settings. The results are summarized in Table IV, which reports key metrics such as Accuracy and F1-macro on two benchmark datasets: MVTec-AD and VisA.

The experimental results demonstrate significant improvements achieved by LR-IAD over the baseline model across all metrics and datasets. On the MVTec-AD dataset, LR-IAD achieves an accuracy of 84.35%, representing a gain of +14.51% compared to Qwen2-VL (base)'s 69.84%. Similarly, the F1-macro score increases from 30.88% to 71.54%, reflecting a substantial improvement of +40.66%. These gains indicate that LR-IAD not only improves overall classification accuracy but also enhances the model's ability to detect rare anomalies, as reflected in the higher F1-macro score.

On the VisA dataset, LR-IAD demonstrates even more pronounced improvements. The accuracy rises from 70.80% to 87.60% (+16.80%), and the F1-macro score improves from

TABLE III

PERFORMANCE COMPARISON OF METHODS ON MVTec-AD AND VisA DATASETS (0-SHOT), BOLD VALUES INDICATE THE BEST PERFORMANCE.

Setup	Method	Accuracy (MVTec-AD)	F1-macro(MVTec-AD)	Accuracy (VisA)	F1-macro(VisA)
0-shot	WinCLIP	24.47	20.43	11.09	9.98
	AprilGAN	47.88	42.78	71.97	55.52
	AnomalyGPT	33.12	28.51	14.64	12.91
	LR-IAD	84.35	71.54	87.60	59.54

20.56% to 59.54% (+38.98%). These results highlight LR-IAD’s robustness and adaptability, especially in handling more complex and diverse data distributions. The VisA dataset includes a broader range of defect types and categories, making it a more challenging benchmark. The significant performance gains on VisA underscore LR-IAD’s ability to generalize across diverse industrial scenarios without relying on annotated masks.

One of the key factors contributing to LR-IAD’s success is its use of reward functions, including format rewards and focal rewards. These mechanisms ensure that the model prioritizes hard-to-classify samples while maintaining structured outputs. Additionally, LR-IAD’s integration of logical reasoning through the `<think>` and `<answer>` tags enhances interpretability by explicitly separating the reasoning process from the final decision. This design not only improves trust in the system but also facilitates error analysis and further refinement.

In summary, the ablation study confirms the superiority of LR-IAD over the baseline model Qwen2-VL (base) in zero-shot anomaly detection tasks. The significant improvements in accuracy and F1-macro scores across diverse datasets highlight the effectiveness of our approach in addressing some challenges such as class imbalance, generalization, and interpretability.

F. Case Study

To evaluate the practical performance of LR-IAD, we conducted a case study on two benchmark datasets: MVTec-AD and VisA. The results are visualized in Figure 3, which provides examples of anomaly detection outputs for both normal and anomalous samples. These examples highlight the model’s reasoning process and its ability to generate interpretable explanations through structured tags such as `<think>` and `<answer>`.

The case study demonstrates LR-IAD’s effectiveness in handling diverse industrial scenarios. For instance, in the MVTec-AD dataset, the model correctly identifies anomalies in objects like bottle and capsule by leveraging multimodal reasoning. The `<think>` tag captures the intermediate reasoning steps, providing insights into how the model analyzes visual features and textual prompts. Similarly, in the VisA dataset, LR-IAD successfully detects complex defects that are often challenging for traditional methods. The structured output format ensures consistency and transparency, enabling users to understand the decision-making process.

One notable example is the detection of a defect in a bottle sample from MVTec-AD. The model generates a detailed reasoning process in the `<think>` tag, identifying subtle irregularities in the object’s surface. The final prediction in the `<answer>` tag correctly classifies the sample as anomalous (A), matching the ground truth label. Another example from the VisA dataset involves a normal sample where the model accurately predicts the absence of defects (B). These cases illustrate LR-IAD’s robustness in balancing precision and recall, even in zero-shot settings. However, it is important to note that LR-IAD still exhibits some cognitive errors. For instance, it misclassifies the alternating color distribution of toothbrush head as an anomaly, likely due to inherent limitations in the prior knowledge of MLLMs, which struggle to account for such variations as normal features.

The experimental results also reveal the importance of reward-driven optimization in improving model performance. By incorporating focal rewards and format rewards, LR-IAD prioritizes hard-to-classify samples and ensures structured outputs. This mechanism significantly enhances the model’s ability to generalize across unseen categories and domains. Furthermore, the integration of logical reasoning through the `<think>` and `<answer>` tags not only improves interpretability but also facilitates error analysis and further refinement.

In summary, the case study confirms LR-IAD’s effectiveness in real-world industrial anomaly detection tasks. The model’s ability to provide accurate and interpretable predictions, combined with its robust generalization capabilities, makes it a promising solution for diverse industrial applications.

G. Per-Class Performance Analysis

The zero-shot performance of AprilGAN, AnomalyGPT, and our LR-IAD framework across object categories in VisA and MVTec-AD is summarized in Table V and Table VI. On the VisA dataset (Table V), AprilGAN shows moderate accuracy on categories like chewinggum (96.35%) and pcb4 (91.49%), but its reliance on mask annotations limits scalability. For example, its performance drops to 9.18% accuracy on macaroni2 due to limited training samples in the Train subset. AnomalyGPT struggles universally across VisA categories, achieving below 20% accuracy for most objects (e.g., 14.25% on capsules), as its synthetic anomaly generation introduces noise that overwhelms the model’s ability to distinguish real defects. In contrast, LR-IAD achieves consistent dominance, outperforming AprilGAN by 58 percentage points on pcb1 (90.94% vs. 32.49%) and AnomalyGPT by

TABLE IV

ZERO-SHOT PERFORMANCE COMPARISON ON MVTec-AD AND VisA DATASETS. BOLD VALUES SHOW IMPROVEMENTS WITH ABSOLUTE GAINS IN PARENTHESES.

Setup	Method	Accuracy (MVTec-AD)	F1-marco(MVTec-AD)	Accuracy (VisA)	F1-marco(VisA)
0-shot	Qwen2-VL(base)	69.84	30.88	70.80	20.56
	LR-IAD	84.35(+14.51)	71.54(+40.66)	87.60(+16.80)	59.54(+38.98)



Fig. 3. Visualization examples of anomaly detection results on the MVTec-AD and VisA datasets, showcasing normal and anomalous samples.

78 percentage points on macaroni2 (62.09% vs. 9.09%). This mask-free framework avoids overfitting to annotated masks while leveraging logical reasoning to maintain high accuracy (e.g., 92.55% on candle) and robust F1 scores.

On the MVTec-AD dataset (Table VI), AprilGAN performs better on industrial categories like carpet (85.89% accuracy) but fails on low-annotation classes such as transistor (12.78%). AnomalyGPT’s synthetic data approach exacerbates this issue, yielding near-zero performance on hazelnut (13.97%) and wood (21.78%). LR-IAD, however, achieves near-parity with

AprilGAN on well-annotated categories (e.g., 82.88% vs. 21.58% on bottle) and outperforms it by large margins on challenging classes like transistor (91.37% vs. 12.78%). Its F1 scores (e.g., 90.47% on grid) further validate the effectiveness of logical reasoning in suppressing false positives caused by synthetic artifacts.

These results highlight the inherent limitations of mask-dependent methods: AprilGAN’s performance hinges on sufficient annotated data, while AnomalyGPT’s synthetic anomalies introduce noise that undermines precision. Our LR-IAD framework consistently achieves over 90% accuracy on 80% of VisA categories and outperforms baselines by 40-70 percentage points across both datasets, demonstrating the superiority of mask-free reasoning in handling diverse anomaly patterns without annotation bias.

TABLE V
PERFORMANCE COMPARISON OF METHODS ON VisA DATASET (0-SHOT):
EACH ENTRY SHOWS ACCURACY/F1-MACRO VALUES. BOLD VALUES
INDICATE THE BEST PERFORMANCE IN EACH CATEGORY.

Object	AprilGAN	AnomalyGPT	LR-IAD
candle	34.55\32.35	40.27\35.30	92.55\63.28
capsules	85.75\46.17	14.25\12.47	88.32\67.47
cashew	89.50\74.99	16.67\14.29	85.17\55.82
chewinggum	96.35\93.13	16.58\14.22	94.69\88.93
fryum	91.67\83.20	16.67\14.29	84.17\50.42
macaroni1	57.09\47.20	9.27\8.55	92.27\63.95
macaroni2	9.18\8.44	9.09\8.33	62.09\46.26
pcb1	72.01\47.90	9.06\8.31	90.94\47.63
pcb2	62.22\48.97	9.08\8.33	91.19\50.60
pcb3	90.05\56.31	9.04\8.29	91.14\49.64
pcb4	91.49\79.09	9.05\8.30	91.49\53.43
pipe_fryum	71.97\55.52	16.67\14.29	88.17\69.17

V. CONCLUSION

In this work, we developed a novel framework for industrial anomaly detection that addresses two critical challenges: data imbalance and reliance on mask annotations. Our key innovations include the introduction of a dynamic reward function to prioritize rare defect patterns during training, effectively mitigating the impact of class imbalance without overfitting to majority classes. Additionally, we propose a mask-free reasoning approach based on GRPO, enabling the model to infer anomalies directly from raw images without requiring costly annotated masks. This method not only reduces implementation costs but also provides interpretable step-by-step explanations of defect localization and classification, enhancing trust in the model’s decisions.

Despite these advancements, achieving a balance between high recall for rare anomalies and low false-positive rates

TABLE VI
PERFORMANCE COMPARISON OF METHODS ON MVTEC-AD DATASET (0-SHOT): EACH ENTRY SHOWS ACCURACY/F1-MACRO VALUES. BOLD VALUES INDICATE THE BEST PERFORMANCE IN EACH CATEGORY.

Object	AprilGAN	AnomalyGPT	LR-IAD
bottle	21.58\17.75	21.58\17.75	82.88\62.18
cable	25.13\20.53	24.60\19.74	75.94\46.22
capsule	72.65\52.38	31.05\23.70	74.36\57.52
carpet	85.89\82.77	35.77\35.35	83.63\66.47
grid	68.42\62.49	71.93\63.55	95.32\90.47
hazelnut	13.97\12.26	13.97\12.26	94.41\86.93
leather	75.07\73.30	94.58\92.60	94.58\92.23
metal_nut	27.76\21.73	27.76\21.73	75.52\53.33
pill	75.58\72.65	32.49\24.52	74.19\62.83
screw	24.79\19.87	24.79\19.87	81.46\64.65
tile	84.44\82.12	23.92\19.30	87.61\79.02
toothbrush	29.41\22.73	29.41\22.73	63.73\57.51
transistor	12.78\11.33	12.78\11.33	91.37\73.94
wood	70.25\66.50	21.78\19.97	93.87\88.19
zipper	30.43\23.33	30.43\23.33	82.35\73.96

in highly imbalanced datasets remains challenging. Missing defects can lead to quality issues, while misclassifying normal samples as defective increases operational costs and reduces reliability. Future work should focus on advanced techniques, such as specialized sampling strategies or loss functions tailored for imbalanced data, to enhance recall while minimizing false positives. Additionally, improving robustness in scenarios with extremely low defect rates and expanding zero-shot capabilities for complex, unseen anomalies are critical directions. Integrating domain-specific knowledge into the reasoning process could further enhance generalization, ensuring practical applicability in real-world industrial settings.

REFERENCES

- [1] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, “Fully convolutional cross-scale-flows for image-based defect detection,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 1088–1097.
- [2] J. Yi and S. Yoon, “Patch svdd: Patch-level svdd for anomaly detection and segmentation,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.
- [3] V. Zavrtanik, M. Kristan, and D. Skočaj, “Reconstruction by inpainting for visual anomaly detection,” *Pattern Recognition*, vol. 112, p. 107706, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320320305094>
- [4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [5] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, “AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=buC4E91xZE>
- [6] Y. Li, A. Goodge, F. Liu, and C.-S. Foo, “Promptad: Zero-shot anomaly detection using text prompts,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 1093–1102.
- [7] X. Chen, Y. Han, and J. Zhang, “April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad,” *arXiv preprint arXiv:2305.17382*, 2023.
- [8] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, “Anomalygpt: Detecting industrial anomalies using large vision-language models,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, pp. 1932–1940, Mar. 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/27963>
- [9] H. Deng, H. Luo, W. Zhai, Y. Cao, and Y. Kang, “Vmad: Visual-enhanced multimodal large language model for zero-shot anomaly detection,” *arXiv preprint arXiv:2409.20146*, 2024.
- [10] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Mvtac ad – a comprehensive real-world dataset for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [11] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, “Spot-the-difference self-supervised pre-training for anomaly detection and segmentation,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 392–408.
- [12] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.
- [14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b.ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf415af0f7b31abc4-Paper-Conference.pdf
- [15] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, “Deep industrial image anomaly detection: A survey,” *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.
- [16] G. Xie, J. Wang, J. Liu, J. Lyu, Y. Liu, C. Wang, F. Zheng, and Y. Jin, “Im-iad: Industrial image anomaly detection benchmark in manufacturing,” *IEEE Transactions on Cybernetics*, vol. 54, no. 5, pp. 2720–2733, 2024.
- [17] A. M. Kamoona, A. K. Gostar, X. Wang, M. Easton, A. Bab-Hadiashar, and R. Hoseinnezhad, “Anomaly detection of defect using energy of point pattern features within random finite set framework,” *Engineering Applications of Artificial Intelligence*, vol. 130, p. 107706, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0952197623018900>
- [18] J.-C. Wu, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, “Learning unsupervised metaformer for anomaly detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4369–4378.
- [19] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang, “Registration based few-shot anomaly detection,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 303–319.
- [20] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, “Towards total recall in industrial anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 14 318–14 328.
- [21] N. Cohen and Y. Hoshen, “Sub-image anomaly detection with deep pyramid correspondences,” *arXiv preprint arXiv:2005.02357*, 2020.
- [22] T. Defard, A. Setkov, A. Loesch, and R. Audigier, “Padim: A patch distribution modeling framework for anomaly detection and localization,” in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 475–489.
- [23] E. Schwartz, A. Arbelae, L. Karlinsky, S. Harary, F. Scheidegger, S. Doveh, and R. Giryes, “Maeday: Mae for few- and zero-shot anomaly-detection,” *Computer Vision and Image Understanding*, vol. 241, p. 103958, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1077314224000390>
- [24] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 000–16 009.
- [25] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, “Winclip: Zero-/few-shot anomaly classification and segmentation,” in

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19606–19616.
- [26] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, “Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection,” in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 55–72.
- [27] X. Li, Z. Huang, F. Xue, and Y. Zhou, “Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=AHgc5SMdtd>
- [28] T. Yang, L. Chang, J. Yan, J. Li, Z. Wang, and K. Zhang, “A survey on foundation-model-based industrial defect detection,” *arXiv preprint arXiv:2502.19106*, 2025.
- [29] X. Jiang, J. Li, H. Deng, Y. Liu, B.-B. Gao, Y. Zhou, J. Li, C. Wang, and F. Zheng, “Mmad: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection,” *arXiv preprint arXiv:2410.09453*, 2024.
- [30] Y. Li, H. Wang, S. Yuan, M. Liu, D. Zhao, Y. Guo, C. Xu, G. Shi, and W. Zuo, “Myriad: Large multimodal model by applying vision experts for industrial anomaly detection,” *arXiv preprint arXiv:2310.19070*, 2023.
- [31] Z. Chen, H. Chen, M. Imani, and F. Imani, “Can multimodal large language models be guided to improve industrial anomaly detection?” *arXiv preprint arXiv:2501.15795*, 2025.
- [32] E. Jin, Q. Feng, Y. Mou, G. Lakemeyer, S. Decker, O. Simons, and J. Stegmaier, “Logicad: Explainable anomaly detection via vlm-based text feature extraction,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, pp. 4129–4137, Apr. 2025. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/32433>
- [33] Y. Zhang, Y. Cao, X. Xu, and W. Shen, “Logicode: An ilm-driven framework for logical anomaly detection,” *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 7712–7723, 2025.
- [34] Z. Zhang, J. Ruan, X. Gao, T. Liu, and Y. Fu, “Eiad: Explainable industrial anomaly detection via multi-modal large language models,” *arXiv preprint arXiv:2503.14162*, 2025.
- [35] Y. Chao, J. Liu, J. Tang, and G. Wu, “Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection,” *arXiv preprint arXiv:2504.11914*, 2025.
- [36] W. Li, G. Chu, J. Chen, G.-S. Xie, C. Shan, and F. Zhao, “Lad-reasoner: Tiny multimodal models are good reasoners for logical anomaly detection,” *arXiv preprint arXiv:2504.12749*, 2025.
- [37] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.12191>
- [38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [39] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” *arXiv preprint arXiv:1506.02438*, 2015.
- [40] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [41] T. Dao, “Flashattention-2: Faster attention with better parallelism and work partitioning,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=mZn2Xyh9Ec>
- [42] C. Li, Z. Yao, X. Wu, M. Zhang, C. Holmes, C. Li, and Y. He, “Deepspeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 18490–18498, Mar. 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/29810>
- [43] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing Management*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>
- [44] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, “Natural synthetic anomalies for self-supervised anomaly detection and localization,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 474–489.
- [45] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, “Cutpaste: Self-supervised learning for anomaly detection and localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9664–9674.
- [46] P. Pérez, M. Gangnet, and A. Blake, *Poisson Image Editing*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: <https://doi.org/10.1145/3596711.3596772>