Towards Explainable Industrial Anomaly Detection via Knowledge-Guided Latent Reasoning

Peng Chen[1] **Wenqiang Wang**[1] **Yunkang Cao**[2] **Chengliang Liu**[3]
Wei Wang[1] **Mingbo Yang**[1] **Li Shen**[1] **Wenqi Ren**[1]
Xiaochun Cao[1] **Chao Huang**[1*]

[1]Shenzhen Campus of Sun Yat-sen University [2]Hunan University [3]University of Macau

{chenp299, wangwq69, yangmb3}@mail2.sysu.edu.cn
liucl1996@163.com, caoyunkang0207@gmail.com
{wangwei29, shenli6, renwq3, caoxiaochun, huangch253}@mail.sysu.edu.cn

Abstract

Industrial anomaly detection demands precise reasoning over fine-grained defect patterns. However, existing multimodal large language models (MLLMs), pretrained on general-domain data, often struggle to capture category-specific anomalies, thereby limiting both detection accuracy and interpretability. To address these limitations, we propose Reason-IAD, a knowledge-guided dynamic latent reasoning framework for explainable industrial anomaly detection. Reason-IAD comprises two core components. First, a retrieval-augmented knowledge module incorporates category-specific textual descriptions into the model input, enabling context-aware reasoning over domain-specific defects. Second, an entropy-driven latent reasoning mechanism conducts iterative exploration within a compact latent space using optimizable latent think tokens, guided by an entropy-based reward that encourages confident and stable predictions. Furthermore, a dynamic visual injection strategy selectively incorporates the most informative image patches into the latent sequence, directing the reasoning process toward regions critical for anomaly detection. Extensive experimental results demonstrate that Reason-IAD consistently outperforms state-of-the-art methods across multiple tasks. The code will be publicly available at https://github.com/chenpeng052/Reason-IAD.

1 Introduction

Industrial anomaly detection plays a critical role in ensuring the safety and reliability of modern manufacturing systems [1, 2, 3], aiming to accurately identify anomalous samples that deviate from normal patterns in complex visual data [4, 5, 6]. However, conventional methods typically rely on carefully curated, task-specific datasets, rendering them highly sensitive to domain priors and limiting their generalization across diverse scenarios and product categories [7, 8, 9]. In practice, substantial distribution discrepancies across different industrial products and imaging conditions introduce severe domain shifts, hindering their deployment in real-world manufacturing environments.

In recent years, vision-language models have demonstrated considerable potential for industrial anomaly detection due to their strong multimodal representation and generalization capabilities [10, 11, 12]. CLIP-based approaches align visual features with textual semantics, enabling anomaly detection under zero-shot and few-shot settings [13, 14]. For instance, AnomalyCLIP [15] introduces category-agnostic textual embeddings to reduce reliance on manually designed priors,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

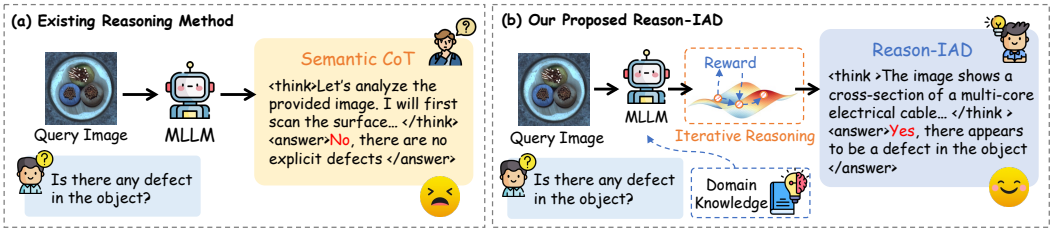


Figure 1: Comparison between existing reasoning methods and the proposed Reason-IAD. (a) Existing methods conduct reasoning through explicit chains of thought. (b) Reason-IAD retrieves domain-specific knowledge and identifies anomalies via iterative latent reasoning.

while AdaCLIP [7] and BayPFL [8] replace static prompts with learnable textual embeddings to enhance semantic expressiveness and model adaptability. Despite these advances, the performance of such methods remains constrained by the representational capacity of pretrained models [16, 17, 5]. When confronted with complex industrial scenes and fine-grained defect patterns, they often fail to achieve reliable anomaly recognition [18, 10, 19, 20]. Moreover, these approaches generally lack explicit reasoning mechanisms, preventing evidence-based analysis as human experts would do, thus limiting their applicability in high-reliability industrial scenarios.

Recent advances in multimodal large language models (MLLMs) have opened a promising avenue for developing general-purpose, expert-level industrial anomaly detection systems [21, 22, 23], owing to their strong pretraining and reasoning capabilities [13, 24, 25]. MMAD [1] introduced the first systematic benchmark for anomaly detection reasoning, evaluating models' reasoning abilities across multiple dimensions. Building on this, Anomaly-R1 [26] leverages the GRPO training strategy to construct an end-to-end anomaly detection framework that replaces traditional expert systems, while AD-FM [27] adopts a two-stage reasoning paradigm, combining supervised fine-tuning (SFT) with GRPO to progressively adapt pretrained multimodal models to anomaly detection tasks. However, as illustrated in Figure 1 (a), these approaches typically rely on explicitly generating long reasoning chains (chain-of-thought) to reach decisions, resulting in substantial inefficiency and limited scalability [28, 29, 30]. Furthermore, since the underlying models are primarily pretrained on general-domain data, they often lack task-specific domain knowledge for industrial anomaly detection, which constrains the reliability and stability of their reasoning outcomes.

To address these challenges, we propose Reason-IAD, a training-free framework for explainable industrial anomaly detection that leverages knowledge-guided dynamic latent reasoning. As illustrated in Figure 1(b), Reason-IAD consists of two key components. First, a retrieval-augmented knowledge module incorporates category-specific textual descriptions into the model input, enabling context-aware reasoning over domain-specific defects. Second, an entropy-driven latent reasoning mechanism conducts iterative exploration within a compact latent space using optimizable latent think tokens, facilitating confident and stable reasoning. In addition, a dynamic visual injection strategy selectively injects the most informative image regions into the latent sequence, guiding the reasoning process toward critical visual evidence for anomaly detection. We evaluate Reason-IAD across seven sub-tasks under one-shot and zero-shot settings. Extensive experiments demonstrate that Reason-IAD consistently outperforms existing methods, highlighting its effectiveness and interpretability.

Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first work to leverage MLLMs for training-free reasoning in the context of industrial anomaly detection.
- We propose Reason-IAD, a knowledge-guided dynamic latent reasoning framework for explainable industrial anomaly detection, enabling effective anomaly perception through iterative reasoning guided by visual evidence and entropy-driven reward optimization.
- Extensive experiments demonstrate the effectiveness and generality of Reason-IAD, providing an efficient and practical solution for applying MLLMs to industrial anomaly detection.

2 Related Work

2.1 Anomaly Detection

Anomaly detection aims to identify abnormal patterns that deviate from the normal data distribution [27, 31, 32]. However, substantial variability in appearance and the scarcity of anomalous samples make anomaly detection a particularly challenging task [33, 34, 35]. Traditional unsupervised methods learn representations solely from normal samples and identify anomalies during testing [36, 37, 38]. While effective to some extent, these approaches often rely on category-specific models, which constrain their generalization and practical applicability in real-world industrial scenarios. Recent studies have explored vision–language models for zero-shot and few-shot anomaly detection by aligning visual representations with textual semantics [10, 39, 40]. For example, WinCLIP [16] leverages CLIP [41] to measure similarity between image patch features and hand-crafted text prompts. To improve cross-modal alignment, AdaCLIP [7] and BayPFL [8] further incorporate image information into the prompt construction process, enhancing the model's capacity to capture anomaly-related semantics. Despite the performance gains of these approaches, their inference pipelines predominantly rely on static similarity matching, and the limited reasoning ability of base models remains a key bottleneck.

2.2 Visual Reasoning

Visual reasoning is a fundamental challenge in artificial general intelligence, requiring models to perform complex cognitive reasoning grounded in visual perception [42, 43]. This capability is essential for a wide range of tasks, including visual counting, geometric problem solving, and anomaly detection [44, 45, 46]. Early approaches primarily relied on program synthesis or neuro-symbolic reasoning frameworks [47, 48]. Recent advances in vision–language models have enabled the integration of large language models to enhance visual reasoning. Representative works such as Reason-RFT [24] adopt a two-stage training strategy combining supervised fine-tuning and reinforcement learning to improve visual reasoning. In anomaly detection, Anomaly-R1 [26] introduces GRPO-based reasoning alignment, while AD-FM [27] bridges anomaly perception and explanation. Triad [49] further incorporates expert-guided region awareness with chain-of-thought reasoning to improve defect understanding. However, most existing approaches rely heavily on large-scale reconstructed datasets for fine-tuning, resulting in high training costs and limited scalability in real-world applications.

2.3 Latent Reasoning

Recent research on advanced reasoning in multimodal large language models has evolved from explicit text-based reasoning toward more efficient latent space reasoning approaches [28, 50, 51]. In latent space reasoning, the intermediate "thinking process" is encoded as continuous vectors in the model's hidden space rather than as discrete textual representations, aiming to achieve performance comparable to explicit chain-of-thought (CoT) reasoning while using a more compact and efficient representation [52, 53, 54]. Some methods enhance latent representations through carefully designed training strategies to enable more effective interactive reasoning [55, 56, 57], while others adopt training-free approaches that guide latent activations to refine the reasoning process [58, 59, 60]. However, existing methods generally lack knowledge-guided reasoning and the integration of visual context, limiting their effectiveness on tasks that require external knowledge or visual evidence. In contrast, our approach retrieves relevant knowledge and dynamically integrates visual evidence into the latent reasoning process, improving performance in industrial anomaly detection.

3 Methodology

3.1 Problem Definition

In this work, we study a multimodal anomaly detection question-answering (QA) task, in which the model answers questions based on a given industrial image.

Formally, given a query image $I \in \mathcal{I}_{test}$ depicting an industrial object and a question $Q \in \mathcal{Q}_{test}$, the model aims to predict the correct answer A. Each question corresponds to a specific anomaly reasoning objective, demanding varying levels of visual perception and reasoning.

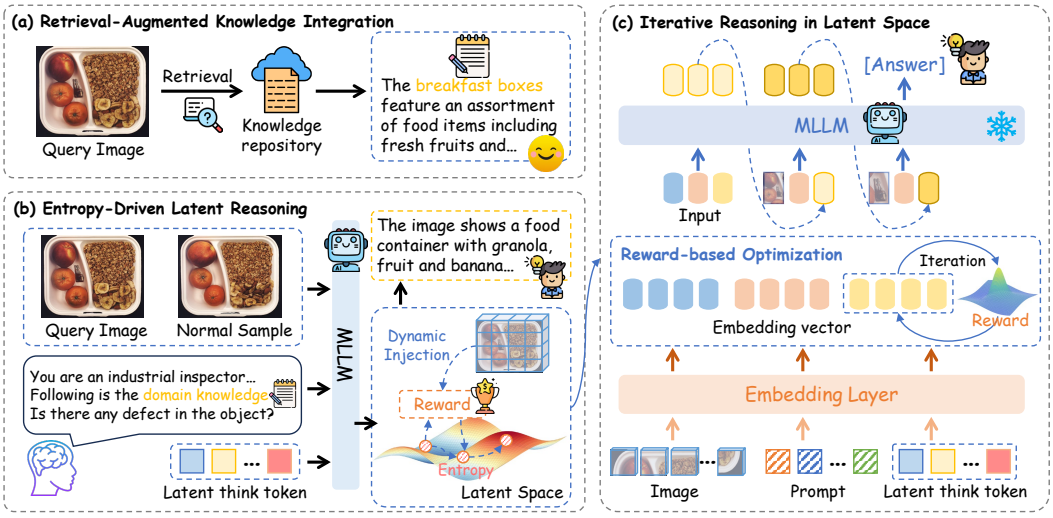


Figure 2: Overview of the proposed Reason-IAD. (a) Given a query image, Reason-IAD retrieves the most relevant category-specific descriptions and incorporates them into the model prompt to enhance anomaly awareness. (b) An entropy-guided latent reasoning module iteratively refines latent think tokens and dynamically injects visual evidence to improve reasoning accuracy. (c) Illustration of the iterative latent-space reasoning process, in which reward signals and visual cues progressively guide the model toward the final prediction.

Our proposed method supports two settings: (1) **One-shot setting**: the model receives a query image I and a question Q, along with a reference image $I_r \in \mathcal{I}_{ref}$ of a normal object from the same domain. The model is required to generate both the answer and a corresponding explanation. (2) **Zero-shot setting**: the model only receives the query image I and the question Q, without any reference information, and is expected to produce the answer along with the relevant explanation.

3.2 Reason-IAD

In this paper, we propose Reason-IAD, a knowledge-guided dynamic latent reasoning framework for explainable industrial anomaly detection. As illustrated in Figure 2, Reason-IAD comprises two principal components: (1) a Retrieval-Augmented Knowledge Integration (RAKI) module that retrieves category-specific textual descriptions and incorporates them into the model prompt, enabling context-aware reasoning over domain-specific defects; and (2) an Entropy-Driven Latent Reasoning mechanism (EDLR) that performs iterative exploration in a compact latent space using optimizable latent think tokens. Furthermore, inspired by human cognitive behavior in revisiting visual evidence under uncertainty, we introduce a dynamic visual injection strategy that selectively incorporates the most informative image patches into the latent sequence, guiding the reasoning process toward regions that are critical for anomaly detection.

3.3 Retrieval-Augmented Knowledge Integration

Existing multimodal large language models are predominantly pretrained on general-domain data, which often limits their ability to understand industrial anomalies. To improve both reasoning accuracy and explainability in industrial anomaly detection tasks, we propose a retrieval-augmented mechanism that incorporates category-specific knowledge to enrich the model input prompt.

Specifically, given a query image I and its associated question Q, we construct a category-oriented knowledge repository $\mathcal{K} = \{(c_i, \text{desc}_i)\}_{i=1}^{N}$, where c_i denotes the i-th category label (e.g., `cable`, `breakfast box`), and desc_i provides a detailed textual description of the category, encompassing both normal characteristics and defect patterns. This repository serves as an external knowledge source for contextual augmentation. To identify the most relevant category information, the query image is first encoded into a visual embedding $\mathbf{v} \in \mathbb{R}^d$ using a visual encoder f_v, while each category

label c_i is encoded into a textual embedding $\mathbf{k}_i \in \mathbb{R}^d$ using a text encoder f_t, projecting both modalities into a shared semantic space:

$$\mathbf{v} = f_v(I), \quad \mathbf{k}_i = f_t(c_i). \tag{1}$$

We then evaluate the semantic alignment between the query image and each category label by computing their cosine similarity:

$$s_i = \frac{\mathbf{v}^\top \mathbf{k}_i}{\|\mathbf{v}\| \, \|\mathbf{k}_i\|}, \quad i = 1, \ldots, N, \tag{2}$$

where s_i denotes the similarity score between the query image and the i-th category. According to the similarity scores $\{s_i\}$, we retrieve the top-k categories with the highest scores. The corresponding category descriptions are then incorporated into the prompt, forming an enriched contextual representation that provides category-aware guidance for reasoning and explanation generation.

3.4 Entropy-Driven Latent Reasoning

Industrial anomaly detection often demands sophisticated reasoning over subtle visual patterns and contextual knowledge. Traditional approaches, such as Chain-of-Thought (CoT) prompting, rely on generating explicit textual reasoning steps to guide model predictions, but they typically entail lengthy intermediate reasoning processes. Inspired by the intertwined nature of perception and reasoning in human cognition, we propose Reason-IAD, an entropy-driven latent reasoning framework. Within this framework, the model dynamically integrates visual evidence from the query image with category-specific knowledge from an external repository in the latent space, producing a compact and interpretable reasoning representation that enables robust industrial anomaly detection.

Latent Think Tokens. To facilitate dynamic exploration in the latent space, we introduce a set of latent think tokens, denoted as $\mathcal{Z} = \{z_1, z_2, \ldots, z_m\}$, which are initialized as optimizable vector embeddings and inserted into the original multimodal input sequence. During each iteration of latent reasoning, we incorporate controlled stochasticity to expand the search space while preserving the stability of the representations. Specifically, we adopt additive Gaussian noise as a local perturbation strategy. Given the latent think tokens $\mathcal{Z}^{(n)}$ at iteration n, the perturbed state is defined as:

$$\mathcal{Z}'^{(n)} = \mathcal{Z}^{(n)} + \xi^{(n)}, \quad \xi^{(n)} \sim \mathcal{N}(0, \sigma^2 I), \tag{3}$$

where σ^2 is the variance of the Gaussian noise, which controls the perturbation magnitude, and $\xi^{(n)}$ represents the Gaussian noise sampled at iteration n to encourage diverse latent trajectories. This strategy introduces controlled exploration in the latent space while preserving the semantic consistency of the original representations, facilitating more effective and robust latent reasoning.

Entropy-Guided Reward Optimization. In multimodal anomaly detection, the objective is to identify and localize anomalous regions given a query image along with its corresponding textual prompt. Let $Q = \{q_1, \ldots, q_l\}$ denote the sequence of input prompt tokens, and $I = \{i_1, \ldots, i_t\}$ represent the set of visual embeddings extracted from the input image. A MLLM parameterized by θ integrates these multimodal inputs to generate a latent reasoning sequence $\mathcal{G} = \{g_1, \ldots, g_N\}$ that supports anomaly detection. The generation process can be formally described using the following autoregressive factorization:

$$\pi_\theta(\mathcal{G}|Q, I) = \prod_{n=1}^{N} \pi_\theta(g_n|\mathcal{G}_{<n}, Q, I), \tag{4}$$

where $\mathcal{G}_{<n}$ denotes the sequence of tokens generated prior to step n. Accurate reasoning in industrial anomaly detection requires not only reliable predictions but also well-calibrated uncertainty over the model's internal reasoning states. Existing latent reasoning frameworks typically lack explicit mechanisms to regulate the uncertainty of intermediate representations, potentially resulting in ambiguous or unstable predictions. To address this limitation, we propose an entropy-guided reward mechanism that leverages predictive uncertainty as an intrinsic feedback signal for optimizing latent think tokens. Specifically, for each query instance, we quantify the model's predictive uncertainty by computing the entropy of its probability distribution as:

$$\mathcal{H}(P_i^{(n)}) = -\sum_{j=1}^{C} P_i^{(n)}(j) \log P_i^{(n)}(j), \tag{5}$$

Table 1: Performance comparison of commercial and open-source MLLMs on the MMAD benchmark under the standard one-shot setting. The best results are highlighted in bold, and all metrics are reported as accuracy (%).

Type	Model	Scale	Anomaly	Defect			Object		Average
			Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Random	Random Chance	-	50.00	25.00	25.00	25.00	25.00	25.00	25.00	28.57
Human	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
Commercial	Claude-3.5-sonnet	-	60.14	60.14	48.81	67.13	79.11	85.19	79.83	68.62
	Gemini-2.5-pro	-	83.07	**73.86**	67.20	**79.97**	86.27	94.88	83.08	**81.19**
	Gemini-2.5-flash	-	**83.38**	69.88	63.30	76.41	81.57	94.04	82.00	78.65
	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
Open Source	Qwen3-VL	2B	70.29	50.42	49.35	64.69	78.09	91.45	80.43	69.24
	Qwen2.5-VL	3B	63.03	47.47	54.87	66.66	80.57	86.85	82.84	68.90
	AnomalyR1	3B	60.62	63.56	**70.14**	**80.47**	85.28	92.48	86.15	76.96
	Qwen3-VL	4B	74.02	59.25	62.74	70.09	79.22	92.47	83.59	74.48
	AnomalyGPT	7B	65.57	27.49	27.97	36.86	32.11	29.84	35.82	36.52
	Qwen2.5-VL	7B	71.79	55.23	68.97	64.54	78.90	92.55	84.66	73.81
	LLaVA-1.5	7B	51.33	37.04	36.62	50.60	69.79	68.29	69.53	54.74
	LLaVA-NEXT	7B	57.64	33.79	47.72	51.84	67.93	81.39	74.91	59.32
	LLaVA-OneVision	7B	51.77	46.13	41.85	62.19	69.73	90.31	80.93	63.27
	InternVL2	8B	59.97	43.85	47.91	57.60	78.10	74.18	80.37	63.14
	Qwen3-VL	8B	73.08	62.33	60.34	69.15	80.09	92.53	84.01	74.51
	Qwen3-VL-thinking	8B	52.31	41.50	40.37	45.67	36.02	64.34	53.30	47.64
	GLM-4.1V-Thinking	9B	72.31	72.01	66.47	79.98	84.61	93.79	83.26	78.92
	LLaVA-1.5	13B	49.96	38.78	46.17	58.17	73.09	73.62	70.98	58.68
	Kimi-VL	16B	66.57	71.72	61.88	79.78	82.93	92.18	**86.51**	77.37
	LLaVA-NeXT	34B	57.92	48.79	52.87	71.34	80.28	81.12	77.80	67.16
	InternVL2	76B	68.25	54.22	56.66	66.30	80.47	86.40	82.92	70.75
Reason-IAD	Qwen3-VL	2B	61.90	70.20	54.63	71.51	79.92	94.59	81.61	73.48
	Qwen3-VL	4B	75.05	73.72	66.48	76.83	82.33	94.90	84.41	79.10
	Qwen2.5-VL	7B	70.74	67.97	60.82	72.78	82.62	**97.17**	84.99	76.73
	Qwen3-VL	8B	**75.24**	**75.07**	62.78	76.38	**85.58**	96.31	84.62	**79.43**

where $P_i^{(n)}$ denotes the predicted probability distribution over answers by the i-th latent think token at iteration n. Lower entropy indicates that the model has reached a more confident and deterministic reasoning state, while higher entropy reflects higher epistemic uncertainty arising from unfamiliar or conflicting features. Based on this measure, the global reward at iteration n is defined as the complement of the mean entropy across all latent think tokens:

$$\mathcal{R}(\mathcal{Z}^{(n)}) = 1 - \frac{1}{m}\sum_{i=1}^{m}\mathcal{H}(P_i^{(n)}), \tag{6}$$

where higher reward values incentivize the model to reduce uncertainty in its latent reasoning process, thereby encouraging more stable predictions. To iteratively optimize the latent think tokens toward higher confidence (i.e., lower entropy), we adopt a REINFORCE-based policy gradient method [61]. Specifically, given the stochastic perturbations of latent tokens defined in Eq. (3), the update rule at iteration n is formulated as:

$$\mathcal{Z}'^{(n)} \leftarrow \mathcal{Z}^{(n)} + \eta\nabla_{\mathcal{Z}^{(n)}}J(\mathcal{Z}^{(n)}), \tag{7}$$

where η represents the learning rate, and $J(\cdot)$ denotes the expected reward objective. Following the Policy Gradient Theorem and considering the Gaussian perturbations, the gradient of the objective can be expressed as:

$$\nabla_{\mathcal{Z}}\mathcal{J}(\mathcal{Z}) = \mathbb{E}_{\mathcal{Z}'\sim\pi(\cdot|\mathcal{Z})}\left[\mathcal{R}(\mathcal{Z}')\nabla_{\mathcal{Z}}\log\pi(\mathcal{Z}'|\mathcal{Z})\right] = \mathbb{E}\left[\mathcal{R}(\mathcal{Z}')\frac{\xi}{\sigma^2}\right]. \tag{8}$$

This optimization process encourages the latent tokens to reduce predictive uncertainty, thereby improving the overall reasoning accuracy of the model.

Dynamic Visual Injection. Inspired by human cognition, in which uncertain situations often prompt revisiting visual information, we introduce a dynamic visual injection mechanism within the latent reasoning process. This mechanism selectively incorporates the most informative image patches into the latent sequence alongside the latent think tokens, guided by the reward signal at each iteration.

Specifically, we leverage the attention weights of the latent think tokens to identify t highly relevant image patches. At each iteration, the model resamples t candidate patches $V_{cand} = \{v_1, \ldots, v_t\}$ according to the attention distribution computed from the current latent sequence, and evaluates their contribution by injecting them into the latent sequence. The optimal patch set V_{best} is then updated

Table 2: Performance comparison of open-source MLLMs on the MMAD benchmark under the standard zero-shot setting. The best results are highlighted in bold, and all metrics are reported as accuracy (%).

Type	Model	Scale	Anomaly	Defect			Object		Average
			Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Open Source	Qwen3-VL	2B	63.27	47.71	52.48	63.48	77.82	92.16	80.03	68.14
	Qwen2.5-VL	3B	60.63	47.03	54.44	63.51	79.25	87.48	82.02	67.77
	Qwen3-VL	4B	65.83	53.97	58.16	66.35	77.82	91.98	83.38	71.07
	Qwen3-VL thinking	4B	64.21	39.61	42.98	48.96	38.32	78.18	64.18	53.78
	Qwen2.5-VL	7B	60.35	49.89	57.51	60.41	75.58	93.39	84.80	68.85
	LLaVA-1.5	7B	50.81	36.82	35.94	50.96	70.45	67.33	69.12	54.49
	InternVL3	8B	64.67	50.91	60.06	65.49	77.16	82.01	83.58	69.13
	Qwen3-VL	8B	67.10	57.98	58.86	66.89	77.65	92.94	83.82	72.18
	Qwen3-VL thinking	8B	63.07	45.81	49.56	53.52	40.63	82.51	66.25	57.34
	LLaVA-NeXT	34B	60.31	51.40	55.47	71.62	80.49	81.45	75.08	67.97
	Qwen2.5-VL	72B	66.22	57.99	**62.95**	72.21	81.01	93.92	**86.51**	74.40
	InternVL2	76B	64.30	51.19	54.20	63.46	79.92	89.34	83.48	69.41
Reason-IAD	Qwen3-VL	2B	54.66	68.95	55.28	70.64	79.68	95.28	80.85	72.19
	Qwen3-VL	4B	66.58	70.62	62.82	73.28	**80.88**	93.80	84.88	76.12
	Qwen2.5-VL	7B	59.23	61.02	55.39	66.57	78.37	**97.36**	85.50	71.92
	Qwen3-VL	8B	**67.73**	**73.12**	59.74	**74.69**	80.65	96.78	84.24	**76.71**

based on the current reward $\mathcal{R}$ as follows:

$$V_{best} = \begin{cases} V_{cand}, & \text{if } \mathcal{R} > \mathcal{R}_{best}, \\ V_{best}, & \text{otherwise.} \end{cases} \tag{9}$$

If the current reward $\mathcal{R}$ exceeds $\mathcal{R}_{best}$, the candidate patches are considered to provide more informative visual evidence (e.g., clearer defect patterns or more discriminative local details), and the best patch set V_{best} is updated accordingly. Otherwise, the previous best patches are retained. As shown in Algorithm 1 in the Appendix, through this iterative, reward-guided refinement, the selected visual patches progressively concentrate on regions most relevant to the latent think tokens, thereby guiding the latent reasoning process toward more accurate and efficient anomaly detection.

4 Experiments

4.1 Experiment Setup

Compared Baselines. To comprehensively evaluate the reasoning capabilities of the proposed Reason-IAD, we conduct extensive experiments comparing it against a diverse set of representative models, including both commercial and open-source multimodal large language models (MLLMs) as well as anomaly detection–specific models. The commercial baselines include Claude-3.5-Sonnet [62], Gemini-2.5-Pro [63], GPT-4o [64], and GPT-5-Mini [65]. The open-source baselines include Qwen2.5-VL [66], Qwen3-VL [67], LLaVA-1.5 [68], LLaVA-NeXT [69], LLaVA-OneVision [70], and InternVL2 [71]. In addition, we evaluate Reason-IAD against anomaly detection–specific models, including AnomalyGPT [21] and Anomaly-R1 [26].

Benchmarks. We evaluate Reason-IAD on the MMAD benchmark [1], which assesses performance across seven subtasks: anomaly discrimination, defect classification, defect localization, defect description, defect analysis, object classification, and object analysis. MMAD comprises four datasets: MVTec-AD [72], VisA [73], MVTec-LOCO [74], and GoodsAD [75].

Implementation Details. In this study, we adopt Qwen2.5-VL-7B and Qwen3-VL-2B/4B/8B as the base models. For retrieval, we employ CLIP with a ViT-L/14-336 backbone as the image and text encoders, with top-k set to 2. The number of latent thinking tokens $\mathcal{Z}$ is set to 4, and the default number of reasoning iterations is 10. The learning rate is fixed at 1×10^{-3}. To ensure stable exploration in the latent space, we set the perturbation magnitude to 10%. By default, we employ a one-shot setting with a randomly sampled normal image from the same domain as the reference template. For comparison, we also evaluate a zero-shot setting. All experiments are conducted on four NVIDIA A100 GPUs (80GB). More details are provided in the Appendix.

4.2 Main Results

Table 1 presents a comprehensive comparison of Reason-IAD with state-of-the-art methods under the one-shot setting. Overall, Reason-IAD demonstrates strong performance across all seven tasks. On

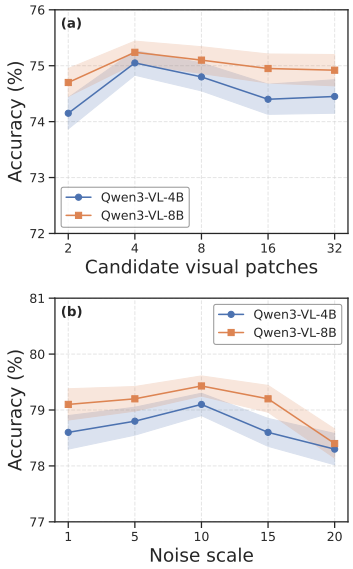


Figure 3: (a) Effect of the number of injected candidate visual patches. (b) Impact of noise magnitude.

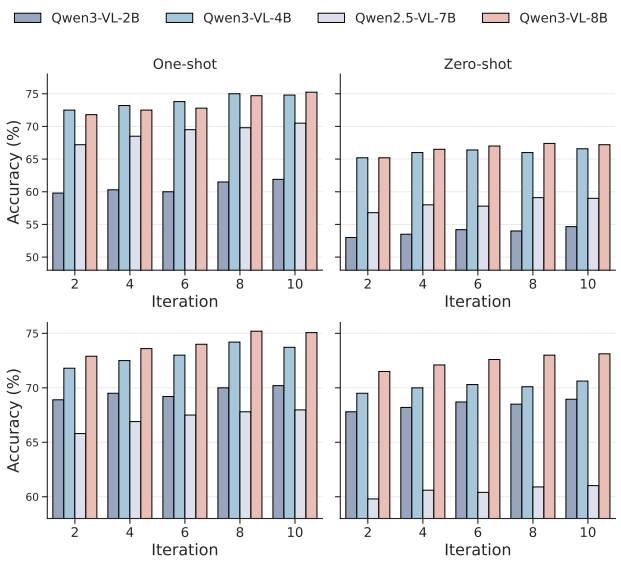


Figure 4: Impact of iteration count on anomaly discrimination (top row) and anomaly classification performance (bottom row) in one-shot and zero-shot settings.

Qwen3-VL-8B, it achieves an average accuracy of 79.43%, improving 4.92% over the base model, and even surpasses the commercial model GPT-4o (74.92%). Although AnomalyGPT is designed for anomaly detection, it underperforms on classification and localization, which may be attributed to catastrophic forgetting when trained on expert outputs. Another domain-specific method, AnomalyR1, incorporates GRPO to enhance anomaly understanding, yet our training-free method still outperforms it by 2.47% in average accuracy, demonstrating stronger reasoning capability. Notably, Reason-IAD surpasses standard human annotators on defect analysis and object classification tasks. Furthermore, while Qwen3-VL-thinking relies on explicit reasoning chains that generate lengthy intermediate outputs, our implicit reasoning achieves a 31.79% gain in performance, highlighting the effectiveness and efficiency of Reason-IAD.

Table 2 reports the results under the zero-shot setting, where no normal reference images are provided. Reason-IAD consistently improves the average performance of the evaluated base models, with Qwen3-VL-4B and Qwen3-VL-8B achieving average accuracy gains of 5.05% and 4.53%, respectively. Notably, our 8B model outperforms InternVL2-76B by 7.30%, further demonstrating the effectiveness of Reason-IAD in anomaly understanding.

4.3 Model Analysis

Ablation Study. To evaluate the contributions of key components in the proposed Reason-IAD, we conduct ablation studies on four public datasets, as summarized in Table 3. Incorporating knowledge guidance consistently improves detection accuracy across all datasets. Specifically, on MVTec, accuracy increases from 85.75% to 88.49%, with an average gain of 3.22% across the four datasets, highlighting the critical role of prior knowledge. Latent reasoning further enhances performance by 1.54%, improving the model's inference capability. Moreover, integrating visual information provides an additional average improvement of 0.32%, indicating that the model effectively leverages visual evidence during reasoning. When all components are combined, Reason-IAD achieves the best overall performance, underscoring the effectiveness of the proposed modules.

Impact of Visual Patch Number. As shown in Figure 3(a), we study the effect of varying the number of injected visual patches on anomaly detection performance. As the number of patches increases, detection accuracy improves steadily; however, further increasing the number leads to performance

Table 3: Ablation study of different components with Qwen3-VL-8B under the one-shot setting on four public anomaly detection datasets, all metrics are reported as accuracy (%).

Module			Datasets			
RAKI	EDLR	V_{patch}	MVTec	VisA	MVTec-LOCO	GoodsAD
✗	✗	✗	85.75	75.44	67.45	69.38
✓	✗	✗	88.49	77.83	70.76	73.83
✓	✓	✗	90.03	79.13	71.95	75.29
✓	✓	✓	**90.36**	**79.46**	**72.29**	**75.61**

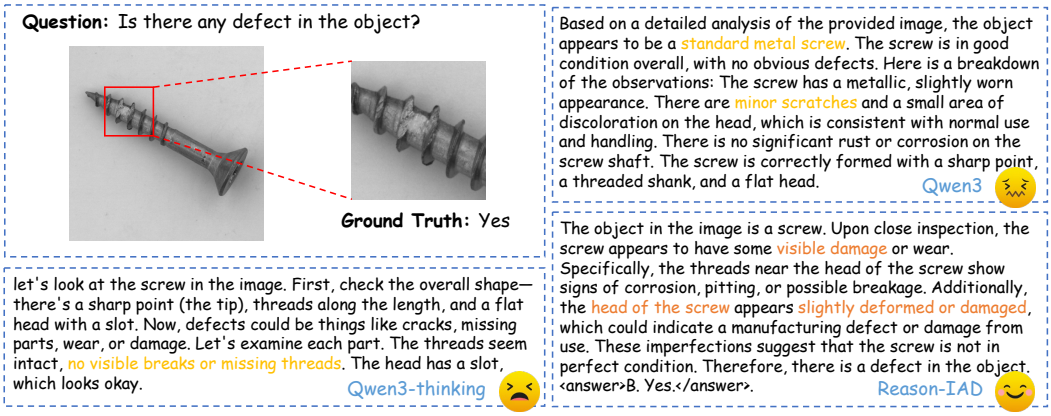


Figure 5: Comparison of model outputs for anomaly detection.

degradation. This suggests that a limited set of fine-grained visual cues is sufficient to effectively guide latent reasoning, whereas excessive patches introduce redundancy that hinders optimization.

Impact of Noise Scale. We further investigate how the perturbation magnitude σ governs the dynamics of latent optimization. As illustrated in Figure 3(b), increasing the initial noise scale enhances exploration, enabling the model to traverse a broader range of latent trajectories and discover higher-confidence reasoning paths. However, when σ becomes excessively large, the perturbations destabilize the optimization process, resulting in performance degradation. This suggests that latent reasoning benefits from a well-calibrated level of perturbation.

Impact of Iteration Number. To evaluate the effect of iteration count on model performance, we conduct experiments on four base models for anomaly detection, as shown in Figure 4. The results indicate that increasing the number of iterations consistently improves anomaly reasoning, demonstrating that iterative optimization effectively enhances latent reasoning. Although minor fluctuations occur in intermediate steps, the models maintain high accuracy and continue to benefit from additional iterations, highlighting their capacity to leverage iterative optimization. Iterations are fixed at 10 to balance performance and efficiency.

Sensitivity to the Number of Latent Tokens. To evaluate how the number of latent tokens influences the capacity of the latent reasoning space, we perform an ablation study by varying the token count, with results shown in Table 4. An insufficient number of tokens limits the representation of intermediate reasoning states, whereas an excessive number enlarges the optimization space and slightly compromises stability. Performance remains relatively stable across configurations, with four latent tokens yielding the best results. This indicates that a moderate number of latent tokens strikes a favorable balance between reasoning capacity and optimization stability.

Table 4: Ablation study on the number of latent tokens.

Latent Tokens	Anomaly Discrimination	Defect Localization	Object Classification
2	73.91	61.12	95.64
3	74.68	62.01	96.02
4	**75.24**	**62.78**	**96.31**
5	74.97	62.43	96.18
6	74.63	62.05	95.91

Table 5: Comparison of retrieval strategies.

Strategy	MVTec	VisA	Avg.
w/o RAKI	87.03	76.88	81.96
Top-1	89.91	78.74	84.33
Top-2	**90.36**	**79.46**	**84.91**
Top-3	89.72	78.96	84.34
Random-2	86.41	75.73	81.07

Table 6: Comparison of reward formulations for latent reasoning optimization.

Reward	MVTec	VisA	Avg.
w/o latent optimization	88.49	77.83	83.16
Random reward	88.12	77.46	82.79
Max-probability reward	89.71	78.83	84.27
Entropy-based reward	**90.36**	**79.46**	**84.91**

Qualitative Analysis. Figure 5 presents a qualitative comparison of different MLLMs on an industrial anomaly detection case. Both Qwen3-VL and Qwen3-VL-thinking fail to correctly identify the defective screw. Specifically, Qwen3-VL-thinking performs step-by-step reasoning but overlooks subtle defects in the screw structure, while Qwen3-VL interprets visible scratches and discoloration as normal wear, leading to an incorrect prediction. In contrast, Reason-IAD correctly identifies the anomaly by recognizing potential damage around the screw head and threads, providing a more accurate and detailed explanation. These results suggest that explicit reasoning alone may be insufficient for fine-grained anomaly recognition, whereas integrating domain-specific knowledge and relevant visual evidence can enhance anomaly understanding and reasoning reliability.

Sensitivity to Knowledge Retrieval Strategies. We evaluate the impact of different retrieval strategies on anomaly reasoning performance. As shown in Table 5, Top-2 achieves the highest average accuracy of 84.91%, outperforming Top-1 by 0.58 percentage points. This suggests that an additional candidate can improve robustness to category-ranking errors. However, increasing the number of retrieved categories to three slightly degrades performance, indicating potential interference from irrelevant knowledge. Moreover, Random-2 underperforms even the setting without RAKI, highlighting the importance of retrieval relevance.

Comparison of Reward Strategies. We investigate the effectiveness of different reward formulations for latent reasoning. As shown in Table 6, random rewards fail to improve performance, whereas both maximum-probability and entropy-based rewards yield consistent gains. In particular, the entropy-based reward achieves the highest average accuracy of 84.91%, surpassing the maximum-probability reward by 0.64 percentage points. Unlike maximum-probability rewards that focus solely on the most likely prediction, entropy-based rewards capture uncertainty across the entire candidate-answer distribution, providing a more comprehensive signal for latent exploration.

5 Conclusion

In this work, we introduce Reason-IAD, a knowledge-guided dynamic latent reasoning framework for explainable industrial anomaly detection. By integrating a retrieval-augmented knowledge module, an entropy-driven latent reasoning mechanism, and a dynamic visual injection strategy, Reason-IAD facilitates context-aware, iterative reasoning over fine-grained industrial defects while effectively leveraging critical visual evidence. Extensive experiments demonstrate that Reason-IAD consistently outperforms state-of-the-art methods, achieving high detection accuracy and interpretable reasoning. In summary, Reason-IAD provides a promising paradigm for training-free, expert-level anomaly detection. Future directions include adaptive knowledge retrieval and advanced latent optimization strategies to further enhance the applicability of MLLMs across diverse industrial scenarios.

Acknowledgments and Disclosure of Funding

Funding. This work was supported in part by the National Natural Science Foundation of China (No. 62301621, No. 62672529, No. 62625602, 62411540034), Shenzhen Science and Technology Program (No. 20231121172359002, 2023A008, NO.SYSRD20250529113401002), Shenzhen General Research Project (No. JCYJ20241202125904007), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011398), and Guangdong Major Project of Basic Research (2023B0303000010).

Competing Interests. The authors declare that they have no competing interests related to this work.

References

[1] 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. *arXiv preprint arXiv:2410.09453*, 2024.

[2] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15204–15213, 2025.

[3] Yifan Fang, Yifei Fang, Ruizhe Chen, Haote Xu, Xinghao Ding, and Yue Huang. Demeaned sparse: Efficient anomaly detection by residual estimate. In *Forty-second International Conference on Machine Learning*, 2025.

[4] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025.

[5] Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. Cnc: Cross-modal normality constraint for unsupervised multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7943–7951, 2025.

[6] Peng Chen, Fangjun Huang, and Chao Huang. Dyc-clip: Dynamic context-aware multi-modal prompt learning for zero-shot anomaly detection. *Pattern Recognition*, page 113215, 2026.

[7] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[8] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025.

[9] Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. Decad: Decoupling anomalies in latent space for multi-class unsupervised anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21568–21577. IEEE, 2025.

[10] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, pages 301–317. Springer, 2024.

[11] Chao Huang, Peng Chen, Jie Wen, Wei Wang, Li Shen, Wenqi Ren, Xiaochun Cao, and Chengliang Liu. Noise-induced cross-modal information interaction and dual-prompt learning for medical image segmentation. *IEEE Transactions on Image Processing*, 2026.

[12] Weichao Cai, Weiliang Huang, Yunkang Cao, Chao Huang, Bob Zhang, Fei Yuan, Jie Wen, and Weiming Shen. Hepe++: Towards vlm-based hybrid explainable prompt enhancement for zero-shot industrial anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[13] Feng Yan, Xiaoheng Jiang, Yang Lu, Jiale Cao, Dong Chen, and Mingliang Xu. Wavelet and prototype augmented query-based transformer for pixel-level surface defect detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23860–23869, 2025.

[14] Weichao Cai, Weiliang Huang, Yunkang Cao, Chao Huang, Fei Yuan, Bob Zhang, and Jie Wen. Towards vlm-based hybrid explainable prompt enhancement for zero-shot industrial anomaly detection. In *IJCAI*, pages 711–719, 2025.

[15] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

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

[17] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16838–16848, 2024.

[18] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024.

[19] Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. Icc: Intra-cluster contraction for pedestrian anomaly detection under one-class classification setting. *IEEE Transactions on Artificial Intelligence*, 2025.

[20] Chao Huang, Weichao Cai, Qiuping Jiang, and Zhihua Wang. Multimodal representation distribution learning for medical image segmentation. In *IJCAI*, pages 4156–4164, 2024.

[21] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

[22] Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. Iad-r1: Reinforcing consistent reasoning in industrial anomaly detection. *arXiv preprint arXiv:2508.09178*, 2025.

[23] Jinyi Xu, Peng Chen, Yunkang Cao, Chengliang Liu, Xinghui Dong, and Chao Huang. Osagen: Object-aware mask priors and multistage decoupled diffusion for industrial anomaly generation. *arXiv preprint arXiv:2607.29533*, 2026.

[24] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[25] Wenqiang Wang, Peng Chen, Yan Xiao, Yangshijie Zhang, Xiaoyue Lu, Jianjie Huang, and Xiaochun Cao. Task-related in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43378–43400, 2026.

[26] Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection. *arXiv preprint arXiv:2504.11914*, 2025.

[27] Jingyi Liao, Yongyi Su, Rong-Cheng Tu, Zhao Jin, Wenhao Sun, Yiting Li, Dacheng Tao, Xun Xu, and Xulei Yang. Ad-fm: Multimodal llms for anomaly detection via multi-stage reasoning and fine-grained reward optimization. *arXiv preprint arXiv:2508.04175*, 2025.

[28] Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning. *arXiv preprint arXiv:2510.23925*, 2025.

[29] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

[30] Wenqiang Wang, Wen Yujia, Yan Xiao, Zhifeng Chen, Yangshijie Zhang, Peng Chen, Mingbo Yang, and Xiaochun Cao. Incomplete in-context learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41629–41650, 2026.

[31] Yuyang Cai, Peng Chen, Feng Qian, Gao Chen, and Jingwen Yan. Botwavenet: semantic segmentation network for remote sensing images based on wavelet transform and attention mechanisms. *Journal of Electronic Imaging*, 34(2):023053–023053, 2025.

[32] Xiaolei Wang, Tianhong Dai, Huihui Bai, Yao Zhao, and Jimin Xiao. Unifying reconstruction and density estimation via invertible contraction mapping in one-class classification. *Advances in Neural Information Processing Systems*, 38:135196–135222, 2026.

[33] Weichao Cai, Weiliang Huang, Lin Tian, Chao Huang, and Jingwen Yan. Multiscale global attention for abnormal geological hazard segmentation. *IEEE Sensors Journal*, 24(10):16961–16971, 2024.

[34] Peng Chen, Kaige Li, Wei Wang, Mingbo Yang, Wenqiang Wang, Li Shen, Fangjun Huang, and Chao Huang. Vfad: Variational semantic prompting meets frequency-adaptive representation learning for zero-shot anomaly detection. *arXiv preprint arXiv:2607.29370*, 2026.

[35] Yaoxuan Feng, Wenchao Chen, Yuxin Li, Bo Chen, Yubiao Wang, Zixuan Zhao, Hongwei Liu, and Mingyuan Zhou. Omiad: One-step adaptive masked diffusion model for multi-class anomaly detection via adversarial distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 16604–16633, 2025.

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

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

[38] Qiao Zhang, Mingwen Shao, Xinyuan Chen, Xiang Lv, and Kai Xu. Wave-mambaad: Wavelet-driven state space model for multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20868–20877, 2025.

[39] 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*, pages 20370–20382, 2025.

[40] Peng Chen and Chao Huang. Wmoe-clip: Wavelet-enhanced mixture-of-experts prompt learning for zero-shot anomaly detection. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 22257–22261. IEEE, 2026.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[42] Mingbo Yang, Wenqiang Wang, Zhaolu Kang, Peng Chen, Yannan Chen, Sunshang Wang, and Yan Xiao. Beyond surface imitation: Contrastive modeling for reasoning path alignment in multimodal in-context learning. *arXiv preprint arXiv:2609.10177*, 2026.

[43] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[44] Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. Search and refine during think: Facilitating knowledge refinement for improved retrieval-augmented reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[45] Peng Chen, Wenxuan He, Feng Qian, Guangyao Shi, and Jingwen Yan. A synergistic cnn-transformer network with pooling attention fusion for hyperspectral image classification. *Digital Signal Processing*, 160:105070, 2025.

[46] Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025.

[47] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pages 279–290. Pmlr, 2020.

[48] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023.

[49] Yuanze Li, Shihao Yuan, Haolin Wang, Qizhang Li, Ming Liu, Chen Xu, Guangming Shi, and Wangmeng Zuo. Triad: Empowering lmm-based anomaly detection with expert-guided region-of-interest tokenizer and manufacturing process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21917–21926, 2025.

[50] Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *arXiv preprint arXiv:2505.16552*, 2025.

[51] Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, et al. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space. *arXiv preprint arXiv:2505.13308*, 2025.

[52] Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*, 2025.

[53] Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*, 2025.

[54] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

[55] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

[56] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[57] Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025.

[58] Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025.

[59] Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*, 2025.

[60] Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

[61] Alessandro Montenegro, Marco Mussi, Matteo Papini, and Alberto Maria Metelli. Last-iterate global convergence of policy gradients for constrained reinforcement learning. *Advances in Neural Information Processing Systems*, 37:126363–126416, 2024.

[62] Anthropic. Claude 3 model family: Opus, sonnet, haiku. `https://www.anthropic.com`, 2024. Model card for Claude-3 series.

[63] 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.

[64] 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.

[65] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[66] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

[68] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[69] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024. *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*, 1(8), 2024.

[70] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[71] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

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

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

[74] 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.

[75] 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.

Appendix

This appendix provides supplementary material for the main paper. We present detailed experimental settings, datasets, prompt design, and additional experimental results, followed by further case study analyses. Finally, we discuss the potential impact and limitations of the proposed method.

The supplementary material is organized as follows:

- Sec. A provides additional details about the evaluation framework, including experimental setup, datasets, prompt design, and an overview of the proposed method.
- Sec. B presents additional experimental results.
- Sec. C provides case studies to further illustrate the effectiveness of the proposed method.
- Sec. D discusses the broader impact and limitations of the proposed approach.

A More Details about Evaluation

A.1 Experiment Setup.

To comprehensively evaluate the reasoning capabilities of the proposed Reason-IAD, we conduct extensive experiments comparing it against a diverse set of representative models. These include both commercial and open-source multimodal large language models (MLLMs), as well as models specifically designed for anomaly detection. The commercial baselines include Claude-3.5-Sonnet [62], Gemini-2.5-Pro [63], GPT-4o [64], and GPT-5-Mini [65]. The open-source baselines include Qwen2.5-VL [66], Qwen3-VL [67], LLaVA-1.5 [68], LLaVA-NeXT [69], LLaVA-OneVision [70], and InternVL2 [71]. In addition, we benchmark Reason-IAD against anomaly detection–specific models, such as AnomalyGPT [21] and Anomaly-R1 [26], to evaluate its performance relative to domain-specialized approaches.

For our experiments, we use Qwen2.5-VL-7B and Qwen3-VL-2B/4B/8B as base models. In the retrieval stage, we employ the CLIP model with a ViT-L/14-336 backbone as the image and text encoder, with top-k set to 2. The number of latent thinking tokens $\mathcal{Z}$ is set to 4, and the default number of reasoning iterations is 10. The learning rate is fixed at 1×10^{-3}, and the perturbation magnitude is set to 10% to ensure stable exploration in the latent space. By default, we adopt a one-shot setting, where a normal image from the same domain is randomly sampled as a reference template to help the model capture the normal appearance of objects. For comparison, we also evaluate the models under a zero-shot setting, which tests their ability to perform pairwise comparisons between industrial images without reference guidance. This evaluation protocol provides a comprehensive assessment of the models' reasoning capabilities in both reference-based and reference-free scenarios. All experiments are conducted on four NVIDIA A100 GPUs (80GB).

A.2 Datasets.

Table 7: Statistics on the composition and quantity of test data.

Image Source	Sampled Images	Generated Questions	Object Categories	Defect Categories
MVTec-AD [72]	1691	8338	15	73
MVTec-LOCO-AD [74]	1566	7624	5	89
VisA [73]	2141	10622	12	67
GoodsAD [75]	2968	13088	6	15
SUM	8366	39672	38	244

We evaluate Reason-IAD on multiple industrial anomaly detection (IAD) datasets with diverse characteristics, covering over 38 product categories and 244 defect types, as summarized in Table 7. Among these, MVTec AD is one of the most widely used benchmarks for IAD, encompassing a variety of object and texture categories; we further leverage fine-grained annotations from Defect Spectrum to provide more detailed defect descriptions. MVTec LOCO AD focuses on logical anomalies, thereby evaluating the model's ability to reason about high-level logical inconsistencies. The VisA dataset contains multiple instances with complex defect patterns, reflecting more challenging real-world IAD

Algorithm 1 Reason-IAD: Knowledge-Guided Dynamic Latent Reasoning

Input: Query image I, question Q, knowledge repository $\mathcal{K} = \{(c_i, \text{desc}_i)\}_{i=1}^{N}$, latent token number m, iteration number N_{iter}, top-k categories
Encode image: $\mathbf{v} \leftarrow f_v(I)$
for each category $c_i \in \mathcal{K}$ **do**
  Encode text: $\mathbf{k}_i \leftarrow f_t(c_i)$
  Compute similarity: $s_i \leftarrow \cos(\mathbf{v}, \mathbf{k}_i)$
end for
Retrieve top-k categories $\{c_{(1)}, \ldots, c_{(k)}\}$ by $\{s_i\}$ and Construct enriched prompt
Initialize latent think tokens $\mathcal{Z} = \{z_1, \ldots, z_m\}$
for $n = 1$ to N_{iter} **do**
  Sample noise $\xi^{(n)} \sim \mathcal{N}(0, \sigma^2 I)$
  $\mathcal{Z}'^{(n)} \leftarrow \mathcal{Z}^{(n)} + \eta \nabla_{\mathcal{Z}^{(n)}} J(\mathcal{Z}^{(n)})$
  Initialize best patch V_{best}
  for $j = 1, \ldots, m$ **do**
    Sample candidate patches V_{cand} according to attention distribution
    Compute entropy $\mathcal{H}^{(n)}$ and reward $\mathcal{R}^{(n)}$ using Eq. (5)–(6)
    if $\mathcal{R}^{(n)} > \mathcal{R}_{\text{best}}$ **then**
      $V_{\text{best}} \leftarrow V_{\text{cand}}$
      $\mathcal{R}_{\text{best}} \leftarrow \mathcal{R}^{(n)}$
    end if
  end for
end for
Return reasoning result

scenarios. GoodsAD primarily consists of industrial goods, and due to the diverse appearances of finished products across different brands, it significantly increases the number of object categories and visual variations.

In our experiments, we adopt the MMAD benchmark as the evaluation protocol, which comprises seven subtasks: (1) Anomaly Discrimination Detection, a binary classification task that determines whether a sample is defective, to evaluate the model's fundamental anomaly detection capability; (2) Defect Classification, which identifies the specific defect type, jointly assessing anomaly recognition and domain knowledge of industrial defect categories; (3) Defect Localization, requiring the model to locate defective regions, where standardized textual descriptions are used instead of explicit segmentation masks; (4) Defect Description, which characterizes defect attributes (e.g., size and color), simulating practical defect inspection scenarios; (5) Defect Analysis, analyzing the potential impact of defects on product quality to estimate their severity; (6) Object Classification, categorizing industrial products to facilitate anomaly detection based on an understanding of normal object characteristics; and (7) Object Analysis, querying the composition, position, and function of products to evaluate fine-grained semantic understanding.

A.3 Overview of the Proposed Method

In this work, we introduce Reason-IAD, a knowledge-guided dynamic latent reasoning framework for explainable industrial anomaly detection, as illustrated in Algorithm 1. By integrating a retrieval-augmented knowledge module, an entropy-driven latent reasoning mechanism, and a dynamic visual injection strategy, Reason-IAD enables context-aware iterative reasoning over fine-grained industrial defects while effectively leveraging critical visual evidence.

Extensive experiments demonstrate that Reason-IAD consistently outperforms state-of-the-art methods, achieving high detection accuracy and improved interpretability. Overall, Reason-IAD provides a promising paradigm for training-free, expert-level industrial anomaly detection.

A.4 Prompt Design

To guide the multimodal large language models (MLLMs) to perform domain-specific anomaly reasoning, we design a structured prompt that explicitly defines the model as an industrial inspection

Table 8: Performance comparison of different MLLMs in Anomaly Discrimination Tasks.

	Model	Scale	Accuracy	Recall	Precision
Human	Human (expert)	-	95.24	94.25	98.89
	Human (ordinary)	-	86.90	87.07	94.35
Commercial	claude-3.5-sonnet	-	60.14	30.87	76.75
	Gemini-1.5-flash	-	58.58	78.63	67.41
	Gemini-1.5-pro	-	68.63	45.47	86.84
	GPT-4o-mini	-	64.33	65.47	73.04
	GPT-4o	-	68.63	67.37	75.68
Open Source	Qwen3-VL	2B	70.29	66.83	79.78
	Qwen3-VL	4B	74.02	57.95	89.41
	AnomalyGPT	7B	65.57	82.11	74.45
	Qwen2.5-VL	7B	71.79	53.74	75.54
	Qwen-VL-Chat	7B	53.65	43.95	65.39
	LLaVA-1.5	7B	51.33	94.79	62.72
	Qwen3-VL	8B	73.08	75.58	79.76
	Cambrian-1*	8B	55.60	22.28	74.10
	SPHINX*	7B	53.13	6.42	99.74
	LLaVA-NEXT-Interleave	7B	57.64	16.58	90.83
	InternLM-XComposer2-VL	7B	55.85	17.94	75.87
	LLaVA-OneVision	7B	51.77	4.90	78.19
	MiniCPM-V2.6	8B	57.31	34.38	70.98
	InternVL2	8B	59.97	30.25	79.22
	LLaVA-1.5	13B	49.96	99.79	62.00
	LLaVA-NeXT	34B	57.92	46.27	69.98
	InternVL2	76B	68.25	55.81	83.52
Reason-IAD	Qwen3-VL	2B	61.90	92.23	69.19
	Qwen3-VL	4B	75.05	68.21	86.40
	Qwen2.5-VL	7B	70.74	55.50	75.95
	Qwen3-VL	8B	75.24	73.77	83.37

expert and incorporates category-level domain knowledge. Specifically, the prompt instructs the model to determine whether the queried image contains defects and to answer the associated questions based on visual evidence. To encourage reliable decision-making, the model is required to perform internal reasoning before producing the final prediction.

Moreover, we inject retrieved domain knowledge describing typical defect types and normal object characteristics into the prompt, enabling context-aware reasoning tailored to industrial scenarios. This design allows the model to leverage both visual observations and prior knowledge, facilitating accurate and interpretable anomaly detection. The prompt template is defined as follows:

> You are an expert industrial inspector responsible for analyzing product images. Your task is to determine whether the query image contains any defects and to answer the related questions. You should first perform the reasoning process internally and then provide the final answer. The following domain knowledge describes typical defect types and normal object characteristics: `\n {Domain_knowledge}`.

B Additional Experimental Results

Analysis of Anomaly Discrimination Performance. To comprehensively evaluate anomaly discrimination performance, we follow the standard binary anomaly detection protocol, treating anomalous samples as the positive class and normal samples as the negative class, and report accuracy, recall, and precision as evaluation metrics.

As shown in Table 8, analysis of recall and precision highlights the reasons for suboptimal performance in certain models. For example, both SPHINX and LLaVAOne-Vision frequently misclassify anomalies as normal, leading to many missed detections, while LLaVA-1.5 shows high recall but low precision, reflecting a substantial false positive rate. Human experts outperform all MLLMs, with

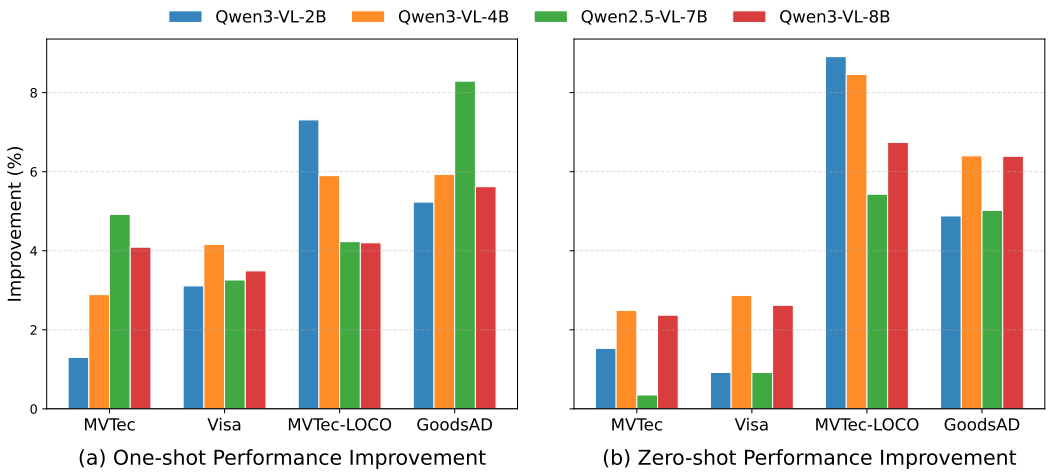


Figure 6: Performance gains of Reason-IAD over baseline models under the one-shot and zero-shot setting.

experts achieving over 94% across metrics and ordinary individuals over 87%. Notably, AnomalyGPT, specifically trained for anomaly detection, outperforms most models but still exhibits a significant false positive issue. Our proposed Reason-IAD effectively mitigates these limitations, enhancing detection performance by leveraging knowledge priors and a latent reasoning strategy. Moreover, the accuracy of all four models exceeds 60%.

Effectiveness of the Proposed Reasoning Strategy. As shown in Figure 6, consistent performance gains are observed under both one-shot and zero-shot settings. Compared with the baseline models, the proposed Reason strategy consistently achieves notable gains across all backbones.

In the one-shot setting, the largest improvement is observed on the GoodsAD dataset with Qwen2.5-VL-7B, reaching 8.29%, followed by Qwen3-VL-2B, which achieves a gain of 7.31% on MVTec-LOCO. In the zero-shot setting, Qwen3-VL-2B and Qwen3-VL-4B obtain improvements of 8.91% and 8.46% on the MVTec dataset, respectively. These results demonstrate that the proposed reasoning strategy significantly enhances anomaly detection performance across different model scales and evaluation settings, indicating strong robustness and generalization capability.

Effect of Retrieval-Augmented Knowledge Integration. Table 9 presents the impact of the proposed Retrieval-Augmented Knowledge Integration (RAKI) on different multimodal large language models (MLLMs). By incorporating externally retrieved knowledge into the reasoning pipeline, RAKI consistently enhances industrial anomaly understanding and reasoning performance.

For Qwen2.5-VL-7B, RAKI significantly improves several defect-oriented tasks, with defect classification increasing from 65.12% to 67.97%, description from 70.36% to 72.78%, and object classification from 95.24% to 97.17%. Similarly, Qwen3-VL-4B achieves substantial gains after introducing RAKI, where defect classification improves from 69.45% to 73.72%, while localization and description rise to 66.48% and 76.83%, respectively.

Notably, the smaller Qwen3-VL-4B benefits even more from RAKI than the larger Qwen2.5-VL-7B, suggesting that retrieval-augmented external knowledge is particularly effective for lightweight MLLMs with limited intrinsic industrial knowledge. Overall, the results demonstrate that RAKI effectively injects domain-specific knowledge into the reasoning process, enabling MLLMs to better understand industrial contexts and generate more accurate anomaly reasoning results.

C Case Study

As shown in Figures 7–13, we further present several qualitative case studies on representative industrial anomaly detection scenarios featuring diverse defect patterns and object categories. These examples are provided to visually demonstrate the effectiveness of the proposed Reason-IAD in

Table 9: Performance comparison of different MLLMs with and without RAKI.

Model	RAKI	Anomaly	Defect				Object	
		Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis
Qwen2.5-VL-7B	–	69.93	65.12	58.94	70.36	80.91	95.24	84.71
	✓	70.74	67.97	60.82	72.78	82.62	97.17	84.99
Qwen3-VL-4B	–	73.86	69.45	64.37	74.21	81.14	93.76	84.03
	✓	75.05	73.72	66.48	76.83	82.33	94.90	84.41

identifying anomalous regions, as well as its interpretability in explaining the underlying defect characteristics through structured reasoning.

In addition, we report several failure cases in Figures 14–15. In Figure 14, for the defect localization task, Reason-IAD correctly identifies the presence of a defect; however, since the defect region is relatively large and spans the lower and right portions of the image, the model incorrectly localizes it to the right region. In Figure 15, the model incorrectly classifies an image with a missing pushpin as normal due to the highly cluttered background and the presence of numerous small-scale pushpin targets. These failure cases indicate that complex spatial layouts and fine-grained small-object anomalies remain challenging for current reasoning-based anomaly detection models.

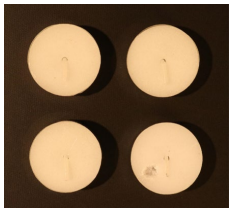
D Impact and Limitation

In this paper, we propose Reason-IAD, a training-free framework for explainable industrial anomaly detection that leverages knowledge-guided dynamic latent reasoning. We believe this work can contribute to industrial anomaly detection research by introducing a new paradigm that integrates external domain knowledge with iterative latent reasoning, enabling more interpretable and controllable anomaly analysis without additional training cost.

Beyond improved detection performance, Reason-IAD also enhances interpretability by explicitly modeling reasoning trajectories in a compact latent space and aligning visual evidence with category-specific knowledge priors. We expect this design to facilitate future research on reasoning-centric visual understanding and training-free multimodal adaptation in industrial scenarios.

However, several limitations still exist. First, the inference efficiency of Reason-IAD is lower compared to single-pass MLLM-based methods, as the multi-step latent reasoning process introduces additional computational overhead. Second, the performance of the framework is partially dependent on the capability of the underlying MLLM to follow retrieval-augmented knowledge and reasoning instructions.

Despite these limitations, we believe Reason-IAD offers a strong foundation for future research on knowledge-driven multimodal reasoning in industrial anomaly detection and can serve as an effective baseline for subsequent advancements in this area.



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

Answer:
Looking at the image, I need to find which candle has a defect. I am examining each candle one by one: Top left candle: appears normal, wick is intact. Top right candle: appears normal, wick is intact. Bottom left candle: appears normal, wick is intact. Bottom right candle: I can see a small, dark, smudged area near the wick, which looks like a burn mark or a defect from previous use. This defect is clearly visible on the bottom right candle. <answer>B. Bottom right candle</answer>

Figure 7: A case study of the defect location task on the Visa dataset.



Question:
Is there any defect in the object?
A. Yes. B. No.

Answer:
The object in the image is a screw. Upon close inspection, the head of the screw shows visible damage or wear, particularly around the slot where a driver would be inserted. The slot appears chipped or broken, which would prevent a standard screwdriver from engaging properly. This is a clear defect that compromises the functionality of the screw.<answer>A. Yes.</answer>

Figure 8: A case study of the anomaly discrimination task on the MVTec-AD dataset.



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

Answer:
Looking at the image, I need to find where the defect is located. The image shows a brown leather-like surface with a characteristic grain pattern. I need to scan the entire image to locate any imperfections. Scanning the image: The top right area appears to have the normal leather texture. The bottom right area also appears normal. The center area has the typical grain pattern. Looking at the center left area, I can see a small, dark, irregular spot that breaks the uniformity of the leather texture. This appears to be a small hole or tear in the material. This defect is located in the center left portion of the image. <answer>C. Center left.</answer>

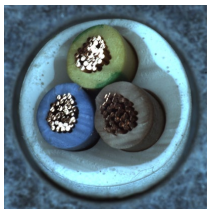
Figure 9: A case study of the defect location task on the Visa dataset.



Question:
There is a defect in the object. What is the appearance of the defect?
A. Peeling paint B. Circular stain
C. Vertical indent D. Horizontal scratch

Answer:
Looking at the image of the Coca-Cola can, I can see a clear defect on the red surface. There is a noticeable white mark running vertically along the side of the can, interrupting the red background. This mark appears to be a scratch or scuff, not a stain, hole, or faded area. It is elongated and follows a vertical path, which matches option A. <answer>A. Peeling paint </answer>

Figure 10: A case study of the defect description task on the GoodsAD dataset.



Question:
What kind of product is in the image?
A. Rubber hose. B. Three-core electrical cable.
C. Plumbing pipe. D. Fiber optic cable.

Answer:
The image shows a cross-section of a cable with three distinct inner conductors, each surrounded by insulation. The conductors appear to be made of copper (shiny, metallic strands), which is typical for electrical wiring. The outer layer is a white sheath, likely for protection and bundling. This structure is characteristic of a multi-core electrical cable used for transmitting electrical power or signals. The presence of three insulated copper conductors strongly indicates a three-core electrical cable, commonly used in electrical installations. <answer>B. Three-core electrical cable.</answer>

Figure 11: A case study of the object classification task on the MVTec-AD dataset.



Question:
Is there any defect in the object?
A. No. B. Yes.

Answer:
The object in the image is a hazelnut. Upon close inspection, there is a noticeable hole or indentation on the surface of the nut, particularly visible on the right side. This appears to be a physical defect, possibly caused by insect damage, a natural blemish, or a break in the shell. Since the question asks whether there is any defect, and a clear imperfection is visible, the correct answer is that there is a defect. <answer>B. Yes.</answer>

Figure 12: A case study of the anomaly discrimination task on the MVTec-AD dataset.



Question:
There is a defect in the object. What is the appearance of the defect?
A. The bottle is dented
B. The label is peeling off
C. The cap is slightly ajar on one side
D. The cap is completely missing

Answer:
Looking at the image, I need to carefully examine the bottle for any defects among the given options. I can see the bottle is standing upright and appears to have a normal, smooth shape without any visible dents or deformations. The label appears to be intact and fully adhered to the bottle. There are no signs of peeling or lifting at any point. Looking closely at the cap, it appears to be securely closed. There is no visible gap or misalignment on any side. The cap is clearly present on top of the bottle. After carefully examining all aspects of the bottle, I can see that there is no defect present among the listed options. The bottle appears to be in perfect condition. Upon closer inspection, I notice that the cap appears to be slightly misaligned or not perfectly seated, which could be interpreted as \"slightly ajar on one side\". This is the most plausible defect among the options, even though it's subtle. <answer>C. The cap is slightly ajar on one side</answer>

Figure 13: A case study of the defect description task on the GoodsAD dataset.



Question:
There is a defect in the object. Where is the defect?
A. Bottom center of the image
B. Left side of the image
C. Right side of the image
D. Top center of the image

Answer:
Looking at the image, I see a close-up view looking down into the opening of a bottle. The bottle appears to be made of dark glass. There are visible cracks and chips along the rim of the bottle opening. Specifically, there are noticeable broken pieces on the right side of the rim, near the bottom right quadrant of the image. The defect is clearly localized to the right side of the object. <answer>C. Right side of the image</answer>

Figure 14: A failure case study of the defect location task on the MVTec-AD dataset.



Question:
Is there any defect in the object?
A. No. B. Yes.

Answer:
The image shows a clear plastic bag containing hardware: two bolts and several nuts and washers. The items appear to be standard fasteners, and there is no visible damage, deformation, or missing parts. The bag itself is intact, and the contents are neatly arranged. There are no obvious defects in the objects shown.<answer>A. No</answer>

Figure 15: A failure case study of the anomaly discrimination task on the MVTec-LOCO dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions and scope of this work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in conclusion and include a section for it in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include formal theoretical results such as theorems or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we include detailed information in the experiment section and in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will make code, data and models public.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We confirmed they are provided in experiment section and in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or statistical significance measures, as the experiments are conducted under fixed settings and follow standard evaluation protocols in anomaly detection benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We confirmed we include this information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and fully adhered to the NeurIPS Code of Ethics throughout our research process, including data usage, experimental design, and reporting. No ethical concerns were identified in the course of this work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential broader impacts of the proposed method. On the positive side, it can improve industrial anomaly detection systems and enhance automation and quality control processes. On the negative side, incorrect or unreliable predictions may lead to missed defects or false alarms in industrial settings, which could affect downstream decision-making.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed method does not involve high-risk or dual-use model or dataset release. The work focuses on anomaly detection using publicly available datasets and does not introduce generative models or systems that require additional safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses publicly available datasets and open-source models. We properly cite all original sources and follow their respective licenses and terms of use. All datasets and assets used in the experiments are standard benchmarks widely adopted in the community.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not introduce new datasets, models, or code assets. The experiments are conducted using publicly available benchmarks and existing models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not include any crowdsourcing studies or human subject experiments. All evaluations are performed on standard public datasets for anomaly detection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects, crowdsourcing, or user studies, and therefore no IRB approval or equivalent ethical review was required.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper utilizes a large language model as part of the reasoning component in the proposed framework. The usage of the LLM is clearly described in the method section, and it is not used for training but as an inference-time module to assist in reasoning for anomaly detection.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.