AirIAD: Agentic Iterative Reasoning for Industrial Anomaly Detection

Huan Yu[1*] **Tiecheng Zhang**[1*] **Jin Wang**[1†] **Jingru Yang**[2] **Kaixiang Huang**[1]
Zhongyang Sun[2] **Guodong Lu**[1] **Shengfeng He**[3]
[1]Zhejiang University [2]Tencent [3]Singapore Management University

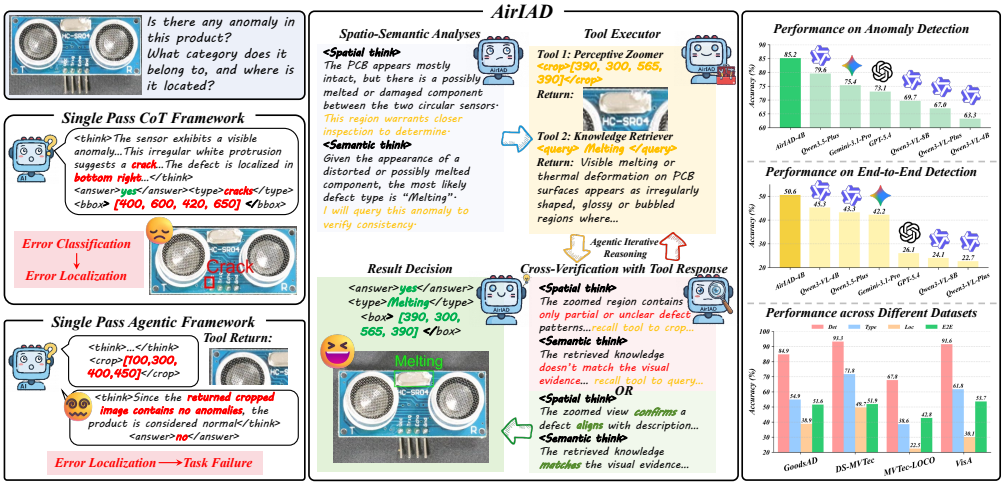


Figure 1: **Overview of AirIAD.** AirIAD formulates fine-grained anomaly analysis as an iterative spatio-semantic matching process. Guided by the SSCV CoT, the agent alternates between targeted visual inspection (Perceptive Zoomer) and semantic retrieval (Knowledge Retriever), using cross-verification to progressively refine its hypothesis until reaching a consistent diagnosis.

Abstract

Multimodal large language models offer a promising foundation for industrial anomaly detection (IAD), yet they remain unreliable for fine-grained diagnosis, where subtle defects require tight alignment between localized visual evidence and domain-specific semantics. Existing single-pass frameworks produce predictions without revisiting or validating intermediate evidence, which leads to spatio-semantic misalignment and error accumulation. More importantly, they lack mechanisms to actively control how additional evidence is acquired and integrated during reasoning, limiting their ability to resolve ambiguous cases. To address these limitations, we propose AirIAD, an agentic iterative reasoning framework that unifies progressive refinement with active information acquisition. AirIAD equips the model with two tools, Perceptive Zoomer and Knowledge Retriever, which enable it to iteratively gather complementary visual and semantic evidence. Through this process, the model performs explicit cross-verification between what is observed and what is known, refining intermediate hypotheses until convergence on a reliable diagnosis. This capability is enabled by a two-stage training pipeline. A supervised fine-tuning stage introduces a Spatio-Semantic Cross-Verification Chain-of-Thought, which structures coordinated perception and knowledge grounding. This is followed by a Spatio-Semantic Group-in-Group Policy Optimization stage, which provides dense step-wise supervision to encourage consistent self-refinement and robust spatio-semantic alignment. Extensive experiments on the

*Equal contribution.
†Corresponding author: dwjcom@zju.edu.cn.

MMAD benchmark show that AirIAD, built on Qwen3-VL-Instruct-4B, achieves state-of-the-art performance in anomaly detection and fine-grained diagnosis. More details at: https://github.com/FisherYuuri/AirIAD.

1 Introduction

Visual anomaly detection, which aims to identify abnormal patterns from images, is a fundamental problem in computer vision with broad applications in industrial inspection and beyond [1, 2]. Recent advances in multimodal large language models (MLLMs) [3–6] have expanded this paradigm by enabling joint spatio-semantic reasoning, offering improved interpretability and generalization to unseen anomalies. To address the domain-specific nature of anomaly patterns, prior works [7–23] incorporate expert knowledge and enhance visual reasoning through retrieval-augmented generation, supervised fine-tuning (SFT), and reinforcement learning (RL), achieving strong performance in binary detection.

Despite these advances, performance drops substantially in fine-grained tasks such as anomaly classification and spatial localization. We attribute this gap to a more fundamental challenge, which we term *open-vocabulary diagnostic visual reasoning*. Unlike query-conditioned settings [24], where targets are explicitly specified, fine-grained anomaly analysis requires the model to simultaneously determine *whether* an anomaly exists, *what* it is, and *where* it occurs, without predefined categories or explicit guidance. This demands precise alignment between localized visual evidence and domain-specific semantics, even for rare or previously unseen defect types.

Current MLLM-based approaches are not well-suited to this setting. Single-pass reasoning produces predictions without revisiting intermediate evidence, making it prone to spatio-semantic misalignment and error accumulation when visual cues are subtle or ambiguous. Moreover, existing methods lack mechanisms to actively control how additional evidence is acquired and integrated during inference. In practice, reliable diagnosis often depends on selectively inspecting informative regions and incorporating relevant domain knowledge, which requires both iterative refinement and goal-driven information acquisition.

Recent visual agents [25–27] address similar challenges by framing reasoning as an iterative, tool-augmented process. Inspired by this paradigm, we propose **AirIAD**, an **a**gentic **i**terative **r**easoning framework for IAD that reformulates fine-grained anomaly analysis as **a process of iterative spatio-semantic matching**. As illustrated in Figure 1, AirIAD operates under a Spatio-Semantic Cross-Verification (SSCV) CoT paradigm. Starting from an initial hypothesis over anomaly category and location, the agent iteratively acquires complementary evidence by invoking two tools: a Perceptive Zoomer for focused visual inspection and a Knowledge Retriever for relevant semantic priors. The agent then cross-verifies the consistency between visual and semantic cues to decide whether to refine its hypothesis or produce a final diagnosis. This closed-loop process progressively aligns what the anomaly is with where it occurs, leading to more reliable fine-grained reasoning.

To effectively train this multi-step reasoning process, we adopt a two-stage pipeline consisting of SFT followed by RL. In the RL stage, we introduce Spatio-Semantic Group-in-Group Policy Optimization (SS-GiGPO), a novel extension of GiGPO [28] that enables step-level credit assignment for iterative spatio-semantic reasoning. SS-GiGPO leverages Spatio-Semantic Anchor State Grouping (SSASG) to cluster intermediate states into functionally comparable groups, forming a hierarchical supervision scheme that jointly evaluates trajectory-level outcomes and step-level improvements. This design encourages consistent self-refinement and stronger spatio-semantic alignment across iterations. Extensive experiments on the MMAD benchmark [7] demonstrate that AirIAD achieves state-of-the-art performance in both anomaly detection and fine-grained diagnosis, significantly outperforming general-purpose MLLMs and specialized baselines. Cross-domain results further highlight its potential as a general framework for open-vocabulary diagnostic visual reasoning.

In summary, our contributions are threefold:

- We propose *AirIAD*, an agentic iterative reasoning framework for IAD. By integrating SSCV CoT, AirIAD actively acquires multimodal evidence at each iteration to verify and refine intermediate spatio-semantic judgments, substantially improving fine-grained anomaly analysis.

- We introduce *SS-GiGPO*, an RL algorithm that enables step-level advantage supervision for multi-step spatio-semantic reasoning. Through SSASG, it establishes a hierarchical grouping structure for intermediate states and assigns higher credit to better-aligned actions.
- Experimental results on the MMAD benchmark demonstrate the superiority of AirIAD over leading general-purpose MLLMs and specialized baselines. Cross-domain evaluations further highlight its potential for open-vocabulary diagnostic visual reasoning tasks.

2 Related Work

Industrial Anomaly Detection. While prior memory- [29–31], reconstruction- [32–36], and CLIP-[37–41]based paradigms have achieved remarkable success in anomaly detection and zero-shot generalization, they often lack explicit interpretability in anomaly judgment and decision-making. To address this limitation, MLLMs have recently emerged as a promising direction for IAD. The MMAD benchmark [7] recently provided a systematic evaluation of this capability. Building upon this, numerous subsequent works [8–18] have introduced robust post-training paradigms based on SFT and RL to enhance anomaly perception, reasoning consistency, and interpretability. To further address the intricacies of fine-grained analysis, recent literature [19–23] has shifted towards agentic and reflective mechanisms. However, research on agentic iterative reasoning for IAD remains in its infancy. In particular, how to effectively orchestrate tool use and iterative reasoning to exploit tool-derived feedback for robust anomaly detection and analysis remains largely underexplored.

MLLM Agentic Reinforcement Learning. RL has emerged as a central paradigm for post-training agentic MLLMs, exemplified by GRPO [42] and its extensions [28, 43–45]. Building on this foundation, recent methods leverage RL to bolster MLLMs' tool utilization and iterative reasoning, effectively supporting open-ended tasks such as autonomous deep research [26], GUI navigation [46, 47], complex visual question answering [27, 48], and recommender systems [49, 50]. Despite this progress, fine-grained anomaly analysis remains underexplored in RL-based agentic MLLMs. Beyond a domain-specific application, IAD serves as a representative testbed for open-vocabulary diagnostic visual reasoning, where models must autonomously discover suspicious regions, infer defect semantics, and maintain spatio-semantic alignment under domain-specific category shifts. Similar challenges also arise in medical imaging analysis [51, 52] and remote sensing interpretation [53, 54], where reliable decisions require coupling fine-grained spatial evidence with specialized semantic knowledge rather than following an explicit target query.

3 Method

3.1 Formulation

Given a query image $\mathbf{I}$, industrial anomaly analysis requires an end-to-end prediction of whether the image is defective, what defect category it belongs to, and where the anomalous region is located. Formally, the model outputs a structured tuple $(\hat{y}_{\text{ano}}, \hat{c}, \hat{\mathbf{b}})$, where $\hat{y}_{\text{ano}} \in \{\text{normal}, \text{anomalous}\}$ denotes anomaly presence, $\hat{c} \in \mathcal{C} \cup \{\text{none}\}$ denotes the predicted defect category over the evaluation-defined candidate set $\mathcal{C}$, and $\hat{\mathbf{b}} \in \mathbb{R}^4 \cup \{\varnothing\}$ is the predicted bounding box. Unlike the MMAD benchmark, which evaluates MLLMs via isolated multiple-choice questions for detection, classification, and coarse localization, our formulation requires a unified fine-grained diagnosis that jointly resolves anomaly presence, defect category, and precise spatial localization.

3.2 AirIAD Agent Framework

Motivation. Conventional paradigms in industrial anomaly analysis typically attempt to estimate the defect category and spatial location in a single forward pass. However, such static inference schemes often suffer from spatio-semantic misalignment and cascading errors, particularly when confronted with complex or ambiguous defects. To address this limitation, AirIAD reformulates anomaly analysis as an **iterative spatio-semantic matching process**. Formally, the goal is to identify the defect category c^* and bounding box $\mathbf{b}^*$ whose localized visual evidence best aligns with the category-conditioned expert knowledge:

$$(c^*, \mathbf{b}^*) = \arg\max_{c \in \mathcal{C},\, \mathbf{b}} p(c, \mathbf{b} \mid \mathbf{I}_{\mathbf{b}}, \mathcal{K}_c)\,, \tag{1}$$

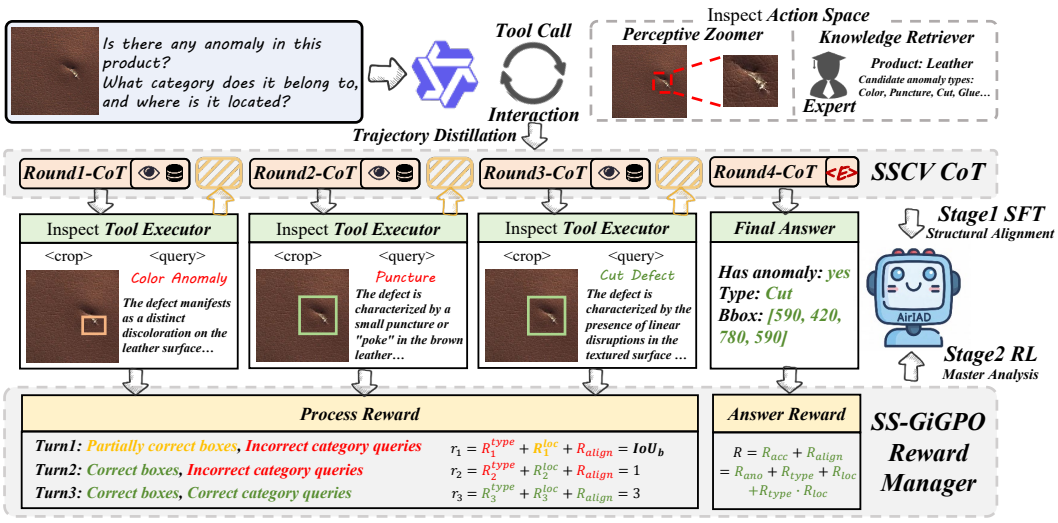


Figure 2: **Training pipeline of AirIAD.** The optimization follows a two-stage curriculum. **Stage 1: Warm-up SFT** instills the foundational SSCV CoT formatting. **Stage 2: Spatio-Semantic GiGPO (SS-GiGPO)** further enhances the model's spatio-semantic reasoning capability through RL.

where $\mathbf{I}_{\mathbf{b}}$ denotes the image crop specified by $\mathbf{b}$, and $\mathcal{K}_c$ represents the expert semantic knowledge retrieved for candidate category c from a predefined expert knowledge base.

Pipeline of AirIAD. Since directly computing this optimal state is computationally intractable, AirIAD approximates it through a progressive MLLM-driven process of hypothesis prediction, evidence verification, and iterative refinement. As shown in Figures 1 and 2, given an input image $\mathbf{I}$ and a set of candidate anomaly options $\mathcal{C}$ as the initial state s_1, the agent generates its initial diagnostic hypothesis $(c^{(1)}, \mathbf{b}^{(1)})$. Subsequently, at each step $t \geq 1$, to validate the current hypothesis $(c^{(t)}, \mathbf{b}^{(t)})$, AirIAD proactively invokes a unified `[INSPECT]` action a_t^{ins} that simultaneously triggers two specialized tools. Specifically, the Perceptive Zoomer utilizes $\mathbf{b}^{(t)}$ to execute a cropping operation, returning a localized visual region $\mathbf{I}_{\mathbf{b}^{(t)}}$; simultaneously, the Knowledge Retriever queries $c^{(t)}$ to fetch its corresponding semantic description $\mathcal{K}_{c^{(t)}}$ from a predefined expert knowledge base. The feedback from these tools is integrated to form the next state s_{t+1}, where the agent performs cross-verification. If the spatio-semantic alignment is weak, AirIAD refines the hypothesis to $(c^{(t+1)}, \mathbf{b}^{(t+1)})$ and re-invokes the tools, engaging in agentic iterative reasoning until a diagnostic consensus is reached. We formalize this process as a trajectory of alternating states and actions:

$$\tau = \left(s_1,\ a_1^{\text{ins}},\ s_2,\ \ldots,\ a_k^{\text{ins}},\ s_{k+1},\ a_{k+1}^{\text{ans}}\right), \tag{2}$$

where a_{k+1}^{ans} yields the final diagnostic tuple $(\hat{y}_{\text{ano}}, \hat{c}, \hat{\mathbf{b}})$, with $(\hat{c}, \hat{\mathbf{b}})$ serving as an empirical approximation of the optimal spatio-semantic target $(c^*, \mathbf{b}^*)$ when an anomaly is detected.

Spatio-Semantic Cross-Verification (SSCV) CoT. The core engine driving this iterative pipeline is the SSCV CoT. At each step t, prior to generating the action a_{t+1}^{ins} or a_{t+1}^{ans}, the agent produces two decoupled reasoning streams enclosed within `<spatial>` and `<semantic>` tags. The `<spatial>` stream analyzes the localized visual crop, verifying its alignment with semantic priors to determine if the spatial boundary $\mathbf{b}^{(t)}$ requires adjustment. Conversely, the `<semantic>` stream reviews the retrieved knowledge, cross-referencing it against the visual evidence to validate or correct the category hypothesis $c^{(t)}$. By forcing the agent to converge on a mutually supportive $(c^{(t+1)}, \mathbf{b}^{(t+1)})$ pair before determining the subsequent tool invocation or final diagnosis, this mechanism rigorously intercepts error accumulation and promotes precise spatio-semantic alignment.

3.3 Training Pipeline

As illustrated in Figure 2, the training pipeline of AirIAD follows a progressive two-stage curriculum. The first stage, Warm-up SFT, aims to instill the foundational SSCV CoT formatting and tool interaction syntax, equipping the model with the basic capacity for multi-step reasoning. Subsequently, the

second stage introduces Spatio-Semantic GiGPO, a specialized RL phase designed to strengthen the agent's analytical capacity to align spatial and semantic cues and orchestrate logical tool invocations.

Stage I: Warm-Up Supervised Fine-Tuning. To initialize the model with the required reasoning and interaction priors, we distill agentic trajectories from Qwen3-VL-235B-A22B [3] to construct the SFT data. Each distilled trajectory comprises a variable number of `[INSPECT]` iterations followed by a final `[ANSWER]`. After removing logically inconsistent or redundant samples, we train the model while masking observation tokens, such that gradients are applied only to reasoning traces and action outputs. This stage equips the model with the SSCV CoT format, allowing the subsequent RL stage to focus on optimizing spatio-semantic alignment.

Stage II: Spatio-Semantic GiGPO. Although GRPO has achieved strong performance in general reasoning tasks, its sparse episode-level rewards are inadequate for optimizing AirIAD, whose multi-step agentic trajectories critically depend on informative intermediate exploration and iterative refinement. As illustrated in Figure 3, we therefore propose Spatio-Semantic GiGPO (SS-GiGPO), which enables fine-grained credit assignment by evaluating the relative advantages of grouped actions under identical states, encouraging progressive self-refinement throughout the reasoning process.

Preliminary: Group-in-Group Policy Optimization. To better address credit assignment in multi-turn agentic RL, GiGPO [28] improves upon GRPO by introducing a two-level hierarchical framework. Specifically, it computes macro advantages at the episode level across complete trajectories, and micro advantages at the step level by dynamically grouping actions that stem from identical environment states. Following this formulation, global outcomes and local decisions are combined via a hierarchical advantage signal:

$$A(a_t^{(i)}) = A_E(\tau^{(i)}) + \omega \cdot A_S(a_t^{(i)}), \tag{3}$$

where ω balances the macro advantage A_E, which evaluates global episode quality, and the micro advantage A_S, which assesses step-level actions within shared states, to provide hierarchical supervision. Following the advantage estimation paradigm of GiGPO detailed in Appendix A, we empirically set $\omega = 1$ to equally weight these signals, with additional sensitivity analysis provided in Appendix C. The clipped policy optimization objective of GiGPO is defined as:

$$\begin{aligned}\mathcal{J}_{\text{GiGPO}}(\theta) = \mathbb{E}_{\substack{x\sim p(X) \\ \{\tau_i\}_{i=1}^{N}\sim\pi_{\theta_{\text{old}}}}} \Bigg[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \min\Big(\rho_\theta(a_t^{(i)})A(a_t^{(i)}), \text{clip}(\rho_\theta(a_t^{(i)}), 1\pm\epsilon)A(a_t^{(i)})\Big)\Bigg] \\ - \beta\mathbb{D}_{\text{KL}}(\pi_\theta(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x)).\end{aligned} \tag{4}$$

where $\rho_\theta(a_t^{(i)})$ is the importance sampling ratio and β is the KL penalty coefficient.

Spatio-Semantic Anchor State Grouping (SSASG). While standard GiGPO relies on exact state matching for step-level grouping, AirIAD operates in a progressive spatio-semantic grounding setting, where intermediate states are defined by continuously varying inspected regions and candidate defect semantics, making exact matches rare. To address this challenge, we introduce Spatio-Semantic Anchor State Grouping (SSASG) to identify functionally equivalent diagnostic contexts. Specifically, we group current actions $\{a_t^{(i)}\}_{i=1}^{N}$ by the similarity of their preceding hypotheses $(c^{(t-1,i)}, \mathbf{b}^{(t-1,i)})$, which constitute the anchor state for the decision $a_t^{(i)}$. Two actions share the same anchor state $\tilde{s} = (\tilde{c}, \tilde{\mathbf{b}})$ if their preceding categories are identical and their spatial regions significantly overlap (e.g., $\text{IoU} > 0.5$). The resulting step-level group $G^S(\tilde{s})$ is defined as:

$$G^S(\tilde{s}) = \left\{ \left(a_t^{(i)}, r_t^{(i)}\right) \;\middle|\; c^{(t-1,i)} = \tilde{c} \;\wedge\; \text{IoU}\Big(\mathbf{b}^{(t-1,i)}, \tilde{\mathbf{b}}\Big) > 0.5 \right\}, \tag{5}$$

where $r_t^{(i)}$ denotes the process reward obtained by the i-th trajectory at step t. By computing the micro advantage within these groups, the agent identifies the most effective actions given the same prior hypothesis, encouraging the retrieval of discriminative evidence and iterative refinement to bridge ambiguous intermediate states with the final diagnostic consensus.

Reward Formulation. To support SS-GiGPO training, the RL objective is driven by two complementary reward streams: an episode-level *Answer Reward* (R), which evaluates the correctness of the final prediction, and a step-wise *Process Reward* (r_t), which assesses the quality of intermediate queries during iterative reasoning and supports the computation of the micro-level advantage A_S.

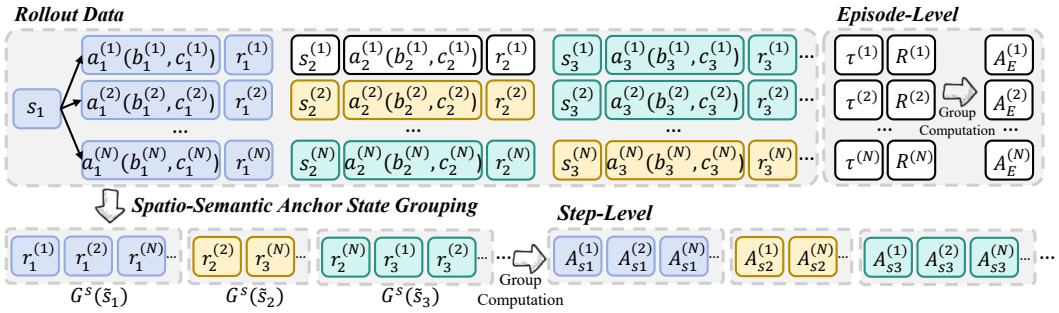


Figure 3: **Overview of Spatio-Semantic GiGPO.** Given a group of environments initialized from the same input, the agent interacts with each environment to produce a set of trajectories. States with the same color denote aligned intermediate spatio-semantic states. SS-GiGPO performs hierarchical group-wise credit assignment at two levels: episode-level relative comparison across trajectories and state-aligned step-level comparison across matched intermediate states.

Answer Reward. The answer reward evaluates the final diagnostic prediction through two complementary components: a base accuracy term R_{acc} and a joint alignment bonus R_{align}:

$$R = R_{\text{acc}} + R_{\text{align}} = \big(R_{\text{ano}} + R_{\text{type}} + R_{\text{loc}}\big) + \big(R_{\text{type}} \cdot R_{\text{loc}}\big), \tag{6}$$

where the base term independently evaluates binary anomaly detection ($R_{\text{ano}} = \mathbb{I}[\hat{y}_{\text{ano}} = y_{\text{ano}_{\text{gt}}}]$), fine-grained defect classification ($R_{\text{type}} = \mathbb{I}[\hat{c} = c_{\text{gt}}]$), and spatial localization quality (R_{loc}), where $\mathbb{I}[\cdot]$ equals 1 if the prediction is correct and 0 otherwise. Crucially, the alignment bonus $R_{\text{align}} = R_{\text{type}} \cdot R_{\text{loc}}$ is activated only when the model simultaneously predicts the correct defect category ($\hat{c} = c_{\text{gt}}$) and localizes it accurately. This multiplicative coupling explicitly encourages semantic correctness and spatial precision to be optimized jointly rather than independently.

Dynamic Localization Reward. To prevent zero-reward stagnation during early training when spatial predictions are inherently noisy, R_{loc} employs a progressive curriculum. We define a dynamic threshold $\theta(\varphi) = 0.1 + 0.4 \cdot \varphi$ that increases linearly with the training progress $\varphi \in [0, 1]$. The localization reward grants full credit once the predicted box surpasses the current threshold, and otherwise provides proportional partial credit according to the achieved IoU:

$$R_{\text{loc}} = \begin{cases} 1 & \text{if } \text{IoU}(\hat{\mathbf{b}}, \mathbf{b}_{\text{gt}}) \geq \theta(\varphi), \\ \text{IoU}(\hat{\mathbf{b}}, \mathbf{b}_{\text{gt}}) & \text{otherwise} \end{cases} \tag{7}$$

Process Reward. To instantiate the micro-level advantage A_S in Eq. 3, we assign a step-wise process reward r_t at each interaction step t:

$$r_t = R_t^{\text{type}} + R_t^{\text{loc}} + R_t^{\text{align}}. \tag{8}$$

where the terms are computed analogously to the answer reward, but over the tool-call parameters, namely the queried defect type and inspected crop. This rule-verifiable reward measures the quality of each tool invocation, allowing AirIAD to compare actions within functionally equivalent states and favor well-aligned refinements over regressive exploration, promoting progressive refinement.

4 Experiments

4.1 Experimental Setup

Datasets & Metrics. We evaluate AirIAD on the MMAD benchmark [7], which amalgamates four industrial datasets, including MVTec AD [55], MVTec LOCO [56], VisA [57], and GoodsAD [58], covering diverse structural and logical defects. We adopt the exact same 20%/80% train-test split as AgentIAD [20] to ensure a fair comparison. We report **Detection Acc.** for binary normal/abnormal recognition, together with **Type Classification Acc.** and **Localization Acc.** at IoU ≥ 0.5. We further emphasize two stricter joint metrics, **Ano-E2E Acc.** and **E2E Acc.**. **Ano-E2E Acc.** counts an anomalous sample as correct only when anomaly detection, type classification, and localization

Table 1: Comparison with state-of-the-art methods on the MMAD benchmark. The best results are in **bold** and second best are underlined, assessed independently within proprietary and open-source categories. Results for AirIAD are reported as $\text{Mean}_{\pm\text{Std}}$ across three random seeds.

Type	Method	Params	Zero-Shot Setting					One-Shot Setting				
			Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
Proprietary	GPT-5.4	–	71.58	56.90	15.97	12.82	28.36	73.06	64.24	19.89	15.66	26.08
	Gemini-3.1-Pro	–	**79.19**	**68.85**	**45.79**	**37.73**	**46.77**	75.36	**77.02**	**52.50**	**44.68**	42.25
	Qwen3-VL-Plus	–	68.27	47.50	29.61	21.78	37.98	67.01	60.83	35.79	24.83	22.71
	Qwen3.5-Plus	–	75.75	59.00	35.01	28.10	43.48	**79.61**	64.17	34.12	27.00	**43.27**
Open-Source	Qwen3-VL-Instruct	4B	57.55	22.32	16.07	10.72	33.53	63.27	28.90	20.35	14.03	45.28
	Qwen3-VL-Thinking	4B	64.61	31.84	13.59	9.57	38.17	65.58	42.64	17.87	13.49	30.33
	Qwen3-VL-Instruct	8B	70.23	42.90	23.39	16.61	37.19	69.70	52.78	22.96	15.64	24.05
	Qwen3-VL-Instruct	32B	70.70	41.31	27.62	19.86	44.62	68.91	54.81	34.92	25.02	33.24
	InternVL3.5	8B	59.73	24.62	10.67	7.47	39.64	49.91	12.82	3.02	2.02	37.98
	LLaVA-OneVision-1.5	8B	66.55	36.55	0.05	0.03	26.48	67.95	40.18	0.08	0.05	23.06
	MiniCPM-V-4.5	8B	65.22	40.34	5.53	4.27	31.31	64.47	55.59	3.40	2.53	8.94
Industrial-Specialized	AnomalyR1	3B	57.92	50.12	–	–	–	59.55	56.74	–	–	–
	JUDO	7B	67.62	49.20	–	–	–	68.89	55.51	–	–	–
	IAD-R1	7B	65.92	32.38	–	–	–	71.95	30.82	–	–	–
	RA-Monitor	8B	68.01	47.11	30.63	11.67	35.42	–	–	–	–	–
Agentic	AgentIAD	4B	69.58	28.36	7.93	3.38	32.52	72.36	32.45	11.65	7.52	36.03
	AirIAD	4B	$82.81_{\pm 0.43}$	$51.04_{\pm 0.72}$	$\mathbf{33.38}_{\pm 0.33}$	$24.00_{\pm 0.08}$	$\mathbf{49.86}_{\pm 0.27}$	$\mathbf{85.03}_{\pm 0.15}$	$57.54_{\pm 0.53}$	$36.43_{\pm 0.17}$	$\mathbf{27.46}_{\pm 0.03}$	$\mathbf{50.40}_{\pm 0.19}$
	Δ (vs. Qwen3-VL-Instruct-4B)		*+25.26*	*+28.72*	*+17.31*	*+13.28*	*+16.33*	*+21.76*	*+28.64*	*+16.08*	*+13.43*	*+5.12*
	AirIAD	8B	$\mathbf{84.26}_{\pm 0.36}$	$\mathbf{53.59}_{\pm 0.60}$	$32.65_{\pm 0.17}$	$\mathbf{24.63}_{\pm 0.05}$	$49.73_{\pm 0.31}$	$82.77_{\pm 0.10}$	$\mathbf{61.76}_{\pm 0.48}$	$\mathbf{36.84}_{\pm 0.21}$	$27.40_{\pm 0.06}$	$44.81_{\pm 0.18}$
	Δ (vs. Qwen3-VL-Instruct-8B)		*+14.03*	*+10.69*	*+9.26*	*+8.02*	*+12.54*	*+13.07*	*+8.98*	*+13.88*	*+11.76*	*+20.76*

are all correct, while **E2E Acc.** further incorporates normal samples to reflect overall detection and analysis performance.

Implementation & Baselines. AirIAD is built on Qwen3-VL-Instruct, with both 4B and 8B variants evaluated. We assess all methods under two standard protocols: *zero-shot setting*, using only the query image, and *one-shot setting*, with an additional normal reference image. Comparisons include general-purpose MLLMs [3–6, 59–61] and industrial-specific methods such as AnomalyR1 [11], Judo [10], IAD-R1 [13], RA-Monitor [21], and AgentIAD [20]. Since AgentIAD does not release official weights, we reproduce its architecture on Qwen3-VL-Instruct-4B and train it on the exact same dataset as AirIAD for a strictly fair comparison. Details of dataset, trajectory synthesis, training schedules, and inference settings are provided in Appendix B.

4.2 Main Results

Superiority over General-Purpose MLLMs. Table 1 details the comprehensive evaluation on the MMAD benchmark. Across all evaluated methods, the AirIAD framework (encompassing both the 4B and 8B variants) achieves the highest anomaly detection accuracy (84.26% for zero-shot, 85.03% for one-shot) and end-to-end accuracy (49.86% for zero-shot, 50.40% for one-shot). Notably, AirIAD consistently surpasses all open-source and industrial-specialized paradigms and significantly outperforms its base model across all metrics (e.g., Qwen3-VL-Instruct-4B, E2E jumps from 33.53% to 49.86% under zero-shot). This substantial margin directly validates the efficacy of our agentic iterative reasoning mechanism and the two-stage training pipeline.

Superiority over Specialized Frameworks. Unlike AnomalyR1 and IAD-R1, which lack bounding box localization for end-to-end deployment, RA-Monitor and AgentIAD support explicit localization but remain severely bottlenecked by inadequate visual grounding. For instance, RA-Monitor's reliance on post-hoc reflection yields a mere 11.67% Ano-E2E. By outperforming RA-Monitor by +12.33% in Ano-E2E and AgentIAD by +17.34% in E2E, AirIAD demonstrates the effectiveness of rigorous iterative spatio-semantic alignment for accurate industrial inspection.

Diagnostic Bias in MLLM-based IAD. While the one-shot protocol, explicit-reasoning paradigms, and massive parameter scales demonstrate superior capabilities in fine-grained anomaly analysis, we observe that they tend to suffer from a *pessimistic diagnostic bias*. By over-analyzing benign variations, they often misclassify normal samples, as reflected by Gemini's normal-image accuracy dropping to 60.93% despite its 90.84% anomaly detection rate, ultimately degrading E2E performance. These inherent limitations in single-pass inference resist brute-force parameter scaling or generic CoT enhancements. In contrast, AirIAD's agentic iterative paradigm effectively mitigates this tendency: by incorporating external information intervention, it empowers the model to reconsider and reduce potential misjudgments, significantly elevating overall diagnostic accuracy.

4.3 Ablation Studies

Unless otherwise specified, all ablation experiments are conducted using Qwen3-VL-Instruct-4B as the base model and evaluated under the one-shot setting with a fixed random seed (Seed 42).

Reasoning Format and Framework Design. Table 2 studies the effects of reasoning format and agentic structure. All variants follow the same two-stage pipeline, using format-specific warm-up SFT followed by outcome-level GRPO. Under the *single-pass* setting, SSCV CoT consistently outperforms both the no-CoT and free-form CoT baseline, showing that explicitly structuring spatial and semantic reasoning provides a stronger inductive bias for anomaly analysis. Introducing the *agentic loop* further improves all reasoning formats, indicating that iterative tool-augmented interaction is complementary to internal reasoning. Notably, combining SSCV CoT with the agentic loop yields the best performance across all metrics, confirming that structured spatio-semantic reasoning and iterative evidence acquisition are mutually reinforcing.

Table 2: Ablation on reasoning format and framework design.

Structure	CoT Format	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
Single-pass	None	69.80	32.20	21.30	18.71	40.69
	Free-form	72.36	43.43	26.16	19.63	41.05
	SSCV	**75.63**	**46.53**	**29.23**	**22.09**	**43.09**
Agentic Loop	None	70.72	37.09	22.70	20.60	43.58
	Free-form	77.53	47.33	30.38	21.65	44.39
	SSCV	**80.94**	**55.08**	**33.25**	**25.26**	**45.92**

Tool Design. To dissect the individual contribution of each tool and examine their synergistic coupling, we benchmark alternative invocation strategies under the same SFT and GRPO training protocol. As shown in Table 3, providing all domain knowledge upfront (*All Know.*) degrades E2E performance relative to the *No tools* baseline, indicating that uncurated knowledge injection introduces distracting semantic noise. Furthermore, deploying either active tool in isolation (*Zoomer* or *Retriever*) yields modest improvements, as each addresses spatial inspection or semantic reference without mutual reinforcement. While sequential execution (*Cascaded*) provides incremental gains, it remains suboptimal due to unidirectional error propagation. In contrast, our synchronously **Coupled** design achieves superior results across all metrics, confirming that interleaving active visual zooming with hypothesis-driven knowledge retrieval within an iterative verification loop is essential for reliable spatio-semantic alignment.

Table 3: Ablation on tool invocation design.

Tool Config.	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
No tools	75.63	46.53	29.23	22.09	43.09
All Know.	73.27	45.82	28.15	20.50	42.92
Zoomer	79.84	47.25	30.43	22.83	43.25
Retriever	79.92	49.24	29.31	22.93	43.38
Cascaded	80.34	50.88	31.17	23.60	44.73
Coupled	**80.94**	**55.08**	**33.25**	**25.26**	**45.92**

Training Stage. Table 4 dissects the progressive impact of our two-stage training scheme. Warm-up SFT with SSCV primarily serves as behavioral initialization: by fitting demonstration trajectories, the model effectively learns structured reasoning patterns, syntactic protocols, and tool invocation formats, which improves classification accuracy from 28.90% to 34.83%. However, pure imitation learning merely mirrors the superficial form of reasoning rather than mastering autonomous visual analysis. Lacking active self-correction mechanisms, the model is prone to spatial hallucination and compounding observation drift on fine-grained grounding tasks, causing localization to degrade sharply from 20.35% to 13.05%. SS-GiGPO fundamentally resolves this limitation: driven by environmental rewards and step-wise spatial credit assignment, it forces the agent to explore and verify hypotheses against fine-grained visual evidence rather than relying on heuristic shortcuts. This converts rigid trajectory imitation into genuine spatio-semantic reasoning, substantially boosting localization to 36.32% and overall E2E diagnosis to 50.58%.

Table 4: Ablation on training stage progression.

Training Stage	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
Baseline	63.27	28.90	20.35	14.03	45.28
+ SFT w/ SSCV	64.98	34.83	13.05	10.67	40.73
+ SS-GiGPO	**85.17**	**57.56**	**36.32**	**27.49**	**50.58**

Optimization & Credit Assignment. Table 5 shows the progressive benefits of our optimization design. Starting from vanilla GRPO with sparse episode-level rewards, adding the *Joint Alignment Reward* and *Dynamic Threshold* consistently improves all metrics, indicating that explicitly coupling anomaly categorization with adaptive spatial grounding yields stronger supervision. To verify the importance of hierarchical credit assignment, we compare SS-GiGPO with a naive dense-reward baseline (+ *Dense Reward*) that directly sums step-wise rewards into the trajectory return. Despite providing denser supervision, directly summing step-wise rewards degrades both fine-grained metrics and overall E2E. Without state-aware credit assignment, intermediate rewards may promote locally plausible but globally suboptimal tool actions, introducing noise into

Table 5: Ablation on optimization strategies and credit assignment.

Config. / Paradigm	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
GRPO (Episode Base)	80.94	55.08	33.25	25.26	45.92
+ Joint Align Reward	81.94	55.34	34.14	26.08	46.73
+ Dynamic Threshold	82.89	56.39	34.91	26.13	47.16
+ Dense Reward (Direct Sum)	82.41	54.67	33.81	24.93	46.83
+ Step-level Credit (SS-GiGPO)	**85.17**	**57.56**	**36.32**	**27.49**	**50.58**

fine-grained spatio-semantic alignment. In contrast, SS-GiGPO uses SSASG to assign step-level credit within functionally equivalent states, leading to clear gains across all metrics and the best overall E2E accuracy of 50.58%.

4.4 Further Analysis

Dynamics of Iterative Active Perception. Figure 4 compares AirIAD-4B's per-turn performance against the Qwen3-VL-Instruct-4B baseline on aligned subsets. As illustrated, AirIAD maintains a significant margin for trajectories spanning ≤ 3 tool invocations, directly validating our active perception policy. Notably, $\sim$90% of samples are resolved with ≤ 1 tool call. The 1-call subset, which constitutes the largest proportion, exhibits a striking $\sim$30% accuracy improvement over the baseline. This suggests that AirIAD can achieve strong diagnostic performance with limited tool-interaction overhead. For the most challenging cases with $\geq$4 tool invocations, accuracy falls slightly below the baseline due to compounded noise and limited training-time reasoning depth. Overall, AirIAD balances precision and operational throughput by dynamically allocating verification steps to sample complexity, averaging merely 0.26s per query. Detailed computational profiling and latency breakdowns are provided in Appendix C.3.

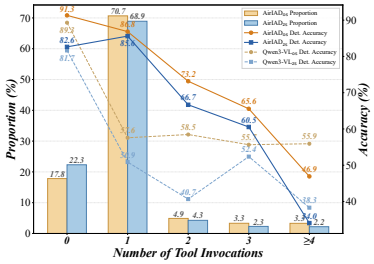


Figure 4: Performance and distribution across reasoning turns.

Self-Refinement via Multi-Turn Reasoning. Table 6 demonstrates the performance improvements achieved by both model scales through multi-turn tool invocations under the one-shot setting. Notably, tool-based verification acts as a robust filter, successfully reclassifying up to 77.3% of normal samples initially flagged as suspicious. For anomalous cases, the iterative process drives profound improvements across both semantic and spatial dimensions. Both models consistently rectify initial category errors in over 16% of multi-turn trajectories. Crucially, spatial refinement scales significantly with model capacity, with AirIAD-8B achieving progressive IoU improvements in 36.1% of cases. Overall, these dynamics highlight that the iterative paradigm effectively bridges the gap from initial ambiguous guesses to accurate diagnostic consensus. A detailed case study illustrating these self-refinement dynamics is provided in Appendix E.

Table 6: Effectiveness of multi-turn reasoning.

Model	Norm. Filtered	Cls. Corrected	IoU Improved
AirIAD-4B	77.3%	16.8%	10.1%
AirIAD-8B	50.0%	16.1%	36.1%

Dynamics of Step-Level Grouping. As illustrated in Figure 5, we track the evolution of step-level groups during AirIAD-4B training to evaluate the stability of SSASG. The effective grouping ratio consistently remains above 94%, demonstrating that our strategy robustly identifies functionally equivalent states in continuous visual space without excessive fragmentation. Furthermore, the largest groups steadily shrink: the top-5 average group size decreases from 14.8 initially to 8.2 by the end of training. This trend indicates that the policy becomes less trapped in repetitive exploratory bottlenecks and is instead capable of resolving spatio-semantic uncertainty through decisive and diversified actions.

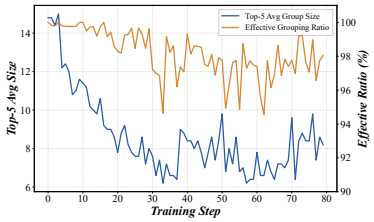


Figure 5: **Dynamics of step-level grouping.** Evolution of the effective grouping ratio and average group size during optimization.

Cross-Dataset Out-of-Distribution Generalization. Table 7 evaluates the cross-dataset out-of-distribution (OOD) generalization of AirIAD. Specifically, the objective of training is not to memorize benchmark-specific defect patterns, but rather to cultivate an iterative spatio-semantic matching capability: actively gathering multimodal evidence via dynamic tool use, cross-verifying visual and categorical cues, and iteratively

Table 7: Cross-dataset out-of-distribution evaluation.

Dataset	Model	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
GoodsAD (ID)	Qwen3-VL-4B-Instruct	54.93	21.18	15.67	12.57	45.90
	AirIAD-OOD-4B	**72.86**$_{+17.93}$	**58.18**$_{+37.00}$	**32.89**$_{+17.22}$	**27.70**$_{+15.13}$	**51.07**$_{+5.17}$
MVTec (OOD)	Qwen3-VL-4B-Instruct	75.39	48.85	37.11	26.52	44.75
	AirIAD-OOD-4B	**83.26**$_{+7.87}$	**64.57**$_{+15.72}$	**38.47**$_{+1.36}$	**27.36**$_{+0.84}$	**47.30**$_{+2.55}$
MVTec-LOCO (OOD)	Qwen3-VL-4B-Instruct	56.39	17.36	12.92	6.73	40.89
	AirIAD-OOD-4B	**62.69**$_{+6.30}$	**41.32**$_{+23.96}$	**16.29**$_{+3.37}$	**11.31**$_{+4.58}$	**44.80**$_{+3.91}$
VisA (OOD)	Qwen3-VL-4B-Instruct	70.16	28.34	15.53	9.01	47.97
	AirIAD-OOD-4B	**75.08**$_{+4.92}$	**53.85**$_{+25.51}$	**17.05**$_{+1.52}$	**9.66**$_{+0.65}$	**51.49**$_{+3.52}$
OOD Overall	Qwen3-VL-4B-Instruct	67.88	32.70	22.65	14.74	44.94
	AirIAD-OOD-4B	**74.12**$_{+6.24}$	**54.20**$_{+21.50}$	**24.64**$_{+1.99}$	**16.58**$_{+1.84}$	**48.26**$_{+3.32}$
Overall	Qwen3-VL-4B-Instruct	63.27	28.90	20.35	14.03	45.28
	AirIAD-OOD-4B	**73.67**$_{+10.40}$	**55.52**$_{+26.62}$	**27.36**$_{+7.01}$	**20.25**$_{+6.22}$	**48.84**$_{+3.56}$

refining diagnostic conclusions. To rigorously assess OOD transfer, AirIAD-OOD-4B is trained exclusively on GoodsAD (553 SFT trajectories and 134 RL instances) and directly evaluated on three unseen benchmarks. As reported, it achieves robust cross-dataset OOD transfer, improving overall OOD detection by 6.24% and classification by 21.50% over the base model. This confirms that AirIAD's iterative spatio-semantic matching operates as a domain-invariant diagnostic mechanism, functioning effectively beyond the visual and semantic distributions of the training domain.

Cross-Domain Generalization of Diagnostic Reasoning. Beyond industrial inspection, AirIAD demonstrates strong cross-domain generalization to open-vocabulary visual reasoning. To evaluate this capability without domain-specific fine-tuning, we formulate a cross-domain medical benchmark by pairing pathological scans from the Figshare brain tumor dataset [62] with healthy control slices from the BTMRI dataset [63], evaluating models solely through prompt modification. Notably, while baselines receive exhaustive textual profiles of candidate lesion types upfront, AirIAD receives only category names and autonomously retrieves diagnostic profiles through its active tool-use loop. As shown in Table 8, AirIAD consistently outperforms both the base model and the reasoning variant across all evaluation dimensions, establishing substantial leads in overall E2E diagnosis. These results confirm that AirIAD develops a transferable, domain-agnostic spatio-semantic reasoning capability.

Table 8: Cross-domain evaluation on brain tumor MRI diagnosis.

Model	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
Qwen3-VL-4B-Instruct	59.91	27.61	40.11	19.61	21.68
Qwen3-VL-4B-Thinking	80.53	46.74	37.24	19.88	24.04
AirIAD-4B (Ours)	**92.49**	**48.86**	**68.86**	**36.95**	**40.55**

Additional Ablations and In-Depth Analyses. Comprehensive analyses investigating SS-GiGPO hyperparameter sensitivity, alternative base models, and knowledge base resilience under progressive degradation are detailed in Appendix C.

5 Conclusion

In this work, we introduced AirIAD, an agentic framework that reformulates MLLM-based industrial anomaly detection as an iterative search for optimal spatio-semantic correspondence. Central to AirIAD is the SSCV CoT reasoning paradigm, which encourages the model to cross-verify spatial and semantic information for aligned reasoning, actively invoking tools to acquire targeted evidence for verification. To optimize this agentic capability, we proposed SS-GiGPO, a novel RL paradigm tailored for continuous visual environments that explicitly incentivizes progressive self-refinement by delivering fine-grained step-level credit assignment. Extensive evaluations on the MMAD benchmark confirm that our AirIAD achieves state-of-the-art performance in fine-grained industrial anomaly analysis. Beyond these empirical gains, cross-domain evaluations demonstrate the robust generalizability of this paradigm, highlighting its potential applicability in diverse real-world fields. The limitations of our method are discussed in Appendix D.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (52475033), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2026LDC01046(GZ)), and the Ningbo Science and Technology Project (2025Z058, 2024Z295). The authors declare no competing interests.

References

[1] Z. Li, Y. Yan, X. Wang, Y. Ge, and L. Meng, "A survey of deep learning for industrial visual anomaly detection," *Artificial Intelligence Review*, vol. 58, no. 9, p. 279, 2025.

[2] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.

[3] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[4] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[5] T. Yu, Z. Wang, C. Wang, F. Huang, W. Ma, Z. He, T. Cai, W. Chen, Y. Huang, Y. Zhao *et al.*, "Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe," *arXiv preprint arXiv:2509.18154*, 2025.

[6] X. An, Y. Xie, K. Yang, W. Zhang, X. Zhao, Z. Cheng, Y. Wang, S. Xu, C. Chen, D. Zhu *et al.*, "Llava-onevision-1.5: Fully open framework for democratized multimodal training," *arXiv preprint arXiv:2509.23661*, 2025.

[7] X. Jiang, J. Li, H. Deng, Y. Liu, B.-B. Gao, Y. Zhou, J. Li, C. Wang, and F. Zheng, "Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection," in *The Thirteenth International Conference on Learning Representations*, 2025.

[8] X. Yao, Z. Qian, C. Shi, J. Song, and C. Zhang, "Mmr-ad: A large-scale multimodal dataset for benchmarking general anomaly detection with multimodal large language models," *arXiv preprint arXiv:2604.10971*, 2026.

[9] K. Zhang, Z. Zhang, X. Sun, J. Nie, Q. Chen, H. Hao, J. Guo, and J. Zhang, "Adseeker: A knowledge-infused framework for anomaly detection and reasoning," *arXiv preprint arXiv:2508.03088*, 2025.

[10] H. Kang, W. Lee, J. Kim, and H. Park, "Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa," in *The Fourteenth International Conference on Learning Representations*, 2026.

[11] Y. Chao, J. Liu, J. Tang, and G. Wu, "Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection," *arXiv preprint arXiv:2504.11914*, 2025.

[12] J. Xu, S.-Y. Lo, B. Safaei, V. M. Patel, and I. Dwivedi, "Towards zero-shot anomaly detection and reasoning with multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 370–20 382.

[13] Y. Li, Y. Cao, C. Liu, Y. Xiong, X. Dong, and C. Huang, "Iad-r1: Reinforcing consistent reasoning in industrial anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 8, 2026, pp. 6583–6591.

[14] J. Zhu, Y.-S. Ong, C. Shen, and G. Pang, "Fine-grained abnormality prompt learning for zero-shot anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22 241–22 251.

[15] W. Guan, J. Lan, J. Cao, H. Tan, H. Zhu, and W. Wang, "Emit: Enhancing mllms for industrial anomaly detection via difficulty-aware grpo," *arXiv preprint arXiv:2507.21619*, 2025.

[16] Y. Li, S. Yuan, H. Wang, Q. Li, M. Liu, C. Xu, G. Shi, and W. 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*, 2025, pp. 21 917–21 926.

[17] S. Zhao, Y. Lin, L. Han, Y. Zhao, and Y. Wei, "Omniad: Detect and understand industrial anomaly via multimodal reasoning," *arXiv preprint arXiv:2505.22039*, 2025.

[18] Z. Zhang, J. Ruan, X. Gao, T. Liu, and Y. Fu, "Eiad: Explainable industrial anomaly detection via multi-modal large language models," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[19] X. Peng, X. Huang, and S. H. Choi, "Eagle: Expert-augmented attention guidance for tuning-free industrial anomaly detection in multimodal large language models," *arXiv preprint arXiv:2602.17419*, 2026.

[20] J. Miao, P. Du, Y. Liu, Y. Wang, and Y. Wang, "Agentiad: Tool-augmented single-agent for industrial anomaly detection," *arXiv preprint arXiv:2512.13671*, 2025.

[21] C. Huang, Y. Li, Y. Cao, W. Wang, H. Huang, J. Wen, W. Ren, and X. Cao, "M3-ad: Reflection-aware multi-modal, multi-category, and multi-dimensional benchmark and framework for industrial anomaly detection," *arXiv preprint arXiv:2603.00055*, 2026.

[22] P. Chen, C. Huang, Y. Cao, C. Liu, W. Wang, M. Yang, L. Shen, W. Ren, and X. Cao, "Reasoniad: Knowledge-guided dynamic latent reasoning for explainable industrial anomaly detection," *arXiv preprint arXiv:2602.09850*, 2026.

[23] J. Liao, Y. Su, R.-C. Tu, Z. Jin, W. Sun, Y. Li, X. Xu, D. Tao, and X. Yang, "Ad-fm: Multimodal llms for anomaly detection via multi-stage reasoning and fine-grained reward optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 18, 2026, pp. 15 234–15 242.

[24] L. Xiao, X. Yang, X. Lan, Y. Wang, and C. Xu, "Towards visual grounding: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[25] H. Yao, R. Zhang, J. Huang, J. Zhang, Y. Wang, B. Fang, R. Zhu, Y. Jing, S. Liu, G. Li *et al.*, "A survey on agentic multimodal large language models," *arXiv preprint arXiv:2510.10991*, 2025.

[26] T. D. Team, B. Li, B. Zhang, D. Zhang, F. Huang, G. Li, G. Chen, H. Yin, J. Wu, J. Zhou *et al.*, "Tongyi deepresearch technical report," *arXiv preprint arXiv:2510.24701*, 2025.

[27] Z. Zheng, M. Yang, J. Hong, C. Zhao, G. Xu, L. Yang, C. Shen, and X. Yu, "Deepeyes: Incentivizing" thinking with images" via reinforcement learning," *arXiv preprint arXiv:2505.14362*, 2025.

[28] L. Feng, Z. Xue, T. Liu, and B. An, "Group-in-group policy optimization for LLM agent training," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=QXEhBMNrCW

[29] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 318–14 328.

[30] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International conference on pattern recognition*. Springer, 2021, pp. 475–489.

[31] H. Li, J. Hu, B. Li, H. Chen, Y. Zheng, and C. Shen, "Target before shooting: Accurate anomaly detection and localization under one millisecond via cascade patch retrieval," *IEEE Transactions on Image Processing*, vol. 33, pp. 5606–5621, 2024.

[32] X. Xu, Y. Wang, J. Wang, X. Lei, G. Xie, G. JIANG, and Z. Lu, "FAST: Foreground-aware diffusion with accelerated sampling trajectory for segmentation-oriented anomaly synthesis," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=qqEfm8tlCM

[33] F. Beizaee, G. A. Lodygensky, C. Desrosiers, and J. Dolz, "Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 088–19 097.

[34] H. Zhang, Z. Wang, D. Zeng, Z. Wu, and Y.-G. Jiang, "Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection," *IEEE transactions on pattern analysis and machine intelligence*, 2025.

[35] J. Tebbe and J. Tayyub, "Dynamic addition of noise in a diffusion model for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3940–3949.

[36] H. H. Nguyen, C. N. Nguyen, X. T. Dao, Q. T. Duong, D. P. T. Kim, and M.-T. Pham, "Variational autoencoder for anomaly detection: A comparative study," *arXiv preprint arXiv:2408.13561*, 2024.

[37] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 606–19 616.

[38] B.-B. Gao, Y. Zhou, J. Yan, Y. Cai, W. Zhang, M. Wang, J. Liu, Y. Liu, L. Wang, and C. Wang, "Adaptclip: Adapting clip for universal visual anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 6, 2026, pp. 4095–4103.

[39] T. Gong, Q. Chu, B. Liu, W. Zhou, and N. Yu, "Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 21 220–21 230.

[40] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=buC4E91xZE

[41] Y. Li, S. Zhang, K. Li, and Q. Lao, "One-to-normal: Anomaly personalization for few-shot anomaly detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 78 371–78 393, 2024.

[42] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[43] S. He, L. Feng, Q. Wei, X. Cheng, L. Feng, and B. An, "Hierarchy-of-groups policy optimization for long-horizon agentic tasks," *arXiv preprint arXiv:2602.22817*, 2026.

[44] C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang *et al.*, "Group sequence policy optimization," *arXiv preprint arXiv:2507.18071*, 2025.

[45] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu *et al.*, "Dapo: An open-source llm reinforcement learning system at scale," *arXiv preprint arXiv:2503.14476*, 2025.

[46] R. Luo, L. Wang, W. He, L. Chen, J. Li, and X. Xia, "Gui-r1: A generalist r1-style vision-language action model for gui agents," *arXiv preprint arXiv:2504.10458*, 2025.

[47] Z. Lu, Y. Chai, Y. Guo, X. Yin, L. Liu, H. Wang, H. Xiao, S. Ren, P. Zhao, G. Liu *et al.*, "Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 21, 2026, pp. 17 608–17 616.

[48] J. Hong, C. Zhao, C. Zhu, W. Lu, G. Xu, and X. Yu, "Deepeyesv2: Toward agentic multimodal model," *arXiv preprint arXiv:2511.05271*, 2025.

[49] S. Chen, B. Chen, Y. Luo, C. Yu, Y. Ouyang, L. Cheng, C. Zhuo, Z. Li, and Y. Wang, "Vragent-r1: Boosting video recommendation with mllm-based agents via reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 3, 2026, pp. 1739–1747.

[50] Q. Wang, R. Ding, Y. Zeng, Z. Chen, L. Chen, S. Wang, P. Xie, F. Huang, and F. Zhao, "VRAG-RL: Empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=EeAHhNwXPV

[51] P. Wang, W. Lu, C. Lu, R. Zhou, M. Li, and L. Qin, "Large language model for medical images: A survey of taxonomy, systematic review, and future trends," *Big Data Mining and Analytics*, vol. 8, no. 2, p. 496, 2025.

[52] Y. Nam, D. Y. Kim, S. Kyung, J. Seo, J. M. Song, J. Kwon, J. Kim, W. Jo, H. Park, J. Sung *et al.*, "Multimodal large language models in medical imaging: current state and future directions," *Korean Journal of Radiology*, vol. 26, no. 10, p. 900, 2025.

[53] X. Sun, B. Peng, C. Zhang, F. Jin, Q. Niu, J. Liu, K. Chen, M. Li, P. Feng, Z. Bi *et al.*, "From pixels to prose: Advancing multi-modal language models for remote sensing," *arXiv preprint arXiv:2411.05826*, 2024.

[54] X. Li, J. Li, K. Zhang, Y. Fang, L. Lin, H. Wang, H. Wu, and Z. Fan, "Decoding the delta: Unifying remote sensing change detection and understanding with multimodal large language models," *arXiv preprint arXiv:2604.14044*, 2026.

[55] P. Bergmann, M. Fauser, D. Sattlegger, and C. 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*, 2019, pp. 9592–9600.

[56] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *International Journal of Computer Vision*, vol. 130, no. 4, pp. 947–969, 2022.

[57] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European conference on computer vision*. Springer, 2022, pp. 392–408.

[58] J. Zhang, R. Ding, M. Ban, and L. Dai, "Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2008–2015, 2024.

[59] OpenAI, "Introducing GPT-5.4," https://openai.com/index/introducing-gpt-5-4/, Mar. 2026, accessed: 2026-04-29.

[60] Google, "Gemini 3.1 Pro: A smarter model for your most complex tasks," https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro, Feb. 2026, accessed: 2026-04-29.

[61] Q. Team, "Qwen3. 5-omni technical report," *arXiv preprint arXiv:2604.15804*, 2026.

[62] J. Cheng, "brain tumor dataset," 2017. [Online]. Available: https://doi.org/10.6084/m9.figshare.1512427.v8

[63] AIOmarRehan, "Brain tumor mri dataset," https://huggingface.co/datasets/AIOmarRehan/Brain_Tumor_MRI_Dataset, 2026, hugging Face dataset repository. Accessed: 2026-04-17.

The supplementary material is organized as follows:

- **Appendix A**: Relative Advantage Computation of SS-GiGPO
- **Appendix B**: Implementation Details
- **Appendix C**: Additional Experimental Results and Analysis
- **Appendix D**: Limitations and Future Work
- **Appendix E**: Case Study
- **Appendix F**: Broader Impacts and Ethics

A Relative Advantage Computation of SS-GiGPO

To address the credit assignment challenge in long-horizon industrial diagnosis, SS-GiGPO decouples the optimization objective into two hierarchical tiers: macro-trajectory success and micro-step exploration quality. Specifically, we follow the GiGPO [28] to compute both episode-level and step-level relative advantages by normalizing rewards within their respective groups.

Episode-level Relative Advantage (A_E). To evaluate the global quality of each reasoning trajectory τ_i within a group of N samples generated for the same input image, we compute the episode relative advantage. This signal reinforces trajectories that lead to a correct final diagnosis:

$$A_E(\tau_i) = \frac{R(\tau_i) - \text{mean}(\{R(\tau_j)\}_{j=1}^N)}{F_{\text{norm}}(\{R(\tau_j)\}_{j=1}^N)}, \tag{9}$$

where $R(\tau_i)$ is the total cumulative return of trajectory τ_i, and F_{norm} denotes the standard deviation of the returns within the group.

Step-level Relative Advantage (A_S). To provide fine-grained feedback for intermediate tool-use and reasoning steps, we compute the step relative advantage for each action $a_t^{(i)}$ within its corresponding anchor state group $G^S(\tilde{s})$. For each action $a_t^{(i)} \in G^S(\tilde{s})$, the advantage is:

$$A_S(a_t^{(i)}) = \frac{r_t^{(i)} - \text{mean}\big(\{r_t^{(j)} \mid (a_t^{(j)}, r_t^{(j)}) \in G^S(\tilde{s})\}\big)}{F_{\text{norm}}\big(\{r_t^{(j)} \mid (a_t^{(j)}, r_t^{(j)}) \in G^S(\tilde{s})\}\big)}, \tag{10}$$

where $r_t^{(i)}$ is the step-wise reward representing the immediate quality of the t-th [INSPECT] action.

B Implementation Details

B.1 Training Details

All experiments, including the SFT and SS-GiGPO stages, were conducted on a single compute node equipped with 8× NVIDIA L20 GPUs (48GB VRAM each), dual AMD EPYC 9K84 CPUs (256 vCPUs), and 743 GB of RAM. The software stack includes TencentOS Server 3.2 and CUDA 12.8. Specifically, the SFT stage was implemented using the LLaMA-Factory framework[3]. We employ the Low-Rank Adaptation (LoRA) for efficient parameter tuning. The core hyperparameters used during the warm-up SFT stage are summarized in Table 9.

In the RL stage, we optimize the model using the *verl* framework[4]. To maintain training stability, we freeze the vision tower. The generation engine during rollout is powered by *sglang*[5] with tensor parallelism. The key hyperparameters for this stage are detailed in Table 10.

B.2 Dataset Information

Following AgentIAD [20], we use 20% of the full MMAD benchmark for training and reserve the remaining 80% for evaluation, resulting in 1,600 samples for SFT, 366 samples for RL, and 6,400

[3] `https://github.com/hiyouga/LLaMA-Factory`

[4] `https://github.com/volcengine/verl`

[5] `https://github.com/sgl-project/sglang`

Table 9: Hyperparameters for SFT.

Hyperparameter	Value
Fine-tuning Type	LoRA
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
LoRA Target Modules	All Linear
Optimizer	AdamW
Learning Rate	1.0e-4
LR Scheduler	Cosine
Warmup Ratio	0.1
Gradient Accumulation	8
Per-device Batch Size	1
Global Batch Size	64
Total Epochs	3.0
Precision	BF16

Table 10: Hyperparameters for the SS-GiGPO.

Hyperparameter	Value
Optimizer	AdamW
Actor Learning Rate	1.0e-6
Total Training Steps	80
KL Loss Coef. (β)	0.001
Entropy Coefficient	0.0
Clipping Ratio	0.2
Precision	BF16
Rollout Engine	*sglang* (TP=4)
Rollout Samples (N)	5
Rollout Mode	async
Max Turns	5
Train Batch Size	64
PPO Mini Batch Size	16
PPO Micro Batch Size	2

samples for testing. Unless otherwise specified, the same data split and training-set construction are used for all ablation studies. Furthermore, to reduce token consumption during inference and accelerate training efficiency, all images are resized to a maximum edge of 1024 pixels.

For the SFT stage, the 1,600 training samples are drawn from four source datasets: GoodsAD (614, 38.4%), VisA (392, 24.5%), DS-MVTec (332, 20.8%), and MVTec-LOCO (262, 16.4%). In terms of anomaly status, the split contains 661 normal samples (41.3%) and 939 abnormal samples (58.7%). It covers 38 product categories, with the largest classes being *drink_bottle* (149), *food_bottle* (116), and *food_package* (103). The most frequent anomaly types include *Discoloration* (76), *Scratch* (49), *Deformation* (45), and *Logical Anomaly* (39).

For the RL stage, the 366 training samples are distributed across GoodsAD (134, 36.6%), VisA (100, 27.3%), MVTec-LOCO (69, 18.9%), and DS-MVTec (63, 17.2%). The split includes 136 normal samples (37.2%) and 230 abnormal samples (62.8%). The most frequent anomaly types are *Discoloration* (21), *Surface Damage* (20), and *Deformation* (17).

The SFT set contains 1162 unique defect-type strings, while the RL set contains 433. Such exceptionally high numbers of unique descriptions arise because, in some sub-datasets of MMAD, the anomaly labels are derived from colloquial MLLM-generated responses rather than standardized terminologies. To address this severe linguistic fragmentation, we further normalize and consolidate these raw categories into a more structured taxonomy in the subsequent Knowledge Base Construction stage.

Crucially, to rigorously evaluate the open-vocabulary generalization of our framework, the testing set naturally presents a severe out-of-distribution challenge. At the sample level, 43.4% of the abnormal test samples involve ground-truth defect types that are entirely unseen during training. Furthermore, 98.7% of all test samples contain at least one unseen candidate defect type among their provided options. This highly challenging distribution ensures that the evaluation heavily relies on dynamic knowledge retrieval and active spatio-semantic alignment, rather than memorized category priors.

Cross-Dataset and Cross-Domain Evaluation Sets. To further probe generalizability beyond standard in-benchmark evaluation, we formulate two distinct transfer settings:

(i) *Cross-dataset evaluation:* To verify that AirIAD does not rely on cross-product memorization, the training data is strictly confined to the GoodsAD partition (553 SFT trajectories and 134 RL instances), and the resulting model is directly transferred and evaluated on unseen benchmarks (MVTec AD, MVTec-LOCO, and VisA) under the one-shot setting without any in-domain adaptation.

(ii) *Cross-domain medical evaluation:* To assess open-vocabulary reasoning across disparate visual domains, we pair pathological scans from the Figshare brain tumor dataset [62] with healthy controls from the BTMRI dataset [63]. The Figshare cohort provides 3,064 T1-weighted contrast-enhanced MRI slices covering three lesion categories: 1,426 glioma, 708 meningioma, and 930 pituitary tumor scans. To enable both binary anomaly detection and overall diagnostic evaluation, we introduce 613 normal brain MRI slices from BTMRI as negative controls, forming a unified 3,677-image cross-domain diagnostic benchmark.

B.3 Warm-Up SFT Data Construction

Trajectory Generation. For each training sample, we prompt the teacher model (Qwen3-VL-235B-A22B-Instruct) with system instructions defining the SSCV CoT format and the `[INSPECT]` tool interface. The model is required to generate `<spatial>` and `<semantic>` reasoning blocks prior to any action, execute `[INSPECT]` for $k \sim \mathcal{U}\{2, 10\}$ iterations, and conclude with a final `[ANSWER]` tuple. During each `[INSPECT]` step, the tool crops the hypothesized region using normalized coordinates and retrieves relevant defect profiles from the knowledge base. The resulting paired observation is then appended as the subsequent user turn. To comply with the vision encoder's constraints, crop resolutions are clamped to $[28\times28,\ 1024\times1024]$ pixels.

GT-Guided Correction. To ensure high-quality SFT supervision, we introduce a Ground Truth (GT)-guided correction mechanism for trajectories failing strict success criteria (correct detection, classification, and localization $\mathrm{IoU} \geq 0.5$). During generation, these failed trajectories receive a temporary correction prompt forcing re-evaluation: invoking an additional `[INSPECT]` step for anomalous samples, or directly revising conclusions for normal ones. To prevent compounding errors, the bounding boxes and semantic queries in these corrective steps are explicitly overridden with GT values. Crucially, these artificial prompts are discarded in the final dataset, resulting in seamlessly concatenated, flawless reasoning trajectories that reliably converge to accurate final states.

Data Format and Statistics. Each SFT record comprises a multi-turn, interleaved text-image conversation. Following the MMAD protocol, the initial visual context contains either a single query image (zero-shot) or a query paired with a defect-free reference (one-shot), with subsequent visual inputs being the crops returned by iterative `[INSPECT]` invocations. Defect categories are formalized as multiple-choice options (e.g., `A. scratch`), and tool queries strictly reference these identifiers.

The finalized distilled dataset contains 1,600 trajectories, evenly split between zero-shot and one-shot settings to foster robust generalization. The average trajectory requires $\bar{k} = 3.43$ tool invocations. The vast majority of episodes (86.44%) resolve within 2 to 5 turns—with 3-turn trajectories being the most prevalent (31.62%)—while exceptionally long chains (≥ 7 turns) account for merely 0.62%, demonstrating the teacher model's efficient convergence. Finally, all dynamically cropped sub-images are localized and referenced via path URIs, yielding a fully self-contained multimodal training corpus.

B.4 Knowledge Base Construction and Usage

The external knowledge base $\mathcal{K}$ provides structured defect profiles across all MMAD product categories, serving as the fundamental semantic grounding for the `[INSPECT]` tool. We construct this repository via a rigorous two-stage pipeline.

Stage 1: Taxonomy Canonicalization. Raw anomaly labels across the MMAD datasets often suffer from linguistic fragmentation (e.g., "Scratch", "A scratched surface"). To address this, we employ Qwen3.5-Plus to perform semantic deduplication and canonicalization. For each product class, the LLM processes raw labels to: (i) filter out non-defect noise (e.g., vague meta-labels); (ii) map synonymous surface forms to a unified canonical key, preventing semantic fragmentation; and (iii) instantiate new keys only when no existing cluster applies. Notably, during this process, inevitably unmapped ground-truth entries are intentionally preserved rather than forcefully clustered, serving as a natural testbed to evaluate the framework's fallback robustness. To guarantee discriminative retrieval during inference, we strictly enforce an intra-list uniqueness constraint, ensuring that no two candidate options in a multiple-choice list map to the same canonical key.

Stage 2: Profile Synthesis and Expert Verification. For each canonical key, we synthesize a structured textual profile. Prompted with representative anomaly images, product contexts, and label variants, Qwen3.5-Plus generates a concise 2–4 sentence characterization. These profiles systematically detail: (i) *visual morphology* (e.g., color, texture, shape, scale); (ii) *spatial distribution*; and (iii) *distinguishing features* relative to normal regions or morphologically similar defects. When applicable, the initial defect descriptions provided by the MMAD benchmark are reused as a supplementary seed dictionary. Crucially, all generated descriptions undergo a rigorous human-in-the-loop review by industrial inspection experts to ensure factual accuracy and diagnostic discriminability. The finalized $\mathcal{K}$ encompasses 486 canonical defect entries distributed across 38 product classes.

Inference Retrieval. During inference, each queried candidate option is first mapped to its canonical key via the Stage-1 taxonomy. The corresponding profile is then retrieved from $\mathcal{K}$ to serve as semantic

Table 11: Typical knowledge base examples across multiple datasets, products, and defect types.

Dataset	Product	Defect Type	Profile Snippet
MVTec AD	Zipper	Broken Teeth	Missing or damaged teeth creating a visible gap that disrupts the interlocking pattern.
		Fabric Tear	An irregular, jagged opening or split in the zipper tape fabric, often with frayed threads.
		Missing Component	Absence of the metallic or plastic slider or pull tab, leaving an empty gap along the tape.
	Metal Nut	Corrosion	Irregular, reddish-brown to dark orange patches or streaks with a granular, flaky texture.
		Broken Edge	Irregular, jagged, or missing segment along the otherwise smooth, circular perimeter.
		Threading Error	Distorted, misaligned, or incomplete helical ridges on the threaded portion.
MVTec-LOCO	Pushpins	Structural Anomaly	Bent, misshapen, or broken metal pins, and plastic head defects like breakage or detachment.
		Logical Anomaly	Missing pushpins from designated compartments or additional pushpins disrupting the pattern.
	Connectors	Missing Cable	Entire absence of the expected cable, creating a logical anomaly preventing connection.
		Cable Damage	Visible discontinuity, cut, or severe abrasion along the cable's outer insulation.
VisA	Macaroni	Contamination	Small, irregularly shaped particles or debris loosely adhering to the surface.
		Chip	Small, irregularly shaped, missing fragment along the edge, revealing interior material.
		Scratch	Thin, linear, shallow groove or streak, typically lighter or darker than surrounding material.
	PCB	Melt Damage	Areas exhibiting discoloration or melting, often appearing as reddish patches near components.
		Missing Component	Empty spot with visible solder pads where a designated electronic component is absent.
		Excess Solder	Shiny, silvery-white, globular protrusions, larger than adjacent joints, without defined shape.
GoodsAD	Food Box	Surface Damage	Irregular marks, scratches, dents, or discolorations disrupting the box's surface uniformity.
		Opened Packaging	Tear or disruption at the sealing point, indicating potential tampering or prior opening.
		Missing Component	Empty region where a structural or functional part (e.g., lid tab or divider) should be present.
	Cigarette Box	Opened Flap	Opened or partially lifted flap deviating from flat closure, indicating tampering.
		Surface Damage	Localized deformations, scratches, tears, or material failures like peeling cellophane.
		Color Inconsistency	Areas where printed colors deviate noticeably from expected hue, appearing faded or uneven.
Brain MRI	–	Glioma Tumor	An irregular intra-axial parenchymal mass with ill-defined heterogeneous margins, typically T1 hypo/isointense and T2/FLAIR hyperintense...
		Meningioma Tumor	A well-circumscribed extra-axial dural-based lesion with smooth round or lobulated contour, usually showing strong homogeneous enhancement and possible dural tail sign, compressing adjacent brain tissue rather than infiltrating it.
		Pituitary Tumor	A well-defined sellar or suprasellar mass centered in the pituitary gland within an enlarged sella turcica, typically round or oval with variable signal intensity and post-contrast enhancement.

evidence. Queries lacking a valid mapping trigger a `no_mapping` fallback, explicitly signaling to the agent that no reliable external profile is available for that specific option. In this case, the system provides only the visual description of the normal product as a baseline reference, assisting the agent in detecting and localizing potential anomalies.

B.5 Evaluation Details

Baseline Implementation Details. Different models adopt different coordinate conventions for bounding box prediction, and we unify them for fair IoU computation. For Qwen-VL, Gemini, InternVL3, LLaVA-OneVision, and MiniCPM-V, we prompt the model to output bounding boxes in the $[0, 1000]$ integer coordinate format (x_1, y_1, x_2, y_2), consistent with the grounding convention of mainstream MLLMs. For GPT models, which natively output normalized floating-point coordinates, we prompt for the (x, y, w, h) format in $[0, 1]$ and convert it to (x_1, y_1, x_2, y_2) before evaluation. For AgentIAD and RA-Monitor, we follow their default convention of $[0, 1]$ normalized (x_1, y_1, x_2, y_2) coordinates. Furthermore, due to the category discrepancy between RA-Monitor and the MMAD dataset, we employ Qwen3.5-Plus to evaluate the semantic correctness of its predictions to obtain the final classification accuracy. Additionally, for methods including Judo, IAD-R1, and Anomaly-R1, we strictly adhere to their original inference paradigms by employing MMAD's decoupled multiple-choice format for anomaly detection and classification evaluation.

Ground-Truth Bounding Box Extraction. The MMAD benchmark provides pixel-level segmentation masks for anomalous samples rather than bounding box annotations. We extract ground-truth bounding boxes from these masks using a connected-component analysis procedure: (i) the mask is binarized and connected regions are labeled via `scipy.ndimage.label`; (ii) regions whose area falls below 0.1% of the image area are discarded as noise; (iii) starting from the largest valid region, nearby regions whose boundary distance is within 5% of the longer image dimension are iteratively merged; (iv) the axis-aligned bounding box of the merged region is computed and converted to $[0, 1000]$ integer coordinates. All ground-truth boxes are verified via visual inspection to ensure spatial accuracy before evaluation.

B.6 Prompt Template

To ensure complete reproducibility and provide transparency into the interaction protocol of AirIAD, we detail the exact prompt templates used to orchestrate the agent's behavior.

System Prompt. The system prompt configures the expert persona and enforces a strict XML syntax, explicitly defining the SSCV CoT structure and the exact schemas for the two tool invocations. The exact prompt is as follows:

System Prompt for AirIAD

```
You are an industrial anomaly detection expert equipped with a visual inspection tool.

## Role
Your task is to localize and identify defects in industrial product images. You reason
    iteratively: each turn you propose a spatial crop and a semantic query, then verify
    whether the retrieved evidence is sufficient to conclude.

## Tool
You have access to one tool:

  inspect(crop, query)

  Parameters:
    crop  - [x1, y1, x2, y2], normalized floats in [0,1000], top-left to bottom-right
    query - the LETTER (A, B, C, D, etc.) of the anomaly type from the candidate list
    provided in the user message; choose the letter that best matches your current
    visual hypothesis

  Returns:
    - A zoomed image of the cropped region
```

```
    - Retrieved knowledge: text descriptions of defects matching the query

  Call this tool when you need more visual or semantic evidence. Do NOT call it if you
    have already reached a conclusion.

## Output Format
At every turn, output two brief analysis lines, then either a tool call OR a final
    answer - never both.

Analysis format (required before every action):
<think><spatial>[Evaluate the current visual evidence. Where is the suspicious region?
    Does the crop need adjustment?]</spatial>
<semantic>[Evaluate the retrieved knowledge against the visual. Does the selected
    anomaly type match what you see? Should it be refined?]</semantic></think>

Tool call format (when more evidence needed):
<tool_call>
  <name>inspect</name>
  <crop>[x1, y1, x2, y2]</crop>
  <query>A/B/C/D/...</query>
</tool_call>

Final answer format (when conclusion is reached):
<answer>
  <has_anomaly>yes/no</has_anomaly>
  <anomaly_type>A/B/C/D/... or "none"</anomaly_type>
  </bbox>[x1, y1, x2, y2] or null</bbox>
</answer>

IMPORTANT:
- Use ONLY the LETTER (A, B, C, etc.) to refer to anomaly types, NOT the full name
- If has_anomaly is "no", set anomaly_type to "none" and bbox to null.
```

User Prompt. For each inference instance, the user prompt injects the dynamic context. It provides the specific product category and the corresponding multiple-choice list of defect candidates. To support the one-shot evaluation setting common in industrial benchmarks, it also supplies a defect-free reference image alongside the target query image. The exact prompt is as follows:

User Prompt Template (One-Shot Setting)

```
Evaluate the following image from the class {CLASS_NAME}.

Candidate anomaly types:
{CANDIDATE_LIST}

A normal reference image is provided for comparison. Compare the target image with the
    reference to identify any visual differences. Determine whether this sample is
    normal or abnormal. If abnormal, identify the defect type and localize it.

You may call inspect(crop, query) to zoom into a suspicious region and retrieve
    relevant defect knowledge.

<image>

Here is the normal reference image:
<image>
```

Tool Observation Feedback. The observation template defines how the environment communicates back to the agent after an `[INSPECT]` action is executed. It returns the cropped sub-image concatenated with the semantically matched textual profiles retrieved from the knowledge base. The exact prompt is as follows:

Tool Observation Template (Environment Feedback)

```
Tool result for inspect(crop={CROP_COORDS}, query="{QUERY_LETTER}"):

<image>

Retrieved knowledge for "{CLASS_NAME}":
{RETRIEVED_KNOWLEDGE}

---
Now verify your hypothesis:
- Does the cropped image tightly cover an anomaly?
- Does the retrieved knowledge accurately describe what you see?

If YES to both -> output <answer>
If NO to either -> refine and call inspect again
```

Table 12: Sensitivity analysis of optimization hyperparameters (one-shot). Each group varies one parameter while keeping others at their default values ($\omega = 1$, $\theta(\varphi) = 0.1 \rightarrow 0.5$, $\tau_{iou} = 0.5$).

Param	Value	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
ω	0	82.89	56.39	34.91	26.13	47.16
	0.2	83.38	56.26	35.37	26.57	47.92
	0.5	84.56	56.74	35.76	26.36	48.84
	0.8	**85.33**	**57.82**	35.91	27.16	49.53
	1 (ours)	85.17	57.56	**36.32**	**27.49**	**50.58**
	1.2	84.39	56.51	35.65	26.54	49.80
	1.5	84.06	55.85	35.14	26.29	48.42
τ_{iou}	0.3	83.58	56.13	34.14	26.26	48.66
	0.5 (ours)	**85.17**	**57.56**	**36.32**	**27.49**	**50.58**
	0.7	84.34	56.39	35.24	25.67	49.25
$\theta(\varphi)$	Fixed 0.1	84.31	55.62	33.15	25.42	48.25
	Fixed 0.3	84.55	55.82	34.25	25.06	48.13
	Fixed 0.5	84.94	56.13	35.76	25.95	49.13
	$0.1 \rightarrow 0.9$	84.53	56.87	35.47	26.18	48.72
	$0.1 \rightarrow 0.5$ (ours)	**85.17**	**57.56**	**36.32**	**27.49**	**50.58**

C Additional Experimental Results and Analysis

This section provides extensive empirical analyses to complement Section 4 of the main text, comprehensively examining AirIAD's algorithmic sensitivity, architectural dependencies, operational efficiency, and statistical reproducibility. Specifically, we organize these analyses into four dimensions: (1) **Algorithmic In-Depth Ablations**, dissecting core hyperparameter choices within SS-GiGPO and foundational backbone compatibility; (2) **Knowledge Base Dependencies**, examining AirIAD's reliance on external semantic priors and evaluating its resilience under taxonomic scarcity; (3) **Inference Efficiency**, characterizing per-step computational overhead and token latency under dynamic budget allocation; and (4) **Multi-Seed Stability**, demonstrating the statistical robustness and reproducible convergence of our agentic optimization.

C.1 Algorithmic In-Depth Ablations

Sensitivity to SS-GiGPO Hyperparameters. We study AirIAD's sensitivity to key hyperparameters (ω, τ_{iou}, $\theta(\varphi)$) within SS-GiGPO on the MMAD benchmark under the one-shot setting. As shown in Table 12, the default configuration consistently achieves the optimal trade-off across all metrics.

- **Step-level advantage weight ω:** This parameter balances global episode-level (A^E) and step-level (A^S) advantages. Setting $\omega = 0$ degrades optimization to standard trajectory-level GRPO,

yielding the weakest E2E accuracy (47.16%) due to sparse feedback. While $\omega = 0.8$ maximizes Det. and Cls. accuracy, our default $\omega = 1$ provides the optimal balance for spatial tasks, achieving the best Loc. (36.32%), Ano-E2E (27.49%), and E2E accuracy (50.58%). An overly aggressive $\omega = 1.5$ degrades performance, suggesting that overpowering step-level signals causes the agent to prioritize myopic actions over the overarching goal.

- **Anchor state grouping threshold τ_{iou}:** This defines functional state equivalence in visual spaces. A permissive $\tau_{iou} = 0.3$ introduces credit assignment noise, dropping Ano-E2E to 26.26%. Conversely, a strict $\tau_{iou} = 0.7$ causes excessive state fragmentation, degrading updates back to sparse trajectory-level feedback and dropping E2E to 49.25%. The default $\tau_{iou} = 0.5$ strikes the ideal balance, providing dense and reliable gradients for self-refinement.
- **Dynamic IoU threshold $\theta(\varphi)$:** Fixed thresholds prove suboptimal: a permissive threshold (0.1) fails to enforce precise grounding, while stricter ones (0.3, 0.5) provide overly sparse early signals. An aggressive dynamic schedule ($0.1 \rightarrow 0.9$) also underperforms by over-penalizing moderately accurate boxes in later stages. In contrast, our progressive curriculum ($0.1 \rightarrow 0.5$) achieves the best overall results, effectively guiding the model from initial broad attention to precise visual grounding.

Table 13: Performance of AirIAD across different foundational base backbones.

Configuration	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
AirIAD w/ Qwen3-VL-Thinking-4B	80.97	55.39	32.33	24.67	47.03
AirIAD w/ Qwen3-VL-Instruct-4B	**85.17**	**57.56**	**36.32**	**27.49**	**50.58**

Impact of Foundational Base Models. To examine how the underlying foundation affects agentic policy optimization, Table 13 benchmarks AirIAD initialized with different 4B backbones under the one-shot setting. Both variants underwent identical two-stage SFT and SS-GiGPO training, ensuring strictly comparable tool-use syntax and format compliance. Initializing AirIAD with *Qwen3-VL-Thinking-4B* yields notably lower performance compared to the default *Qwen3-VL-Instruct-4B*: overall E2E accuracy drops from 50.58% to 47.03%, and spatial localization precision degrades from 36.32% to 32.33%. These results indicate that the instruction-tuned backbone offers superior representational plasticity and grounding adaptability for policy optimization, whereas the rigid internal reasoning priors of the thinking variant can introduce friction into the iterative, environment-driven tool interaction loop.

C.2 Knowledge Base Dependencies

Table 14: Performance of AirIAD-4B (one-shot) on abnormal samples with matched vs. unmatched defects in the knowledge base ($\mathcal{K}$).

KB Mapping	Samples	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑
Matched	3,711	83.97	**58.18**	36.08	27.41
Unmatched	196	**87.76**	45.92	**40.82**	**29.08**

Resilience to Knowledge Base Misses. To further investigate AirIAD's reliance on external semantic priors and its robustness when facing incomplete taxonomies, we partitioned the 3,907 abnormal test samples based on whether their ground-truth defect types successfully matched with the predefined profiles in $\mathcal{K}$. As detailed in Table 14, the absence of semantic priors leads to a 12.26% drop in Cls. accuracy, reaffirming that fine-grained defect categorization benefits from the taxonomic coverage of the knowledge base. In contrast, Det. and Loc. performances exhibit remarkable stability, even maintaining competitive accuracies of 87.76% and 40.82% on unmatched queries. This asymmetry reveals a critical property of our framework: spatial anomaly perception operates independently of semantic naming. When faced with novel or uncataloged defect types, AirIAD does not suffer from perceptual blindness; instead, it reliably flags and localizes anomalies by anchoring spatio-visual discrepancies against the normal reference image. This decoupling between spatial grounding and semantic categorization ensures graceful degradation in open-world deployment, guaranteeing reliable defect interception even when encountering zero-day defects outside predefined taxonomies.

Table 15: Ablation on progressive knowledge base coverage.

Setting	KB Coverage	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
Full KB (Baseline)	100%	**85.17**	**57.56**	**36.32**	27.49	**50.58**
Drop 20%	80%	85.05	57.05	**36.32**	**27.54**	50.47
Drop 40%	60%	84.92	56.13	36.29	27.41	49.44
Drop 60%	40%	84.55	55.52	36.29	27.39	48.81
Drop 80%	20%	84.30	53.49	36.27	26.87	48.20
Drop 100% (No KB)	0%	84.06	51.27	36.11	26.34	48.09

Table 16: Inference efficiency and computational overhead across varying tool call steps.

# Calls	% Samples	Prompt Tok.	Compl. Tok.	Time (s)
0	17.8%	1,582	146	0.13
1	70.7%	3,756	306	0.22
2	4.9%	6,608	506	0.41
3	3.3%	9,997	645	0.77
≥ 4	3.3%	14,605	838	1.21
Overall	100.0%	4,073	316	0.26

Progressive Knowledge Base Degradation. To assess behavior under progressive knowledge scarcity, Table 15 reports performance as the knowledge base is randomly trimmed down to complete omission. Across the entire degradation spectrum from 100% to 0% coverage, localization remains exceptionally stable, shifting by only 0.21% (from 36.32% to 36.11%) while detection drops by merely 1.11%. Although classification accuracy gradually declines from 57.56% to 51.27%, the completely unassisted model (Drop 100%) exhibits no catastrophic collapse and retains an overall E2E of 48.09%, consistently outperforming the 45.28% baseline. This resilience indicates that while external profiles in $\mathcal{K}$ enrich fine-grained diagnostic terminology, the underlying MLLM retains substantial parametric reasoning to classify standard anomalies. Taken together with Table 14, these results confirm that visual anomaly identification and spatial localization in AirIAD are firmly anchored in autonomous visual comparison rather than retrieval shortcuts.

C.3 Inference Efficiency and Dynamic Budget Allocation

Table 16 summarizes AirIAD's token consumption and inference time across different numbers of tool invocations. Overall, 88.5% of test instances terminate with at most one tool call: 17.8% require no tool calls and 70.7% require one, with average inference times of 0.13 s and 0.22 s, respectively. The remaining 11.5% involve two or more tool calls, indicating that extended verification is used for a relatively small subset of inputs. Averaged over the test set, AirIAD consumes 4,073 prompt tokens and 316 completion tokens per query, with a measured inference time of 0.26 s under our evaluation configuration. These results characterize the computational cost of iterative verification and show that most predictions are produced with limited tool interaction.

C.4 Multi-Seed Reproducibility and Stability Analysis

Table 17 presents the performance breakdown of AirIAD-4B and AirIAD-8B across three independent random seeds (1, 42, and 2026) under both zero-shot and one-shot protocols. To verify reproducibility, each seed controls training data shuffling and stochastic trajectory exploration during reinforcement learning, while strictly preserving the standard benchmark data split. Across all metrics, model scales, and both evaluation settings, AirIAD demonstrates remarkable consistency, with standard deviations strictly bounded within 0.72. Notably, the Ano-E2E standard deviations do not exceed 0.08 percentage points, confirming that the iterative cross-verification loop converges reliably rather than relying on stochastic exploration luck. Furthermore, relative performance trends and fine-grained gains remain consistent across seeds, confirming the stability of the learned agentic reasoning policy.

Table 17: Multi-seed stability evaluation of AirIAD-4B and AirIAD-8B under zero-shot and one-shot settings on the MMAD benchmark.

Model	Setting	Seed	Det.↑	Cls.↑	Loc.↑	Ano-E2E↑	E2E↑
AirIAD-4B	Zero-shot	1	82.62	50.52	33.20	24.06	50.12
		42	82.50	50.73	33.17	23.91	49.89
		2026	83.30	51.86	33.76	24.03	49.58
		Mean ± Std	**82.81** ± 0.43	**51.04** ± 0.72	**33.38** ± 0.33	**24.00** ± 0.08	**49.86** ± 0.27
	One-shot	1	84.88	57.00	36.35	27.44	50.42
		42	85.17	57.56	36.32	27.49	50.58
		2026	85.03	58.05	36.63	27.44	50.20
		Mean ± Std	**85.03** ± 0.15	**57.54** ± 0.53	**36.43** ± 0.17	**27.46** ± 0.03	**50.40** ± 0.19
AirIAD-8B	Zero-shot	1	83.89	54.21	32.66	24.65	50.02
		42	84.28	53.54	32.81	24.67	49.75
		2026	84.61	53.01	32.48	24.57	49.41
		Mean ± Std	**84.26** ± 0.36	**53.59** ± 0.60	**32.65** ± 0.17	**24.63** ± 0.05	**49.73** ± 0.31
	One-shot	1	82.66	62.22	36.63	27.46	44.61
		42	82.86	61.79	36.86	27.39	44.88
		2026	82.80	61.27	37.04	27.34	44.94
		Mean ± Std	**82.77** ± 0.10	**61.76** ± 0.48	**36.84** ± 0.21	**27.40** ± 0.06	**44.81** ± 0.18

D Limitations and Future Work

Despite promising results, our approach has several limitations that warrant future investigation.

First, while AirIAD demonstrates strong zero-shot generalization without task-specific image tuning, its fine-grained semantic categorization still depends on the vocabulary coverage of the expert knowledge base. For entirely novel anomaly types without prior semantic definitions, the agent may fail to assign an exact diagnostic label. Nevertheless, AirIAD mitigates the traditional closed-set bottleneck: recognizing new defects does not require costly model retraining, but only appending new textual profiles to the knowledge base. Moreover, even without exact semantic matches, AirIAD can still isolate abnormal visual patterns, preserving robust class-agnostic anomaly detection. Future work may explore open-vocabulary anomaly captioning with human-in-the-loop knowledge updates toward autonomous novel defect discovery.

Second, the fundamental inference time of MLLMs poses a challenge for strictly real-time industrial deployment. Notably, AirIAD already exhibits significant efficiency by utilizing a compact 4B-parameter backbone to achieve highly competitive detection results. Furthermore, the latency introduced by visual cropping and knowledge base matching is negligible compared to the model's forward-pass inference. Future work will investigate advanced quantization techniques and edge-computing adaptations for high-throughput on-device deployment.

Finally, prolonged multi-turn exploration on highly ambiguous samples can lead to context saturation. Accumulated intermediate textual feedback and visual crops may dilute attention to the global context, causing degraded predictions near the truncation boundary. Exploring efficient context compression and long-term memory mechanisms for anomaly detection is therefore important for sustaining robust long-horizon reasoning. Moreover, mitigating severe pessimistic diagnostic bias remains an important challenge for practical MLLM-based anomaly detection.

E Case Study

We select representative test samples from the MMAD benchmark to qualitatively compare the iterative reasoning trajectories of AirIAD against two state-of-the-art proprietary models, Gemini-3.1-Pro and GPT-5.4. These examples highlight three distinct diagnostic behaviors: (i) **Robust Disambiguation:** While single-pass models struggle with subtle or ambiguous defects and frequently hallucinate, AirIAD accurately identifies and localizes the anomaly through active tool invocation; (ii) **Iterative Self-Correction:** When the initial hypothesis is flawed, the SSCV CoT process empowers AirIAD to explicitly reject its own erroneous assumptions, refocus on the correct region, and revise its prediction; and (iii) **Progressive Refinement:** AirIAD leverages multi-turn exploration to systematically eliminate false positives, output tighter bounding boxes, and refine its semantic diagnosis before finalizing a conclusion.

To transparently illustrate AirIAD's iterative refinement process, we employ a unified color-coding scheme for both semantic reasoning and visual localization. In the text, red highlights initial hallucinations, orange denotes accurate but unverified intermediate hypotheses that require tool confirmation, and green indicates verified definitive diagnoses. Correspondingly, for visual bounding boxes, **black** represents the Ground Truth (GT), while model predictions are colored red (IoU < 0.3), orange ($0.3 \leq$ IoU ≤ 0.5), or green (IoU > 0.5).

E.1 Comparative Analysis against Proprietary Models

Figure 6 demonstrates AirIAD's advantage in resolving semantic ambiguity through iterative cross-verification. While commercial baselines misidentify the defect as Dent under the ambiguous global view, AirIAD proactively doubts the vague initial evidence and invokes dual tools to cross-verify local textures against expert semantic priors via SSCV CoT. This evidence-driven alignment enables AirIAD to recognize the actual material rupture, successfully correcting the diagnosis to Tear and outputting a correct bounding box under the IoU > 0.5 criterion.

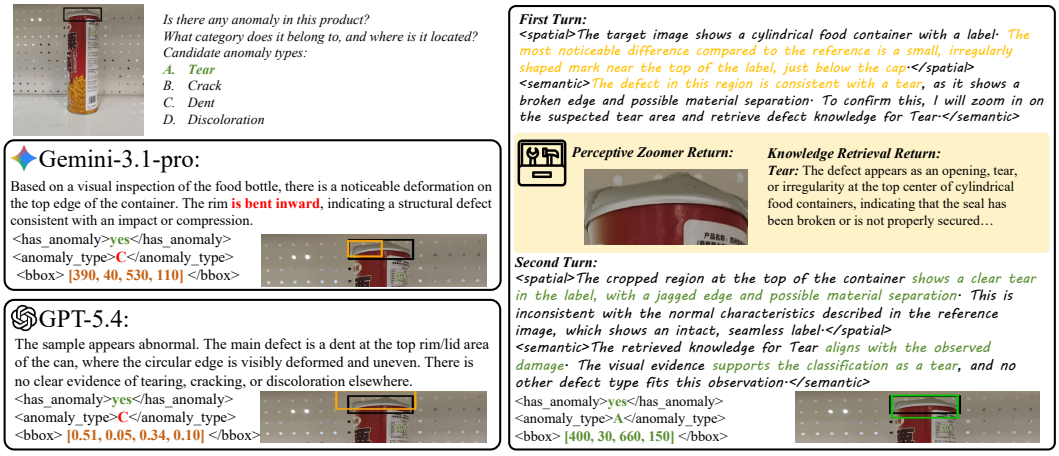


Figure 6: **Qualitative comparison on a challenging case.** While commercial baselines (Gemini 3.1 Pro, GPT 5.4) hallucinate the defect as **Dent** based on a blurred global view, AirIAD utilizes spatio-semantic cross-verification to correctly identify and localize the **Tear** with high precision.

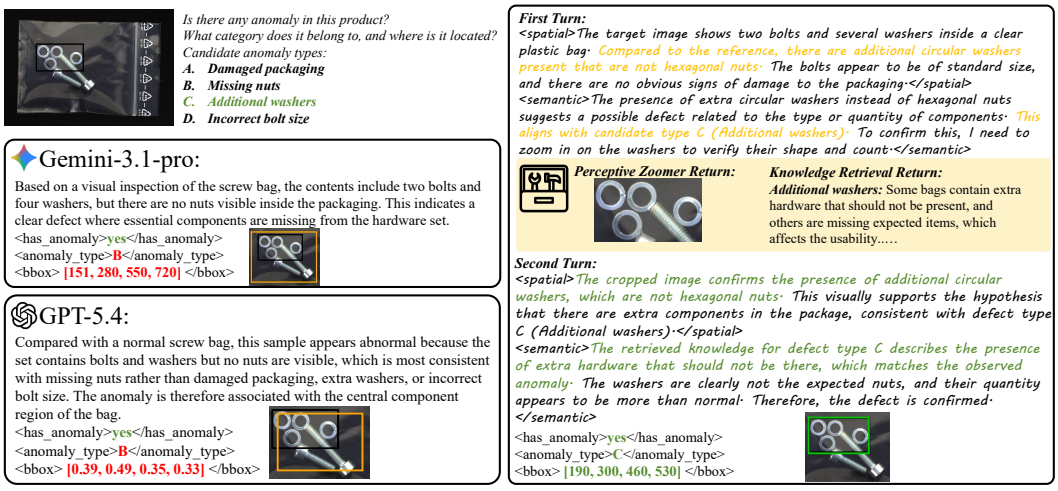


Figure 7: **Qualitative comparison on a logical anomaly case.** While baselines misclassify the screw bag defect as Missing Nuts due to insufficient domain knowledge, AirIAD actively retrieves product specifications to correctly identify the Additional Washers, demonstrating the necessity of interactive knowledge integration.

Figure 7 demonstrates AirIAD's advantage in resolving logical anomalies through interactive knowledge retrieval. For complex industrial scenarios like verifying the contents of a screw bag, relying solely on visual features and internal model weights often leads to misjudgments, as it is difficult to determine correctness without explicit reference standards. Consequently, the baseline models hallucinate the defect as missing nuts. AirIAD overcomes this limitation by actively querying external product specifications. By retrieving and reading the exact manual requirements, the agent cross-verifies the visual component counts against professional domain knowledge, correctly identifying the anomaly as additional washers. This highlights AirIAD's unique capability to transcend pure visual perception by effectively grounding its diagnosis in the retrieved instructional context.

Figure 8 and Figure 9 demonstrate AirIAD's superior capability in detecting extremely subtle anomalies. When confronted with minute defects, baselines exhibit poor classification and localization performance, primarily because granular details are severely diluted in the compressed global view. Conversely, AirIAD overcomes this bottleneck by invoking its Perceptive Zoomer. This tool actively crops and magnifies suspicious regions, fundamentally enhancing the visual perception of local textures. This targeted spatial refinement allows AirIAD, despite its smaller model scale, to capture fine-grained visual features that larger models easily overlook. Consequently, the agent achieves

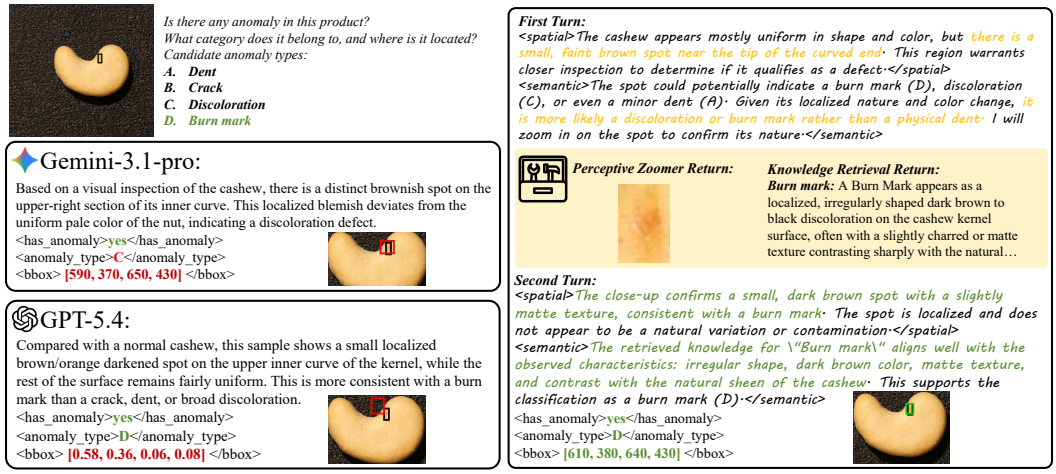


Figure 8: **Qualitative comparison on a subtle anomaly case.** While baselines struggle to locate or classify the minute defect, often outputting low-quality predictions (red), AirIAD leverages its Perceptive Zoomer to enhance local visual clarity, enabling highly precise localization (dark green).

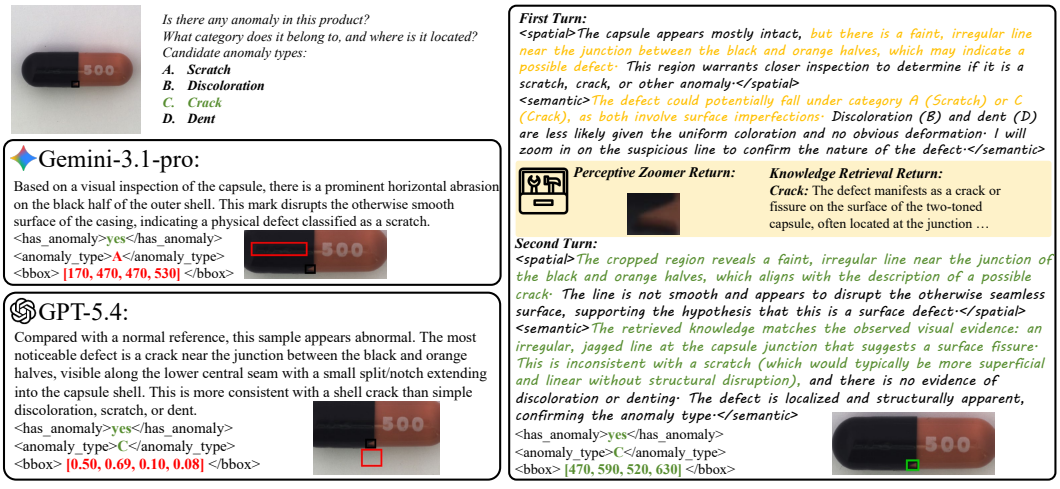


Figure 9: **Qualitative comparison on a subtle anomaly case.** While baselines struggle to locate or classify the minute defect, often outputting low-quality predictions (red), AirIAD leverages its Perceptive Zoomer to enhance local visual clarity, enabling highly precise localization (dark green).

highly accurate categorical diagnosis and precise spatial bounding boxes for barely visible defects, demonstrating the effectiveness of active visual zooming.

Figure 10 and Figure 11 demonstrate AirIAD's advantage in mitigating diagnostic bias on normal samples. When tasked with anomaly detection, baselines often exhibit a strong tendency to over-diagnose, hallucinating nonexistent defects on normal surfaces due to their inherent bias towards finding faults. While AirIAD also generates an initial suspicion regarding a visually ambiguous region, it does not blindly commit to a false positive. Instead, the agent invokes its dual tools to acquire local crops and retrieve standard product specifications. Through the SSCV CoT process, AirIAD rigorously cross-verifies these granular visual features against the retrieved reference standards, confirming that the observed textures are benign manufacturing variations rather than actual defects. Consequently, AirIAD successfully dismisses its initial doubts and correctly classifies the sample as normal, proving its robustness and reliability in preventing false alarms.

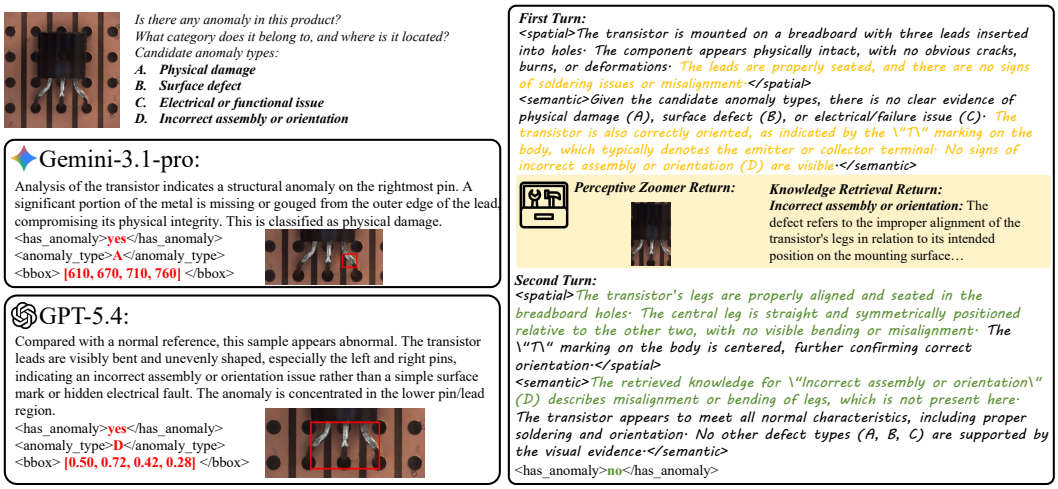


Figure 10: **Qualitative comparison on a normal sample.** Driven by diagnostic bias, commercial baselines hallucinate nonexistent defects (False Positive). Conversely, AirIAD leverages dual-tool cross-verification to resolve its initial suspicion, successfully confirming the sample as Normal and effectively mitigating false alarms.

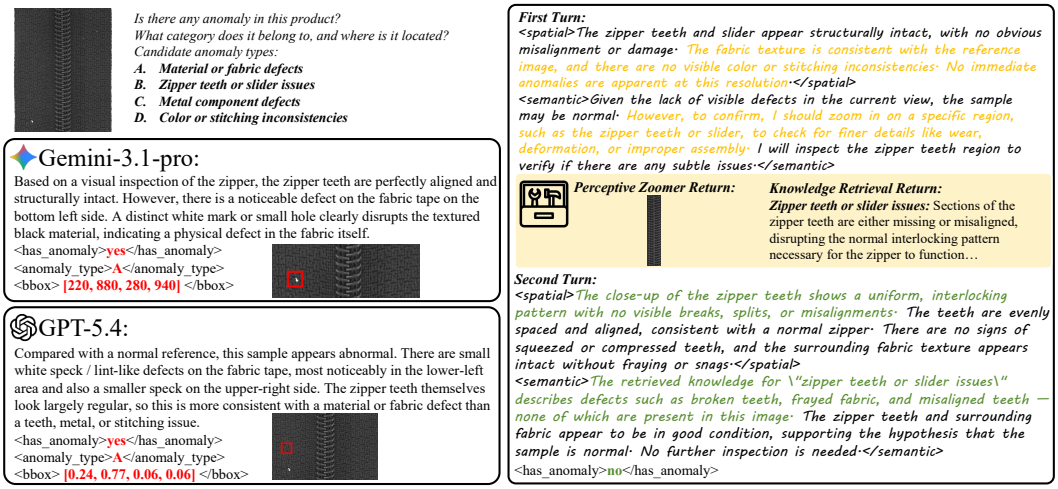


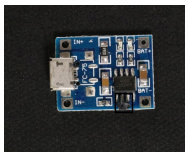
Figure 11: **Qualitative comparison on a normal sample.** Driven by diagnostic bias, commercial baselines hallucinate nonexistent defects (False Positive). Conversely, AirIAD leverages dual-tool cross-verification to resolve its initial suspicion, successfully confirming the sample as Normal and effectively mitigating false alarms.

E.2 Trajectory of Iterative Self-refinement

Figure 12 demonstrates AirIAD's capacity for multi-turn self-correction when its initial judgments are entirely incorrect. In the first round, the agent generates a flawed hypothesis, proposing both an incorrect category and an inaccurate bounding box. However, instead of blindly committing to this error, AirIAD utilizes the retrieved tool observations in the second round to cross-verify its assumptions. Recognizing a distinct mismatch between the visual crop and the semantic prior via SSCV CoT, the agent explicitly rejects its initial diagnosis. It then formulates a new hypothesis and re-invokes the dual tools to gather supporting evidence. In the final round, the newly acquired observations align well with the updated hypothesis, prompting the agent to confidently output the correct diagnosis and precise spatial coordinates. This extended reasoning trajectory highlights the robustness of the iterative framework in recovering from severe initial hallucinations.

Figure 13 further illustrates AirIAD's multi-turn self-correction capability under partial initial success. In the first round, the agent successfully localizes the suspicious region but misclassifies the defect type. Rather than terminating prematurely, AirIAD invokes the dual tools to gather concrete evidence. In the second round, SSCV CoT cross-verifies the visual crop with the retrieved semantic prior, revealing a clear categorical mismatch. The agent then rejects its initial judgment, proposes the correct diagnosis, and re-invokes the tools for validation. By the final round, the new observations confirm the revised hypothesis, enabling a highly accurate joint prediction. This shows that AirIAD can selectively correct semantic errors while retaining useful spatial priors.

Figure 14 demonstrates AirIAD's capability in multi-hypothesis exploration and elimination. In the initial round, the agent identifies two distinct suspicious regions and formulates a separate categorical hypothesis for each. Rather than guessing blindly or aggregating the uncertainties,



Is there any anomaly in this product?
What category does it belong to, and where is it located?
Candidate anomaly types:
A. Loose components
B. Scratches
C. Burn mark
D. Discoloration

First Turn:
<spatial>The USB connector on the left side of the PCB appears slightly misaligned and has visible wear or possible damage at its edges.</spatial>
<semantic>The suspicious region near the USB connector and the dark object near the IC could suggest loose components or physical damage.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Loose components: Loose component appears as a component that is visibly shifted from its designated …

Second Turn:
<spatial>The close-up shows no obvious signs of loose components. There is also a small, dark, irregularly shaped object near the bottom of the IC package.</spatial>
<semantic>The retrieved knowledge for "Loose components" does not match the visual evidence. I should now examine the dark object near the IC to determine if it represents a different type of defect, such as a burn mark or discoloration.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Burn mark: Burn mark is characterized by discoloration typically manifested as a darker area on the surface…

Third Turn:
<spatial>The close-up of the dark object near the IC shows a burnt area, which is inconsistent with normal manufacturing. <spatial>
<semantic>The retrieved knowledge for "Burn mark" aligns with the visual evidence. The dark region near the IC is consistent with a burn mark, which could indicate overheating. This supports the conclusion that the PCB has a defect.</semantic>
<has_anomaly>yes</has_anomaly>
<anomaly_type>C</anomaly_type>
<bbox> [620, 650, 700, 720] </bbox>

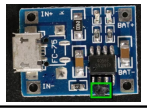
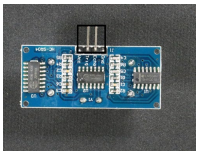


Figure 12: **Qualitative visualization of multi-turn self-correction.** In Round 1, AirIAD generates an erroneous initial hypothesis. In Round 2, through tool invocation and cross-verification, it rejects its own assumption, proposes a corrected hypothesis, and re-invokes the tools. By Round 3, further verification confirms the new diagnosis, yielding a highly accurate final prediction.



Is there any anomaly in this product?
What category does it belong to, and where is it located?
Candidate anomaly types:
A. Damaged components
B. Missing components
C. Incorrect components
D. Misaligned components

First Turn:
<spatial>The PCB appears mostly populated with surface-mount components including ICs and resistors· There is a noticeable gap near the top center where two pins protrude from the board, but no component is visibly soldered or mounted in that area· This region warrants closer inspection to determine if it's a missing component or a misalignment·</spatial>
<semantic>The absence of a component in a designated pad could indicate a missing component (B), while the exposed pins might suggest a misalignment (D) or improper soldering· To confirm, I need to zoom in on the top-center region to examine the solder joints and component placement more closely·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Misaligned components: Misalignment appears as a component (e.g., resistor, capacitor, or IC) shifted horizontally, vertically…

Second Turn:
<spatial>The cropped region shows two vertical pins protruding from the PCB, with no visible component mounted at the corresponding solder pads· The pins appear to be exposed and not soldered to any component, which suggests a possible missing component or a misalignment issue· However, the pins are not visibly overhanging or bridging, which would be more indicative of misalignment (D)·</spatial>
<semantic>The retrieved knowledge for misaligned components (D) emphasizes positional deviation with possible overhanging leads or bridging risks, which are not evident here· The absence of a component at the pad is more consistent with a missing component (B)· To confirm, I need to verify if the pins are meant to be exposed or if a component is expected to be present in that location·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Missing components: A missing component defect appears as an empty solder pad or footprint on the PCB surface where…

Third Turn:
<spatial>The cropped region confirms the absence of a component at the solder pads, with no visible leads, body, or solder joints· The pads are clean and unpopulated, consistent with a missing component defect· The silkscreen reference designators (e·g·, "R12", "C5") nearby further support this observation, as they indicate the expected component locations·</spatial>
<semantic>The retrieved knowledge for missing components (B) accurately describes the visual characteristics observed: clean, unpopulated pads with visible silkscreen outlines and no physical component present· This rules out misalignment (D) and other defect types· The defect is localized to the top-center region of the PCB·</semantic>

<has_anomaly>**yes**</has_anomaly>
<anomaly_type>**B**</anomaly_type>
<bbox> **[370, 140, 540, 350]** </bbox>

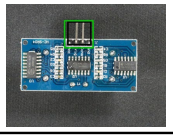


Figure 13: **Qualitative visualization of multi-turn self-correction.** In Round 1, AirIAD accurately localizes the anomaly but generates an incorrect categorical hypothesis. In Round 2, through tool invocation and cross-verification, it recognizes the semantic mismatch, corrects the category, and re-invokes the tools. In the final round, further verification confirms the updated diagnosis, yielding the correct final prediction.

AirIAD systematically isolates and verifies each candidate. Over the subsequent rounds, it invokes the dual tools to gather localized evidence and retrieved semantic definitions. Through the SSCV CoT process, the agent cross-verifies these inputs, explicitly recognizing that one candidate lacks semantic alignment and successfully rejecting it as a false positive. Concurrently, it confirms the supporting evidence for the second region, yielding the correct diagnosis and precise spatial localization. This highlights the framework's robustness in managing complex scenarios with multiple distractors by methodically filtering erroneous assumptions.

Figure 15 illustrates AirIAD's capacity for exhaustive verification and the prevention of premature termination. Initially, the agent successfully localizes the anomalous region but formulates an incorrect categorical hypothesis. Following the first round of tool invocation, the agent utilizes SSCV CoT to cross-verify the visual evidence against semantic priors, successfully recognizing the mismatch and correcting the defect category. However, rather than immediately concluding the



Is there any anomaly in this product?
What category does it belong to, and where is it located?
Candidate anomaly types:
A. Incorrect Label
B. Surface Damage
C. Leakage
D. Deformation

First Turn:
<spatial>The target image shows a cup noodle package with a slightly crumpled or damaged seal at the top edge. This area is not fully sealed, which may indicate a potential leakage risk or surface damage. The rest of the package appears intact, with no obvious deformations.</spatial>
<semantic>The suspicious region at the top of the cup could correspond to either surface damage or a possible leakage. Since the seal is compromised, it is more likely related to leakage rather than just surface damage. I will zoom in on the top edge to confirm whether there is any liquid residue or further damage.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Leakage: Leakage appears as wet, shiny, or discolored patches—often brownish, yellowish, or translucent…

Second Turn:
<spatial>The close-up of the top edge shows a slightly crumpled and uneven seal, but no visible wetness, stains, or residue that would indicate leakage. The material appears intact without obvious signs of liquid seepage. This suggests that while there may be some surface damage, it does not yet meet the criteria for a leakage defect.</spatial>
<semantic>The retrieved knowledge for leakage describes wet, discolored, or sticky patches, which are not present here. The observed damage is more consistent with surface damage, such as a compromised seal or minor tearing. I will now zoom in on the lower part of the label to check for any other potential defects or damage.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Surface Damage: The defect present is characterized by irregularities on the surface of various cylindrical food containers…

Third Turn:
<spatial>The close-up reveals a small, irregular white spot on the green surface, which appears to be a physical defect such as a scratch, abrasion, or minor tear. This is inconsistent with the normal characteristics described, where labels are smooth and free from such imperfections. The defect is localized and does not show signs of leakage or deformation.</spatial>
<semantic>The retrieved knowledge for surface damage aligns with the observed defect: a noticeable irregularity on the surface that disrupts the expected smoothness. There is no evidence of leakage, deformation, or incorrect labeling. The defect is confined to the surface and does not affect the structural integrity of the package. I can now conclude that the anomaly type is surface damage.</semantic>



<has_anomaly>**yes**</has_anomaly>
<anomaly_type>**B**</anomaly_type>
<bbox> **[470, 590, 520, 650]** </bbox>

Figure 14: **Qualitative visualization of multi-hypothesis elimination.** In Round 1, AirIAD generates two distinct anomaly hypotheses for different spatial regions. Through sequential tool invocations and cross-verification, it systematically rejects the false positive while confirming the true defect, ultimately yielding the correct final prediction.

inspection upon confirming this single region, AirIAD exhibits a highly rigorous diagnostic behavior. It re-invokes the perceptive tools to conduct a secondary inspection of other potentially suspicious areas across the global view. After ensuring no additional anomalies are present and re-validating its corrected hypothesis, the agent finally outputs the accurate joint prediction. This extended reasoning trajectory demonstrates that the iterative framework not only facilitates targeted self-correction but also promotes comprehensive scene exploration, effectively minimizing the risk of missed detections in complex industrial environments.

Figure 16 and Figure 17 demonstrate AirIAD's robust reasoning capabilities under high ambiguity. In both complex scenarios, the initial visual evidence is highly ambiguous, prompting the agent to engage in an extensive multi-turn exploration process. Rather than settling for a low-confidence prediction or prematurely terminating the inspection, AirIAD executes three consecutive rounds of tool invocations. Throughout this deep iterative process, the agent continuously adjusts both its spatial bounding boxes and categorical hypotheses. By rigorously applying SSCV CoT at each step, it cross-verifies the newly acquired crops against retrieved semantic priors, systematically eliminating



Is there any anomaly in this product?
What category does it belong to, and where is it located?
Candidate anomaly types:
A. Neck Damage
B. Shiny Spot
C. Leakage
D. Surface Damage

First Turn:
<spatial>The bottle appears mostly intact, but there is a small, dark, irregularly shaped spot near the lower right corner of the label. This region warrants closer inspection to determine if it is a defect.</spatial>
<semantic>The spot could potentially indicate a shiny spot, surface damage, or even a minor leakage. Given its location and appearance, it is most likely a surface defect or a shiny spot. I will zoom in on the spot to confirm its nature.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Shiny Spot: A shiny spot appears as an abnormally bright, highly reflective circular or oval region on the bottle surface…

Second Turn:
<spatial>The cropped region shows a dark, irregularly shaped spot on the label, which does not exhibit the characteristics of a shiny spot as described. The spot lacks the bright, reflective, mirror-like sheen typical of a shiny spot and instead appears more like a physical defect or contamination.</spatial>
<semantic>The retrieved knowledge for \"Shiny Spot\" does not match the observed defect. The spot is not reflective or bright; it is dark and irregular, suggesting it may be a surface damage or possibly a leakage. I will now zoom in further to examine the texture and edges of the spot to determine if it is a surface defect.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Surface Damage: The defect manifests as irregularities or imperfections on the surface of plastic or glass bottles…

Third Turn:
<spatial>The cropped region still shows a dark, irregularly shaped spot on the label, but the texture and edges suggest a possible surface defect rather than a shiny spot. The spot does not exhibit the reflective properties of a shiny spot and instead appears to be a physical imperfection on the surface.</spatial>
<semantic>The retrieved knowledge for \"Surface Damage\" aligns better with the observed defect, as it describes irregularities, discoloration, and potential smudging that disrupt the visual continuity of the product. However, to confirm whether this is a surface damage or another type of defect, I will now check for any signs of leakage by examining the bottle's neck and cap area, as leakage might cause additional visual cues such as wetness or residue.</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Leakage: Leakage appears as wet, shiny, or discolored patches—often darker or stained—on the bottle's bottom…

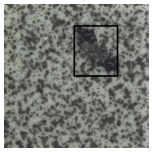
Fourth Turn:
<spatial>The neck and cap area of the bottle appear intact with no visible signs of wetness, residue, or leakage. The seal is properly engaged, and there are no indications of liquid seeping out, which rules out leakage as the defect type.</spatial>
<semantic>The defect observed on the label is consistent with surface damage, as it involves an irregular, dark spot that disrupts the visual continuity of the label. The lack of evidence for leakage and the alignment with surface damage support this conclusion. No other defect types are supported by the visual evidence.</semantic>

<has_anomaly>yes</has_anomaly>
<anomaly_type>D</anomaly_type>
<bbox> [620, 600, 650, 670] </bbox>



Figure 15: **Qualitative visualization of exhaustive verification.** AirIAD initially localizes the anomaly but misclassifies it. After correcting the category through a first round of tool invocation, the agent proactively conducts a secondary inspection of other regions before finalizing its prediction, demonstrating a highly rigorous and comprehensive reasoning process.

erroneous assumptions and refining its focus. This continuous bidirectional refinement gradually resolves the extreme initial ambiguity, enabling the agent to finally converge on the correct joint prediction. These cases highlight the critical necessity of deep iteration and repeated spatio-semantic alignment when confronting highly challenging industrial anomalies.



Is there any anomaly in this product?
What category does it belong to, and where is it located?
Candidate anomaly types:
A. A grey stroke
B. A scratch
C. A crack
D. A chip

First Turn:
<spatial>The target image shows a speckled surface similar to the reference, but there is a distinct dark, irregularly shaped region near the center-right that is not present in the reference· This region appears to be a concentrated area of darker material, possibly indicating a defect·</spatial>
<semantic>Given the irregular, somewhat smeared appearance and the concentration of dark material, it is more likely a scratch or a chip rather than a crack or a simple grey stroke· I will zoom in on the suspicious region to confirm the nature of the defect·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
A scratch : A scratch appears as a thin, linear, elongated defect with irregular edges, typically lighter or darker…

Second Turn:
<spatial>The cropped region shows a dark, irregularly shaped area with a somewhat smeared and uneven texture, which is inconsistent with the uniform speckled pattern seen in the reference image· This suggests a possible defect, but it is not clearly a chip, as there is no obvious sharp edge or concave depression· It also does not resemble a scratch, which would typically be linear and shallow· The defect could still be a chip, but the current evidence is inconclusive·</spatial>
<semantic>The retrieved knowledge for \"chip\" describes a defect with sharp edges and material loss, which is not clearly visible here· The defect might still be a chip, but the current evidence does not fully support this classification· I will refine my hypothesis and consider whether the defect could be a scratch or another type· I will zoom in further on the edges of the dark region to better assess its morphology·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
A chip: A chip appears as a small, irregularly shaped missing piece of material, typically with sharp or jagged edges…

Third Turn:
<spatial>The zoomed-in view of the dark region reveals a somewhat smeared and uneven texture, but there are no clear sharp edges or concave depressions that would indicate a chip· The defect does not exhibit the depth, shadowing, or geometric discontinuity characteristic of a chip· This suggests that the defect is not a chip· Given the irregular shape and lack of linear features, it is less likely to be a scratch either· The defect could still be a grey stroke, which would explain the concentrated dark area without implying material loss·</spatial>
<semantic>The retrieved knowledge for \"chip\" does not accurately describe the observed defect, as there are no signs of material loss or geometric discontinuity· The defect does not match the criteria for a chip· Since the defect is not a chip, and it lacks the linear characteristics of a scratch, it is more plausible that it is a grey stroke· I will now verify this hypothesis by checking if the defect aligns with the characteristics of grey stroke·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
A grey stroke: The Gray Stroke Defect is characterized by an irregularly shaped gray area, often appearing…

Fourth Turn:
<spatial>The cropped region clearly shows an irregularly shaped, darker area that contrasts with the surrounding speckled pattern· The edges of this region are uneven and do not follow the uniform distribution of the dark speckles, which is consistent with the description of a grey stroke defect· The defect is concentrated and disrupts the expected uniformity of the tile's surface, further supporting this classification·</spatial>
<semantic>The retrieved knowledge for \"grey stroke\" accurately describes the observed defect: an irregular, darker area with uneven edges that disrupts the speckled pattern· This aligns with the visual evidence and suggests that the defect is due to potential issues in the glazing process or the presence of foreign materials during manufacturing· No other defect type fits this description as well·</semantic>

<has_anomaly>**yes**</has_anomaly>
<anomaly_type>**A**</anomaly_type>
<bbox> **[470, 150, 800, 480]** </bbox>

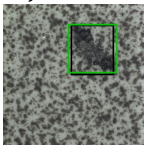


Figure 16: **Qualitative visualization of iterative refinement under high ambiguity.** AirIAD executes three consecutive tool invocations, repeatedly adjusting its spatial localization and categorical hypothesis before converging on the correct final prediction.



Is there any anomaly in this product?
What category does it belong to, and where is it located?
Candidate anomaly types:

A. Damaged packaging
B. Improper arrangement of the items
C. Insufficient quantity of granola
D. Inconsistency in the types of fruits included

First Turn:

<spatial>The target image shows a breakfast box with two peaches and one orange in the left compartment, while the reference image has two oranges· This suggests a possible discrepancy in the quantity or type of fruits· The right compartment appears normal with granola, banana slices, and almonds·</spatial>
<semantic>The anomaly is likely related to the fruit arrangement or quantity· Since the reference has two oranges and the target has two peaches, this points to option B (Improper arrangement of the items)· I will zoom in on the left compartment to confirm the fruit types and count·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Improper arrangement of the items: The primary defects observed across the provided images involve structural anomalies…

Second Turn:

<spatial>The cropped image confirms the presence of two peaches and one orange in the left compartment, which differs from the reference image that contains two oranges· This visual evidence supports the hypothesis that the arrangement of items is improper, as the fruit types and quantities do not match the expected configuration·</spatial>
<semantic>The retrieved knowledge for option B (Improper arrangement of the items) describes structural anomalies related to fruits, such as discoloration or bruising, but does not explicitly mention a mismatch in fruit types or quantities· However, since the defect is not about physical damage but rather a deviation from the expected composition, the knowledge may not fully align· I will now check the granola section to ensure there are no other anomalies, such as insufficient quantity, which could also explain the discrepancy·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Insufficient quantity of granola: The common defects within this batch are primarily categorized as logical anomalies involving…

Third Turn:

<spatial>The granola section appears full and properly arranged, with no signs of insufficient quantity· The dried banana slices and almonds are also present in the expected amounts and positions· This rules out option C (Insufficient quantity of granola) as the primary defect·</spatial>
<semantic>The retrieved knowledge for option C mentions logical anomalies such as missing components or incorrect ratios, which aligns with the observed fruit discrepancy· However, since the granola is not the issue, the defect must be related to the fruit arrangement· The initial hypothesis of option B (Improper arrangement of the items) remains plausible, as the fruit types and quantities differ from the reference· I will now re-examine the fruit compartment to confirm whether the arrangement is indeed improper or if there are other factors at play·</semantic>

Perceptive Zoomer Return:



Knowledge Retrieval Return:
Inconsistency in the types of fruits included: These anomalies often manifest as missing components, incorrect ratios of food items…

Fourth Turn:

<spatial>The cropped image confirms the presence of two peaches and one orange in the left compartment, which contrasts with the reference image that contains two oranges· This visual evidence supports the hypothesis that there is an inconsistency in the types of fruits included, as the fruit composition does not match the expected configuration·</spatial>
<semantic>The retrieved knowledge describes logical anomalies involving discrepancies in the expected content or arrangement of food items, which aligns with the observed fruit discrepancy· The defect is not about physical damage but rather a deviation from the expected composition, which is consistent with the description provided· Since the granola and other items appear normal, and the fruit arrangement is clearly improper, the defect type is confirmed·</semantic>

<has_anomaly>**yes**</has_anomaly
<anomaly_type>**D**</anomaly_type>
<bbox> **[100, 60, 450, 860]** </bbox>



Figure 17: **Qualitative visualization of iterative refinement under high ambiguity.** Similar to the previous case, the agent continuously refines its spatial and semantic judgments over three rounds of cross-verification to resolve high initial ambiguity and output an accurate diagnosis.

F Broader Impacts and Ethics

As a framework for multimodal agentic reasoning in anomaly detection, AirIAD primarily targets industrial manufacturing and does not suffer from direct ethical concerns regarding data privacy. All datasets used in our experiments are publicly available, and all qualitative visualizations are based on industrial product images, which do not infringe on personal privacy. From a positive societal perspective, the deployment of robust anomaly detection systems like AirIAD can significantly enhance manufacturing efficiency, reduce industrial waste, and alleviate the cognitive fatigue of human inspectors. Furthermore, its generalizability holds promise for democratizing high-quality diagnostic assistance in fields like healthcare. However, considering AirIAD's strong cross-domain generalization potential (e.g., to medical imaging), deploying such automated systems in safety-critical domains introduces the risk of automation bias, where human experts might over-rely on MLLM-generated diagnostics. To mitigate potential harms from incorrect predictions, we advocate deploying our framework as a transparent, "human-in-the-loop" assistive tool to augment, rather than replace, expert decision-making.

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 of the paper clearly outline the main contributions and the scope of the research.

   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: The paper adequately discusses the limitations of the work in Appendix D.

   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 proposes a novel agentic framework (AirIAD) and a reinforcement learning optimization paradigm (SS-GiGPO) for industrial anomaly detection. While we rigorously formulate the optimization objectives and hierarchical advantage estimation mathematically, the core contributions are methodological and are validated through extensive empirical evaluations and ablation studies rather than formal theoretical 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: The paper provides a clear and comprehensive explanation of the proposed method. Key setups critical for reproducing the results are meticulously detailed in Section 4.1 and Appendix B.

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: The complete source code, dataset, pre-trained model weights, and detailed instructions for reproducing our main experimental results are publicly available at `https://github.com/FisherYuuri/AirIAD`.

   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: The experiment setting and details are presented in Appendix B.

   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: [Yes]

   Justification: In our main results and the detailed multi-seed reproducibility analysis in Appendix C.4, we report the sample mean and standard deviation across three independent runs with distinct random seeds. The error bars capture the variability stemming from training data shuffling and stochastic trajectory exploration during reinforcement learning.

   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: The information on the computer resources can be found in Appendix B.1.

   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: The research conforms to the NeurIPS Code of Ethics.

   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 societal impact is described in Appendix F.

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: Our method is based on original proprietary or open-sourced MLLMs, the procedure does not introduce additional risks. Users are encouraged to adhere to the safeguards provided by the corresponding MLLM.

    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 creators of the used datasets and models in our paper are properly cited. The proposed dataset is constructed based on MMAD and adheres to its license agreement (CC BY-NC-SA 4.0)

    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: [Yes]

    Justification: The code/dataset provided in the supplementary material is documented with detailed instructions.

    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: This paper does not involve crowdsourcing or research with human subjects. The datasets used for evaluation (e.g., MMAD, BTMRI) are publicly available or annotated internally by the authors, which does not fall under human subject research.

    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: The paper does not involve crowdsourcing nor research with human subjects.

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: As a core component of this work, we provide a detailed description of how the proposed method leverages large models.

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.