COMPARATIVE EVIDENCE POLICY OPTIMIZATION FOR MULTIMODAL INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial anomaly detection (IAD) is a critical task for automated quality inspection and intelligent manufacturing. Recent advances in multimodal large language models (MLLMs) have extended IAD beyond conventional anomaly classification and localization toward unified visual reasoning and natural-language explanation. However, our analysis reveals a fundamental evidence-grounding gap in reference-guided anomaly reasoning: MLLMs often fail to reliably ground their predictions in anomaly-relevant regions and underutilize the normal reference. More importantly, weak anomaly grounding and insufficient Query–Reference comparison substantially degrade anomaly reasoning performance. This observation motivates a simple principle: reliable anomaly reasoning should be grounded in defect-relevant visual evidence and explicitly supported by comparison against the corresponding normal reference. To this end, we first introduce Comparative Attention Supervision (CAS), which establishes anomaly-aware visual grounding by encouraging the generated reasoning to attend to defect-relevant Query regions and their corresponding normal Reference context. We further propose Comparative Evidence Policy Optimization (CEPO), which moves beyond outcome-only optimization by explicitly evaluating whether the generated reasoning is grounded in anomaly-relevant Query–Reference comparative evidence. Experiments on MMAD and MMR-AD demonstrate the effectiveness of our approach across diverse anomaly understanding and localization tasks, as well as its generalization to unseen datasets. Further ablations and qualitative analyses highlight the complementary benefits of anomaly-aware grounding and comparative evidence in improving IAD performance.

1 INTRODUCTION

Industrial anomaly detection (IAD) (Bergmann et al., 2019; Cai et al., 2026) is a fundamental component of automated quality inspection and intelligent manufacturing. Reference-guided anomaly reasoning (Jiang et al., 2025; Yao et al., 2026) has emerged as an important setting for multi-modal anomaly understanding: given a Query image I^q and a normal Reference image I^r, the model identifies potential anomalies and explains its decision. Intuitively, this setting is inherently comparative. A reliable model should identify suspicious regions in I^q, associate them with the corresponding normal patterns in I^r, and reason about their discrepancies before making the final decision.

Recent advances in multimodal large language models (MLLMs) have expanded IAD beyond classification and localization toward visual reasoning and explanation. Researchers explore different ways to enhance anomaly understanding: AnomalyGPT (Gu et al., 2024) and OmniAD (Zhao et al., 2026) introduce anomaly-specific visual modeling. AgentIAD (Miao et al., 2026) incorporates external modules for structured anomaly analysis. Reasoning-oriented methods such as AnomalyR1 (Chao et al., 2025) and AD-FM (Liao et al., 2026) further optimize reasoning trajectories. However, they largely focus on improving what the model predicts or generates, leaving the visual evidence underlying its reasoning insufficiently constrained. This gap is particularly critical in reference-guided IAD, where a correct prediction alone does not guarantee that the model has identified the actual defect, let alone compared it against the corresponding region in I^r.

This lack of evidence supervision creates a fundamental bottleneck: without reliably grounding anomaly-relevant visual evidence and meaningful Query–Reference comparisons, MLLMs struggle

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

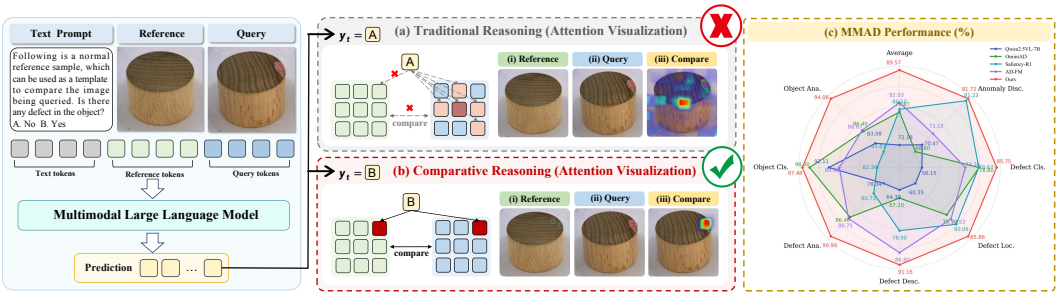


Figure 1: Comparison of reasoning patterns in reference-guided IAD. (a) Traditional reasoning attends to irrelevant Query regions, largely ignores the Reference, and fails to establish meaningful Query–Reference comparison. (b) Comparative evidence reasoning grounds the prediction in anomaly-relevant Query regions, retrieves the corresponding normal Reference evidence, and explicitly compares the two. (c) Performance comparison of existing and our methods on MMAD.

to generate reliable anomaly decisions and reasoning. Existing observations (Jiang et al., 2025) show that providing normal Reference images does not consistently improve performance and can even degrade some models. One research (Gong et al., 2026) points out that MLLMs may underutilize visual information during reasoning. Building on these observations, we reveal two key factors behind this limitation. First, models may fail to ground their reasoning in the actual anomalous regions of I^q, weakening the visual basis for anomaly recognition. Second, even when models identify relevant Query evidence, they may still underutilize the Reference rather than explicitly reason about the discrepancy. As we can see in Fig. 1, stronger anomaly grounding and Query–Reference comparative evidence consistently correspond to better anomaly reasoning performance. These observations motivate a progressive learning paradigm that first establishes reliable anomaly-aware grounding and then encourages the model to reason through explicit Query–Reference comparison.

To address the limitations, we introduce Comparative Attention Supervision (CAS) to learn anomaly-aware visual grounding during supervised fine-tuning (SFT). Rather than relying solely on language generation objectives to determine visual evidence, CAS directly encourages the model's reasoning to attend to defect-relevant regions in I^q and to associate them with the corresponding normal context in I^r. This supervision aligns the reasoning process with visual evidence relevant to the anomaly pattern, and reduces cases where predictions are supported by weak or misplaced visual evidence. Importantly, CAS provides an anomaly-aware grounding initialization for subsequent policy optimization.

To reinforce comparative reasoning, we further introduce Comparative Evidence Policy Optimization (CEPO). Specifically, Comparative Evidence Grounding (CEG) evaluates anomaly-relevant Query–Reference evidence at individual generation positions. These position-level scores are aggregated into a trajectory-level comparative reward and combined with task rewards for policy optimization. Consequently, trajectories that reach the same correct answer can receive different advantages according to the quality of their comparative evidence. CEPO thereby guides policy optimization toward reasoning trajectories grounded in anomaly-relevant visual evidence and its comparison with normal Reference context.

Extensive experiments on two industrial benchmarks, MMAD and MMR-AD, demonstrate the effectiveness of our framework across diverse anomaly reasoning tasks. Besides, numerical analyses show that CAS substantially strengthens anomaly-aware visual grounding, and that CEPO increases the use of meaningful Query–Reference evidence during self-generated reasoning. Comprehensive ablations further validate the respective contributions of CAS and CEPO.

The contributions of this paper are summarized as follows:

- We introduce CAS to establish anomaly-aware visual grounding, and propose CEG to assess whether each reasoning step is supported by meaningful Query–Reference evidence.
- Building on CEG, we propose CEPO to bridge outcome supervision and comparative reasoning, moving policy optimization beyond "what answer is correct" toward "what visual evidence supports the answer" through anomaly-relevant Query–Reference evidence.

- Extensive experiments on MMAD and MMR-AD demonstrate the effectiveness and general applicability of the proposed approach.

2 Related Work

Reference-Guided Industrial Anomaly Detection. Conventional IAD models normal visual patterns through reconstruction, feature embeddings, memory banks, or knowledge distillation (Zavrtanik et al., 2021; Roth et al., 2022; Deng & Li, 2022). Recent MLLMs extend IAD toward anomaly understanding and explanation (Gu et al., 2024; Xu et al., 2025; Li et al., 2025; Zhao et al., 2026). Reference examples and auxiliary memories support retrieval, prototype matching, and cross-instance comparison (Ma et al., 2025; Bhunia et al., 2025; Cai et al., 2026; Xu et al., 2026). JUDO (Kang et al., 2026) juxtaposes Query and normal images for fine-grained inspection, while AD-Copilot (Jiang et al., 2026) uses visual in-context comparison. These approaches establish the value of visual comparison, but do not score attention-based evidence from matched Query and Reference regions at individual generation positions.

Visual Grounding and Evidence Alignment. Visual grounding associates generated language or reasoning with supporting visual evidence. Region-grounded MLLMs connect language with localized regions (Rasheed et al., 2024), while Visual CoT introduces intermediate visual grounding for multimodal reasoning (Shao et al., 2024a). These examples primarily ground reasoning within an individual image. Reference-guided anomaly reasoning additionally requires relating anomalous Query regions to their normal Reference counterparts. Our CAS supports anomaly-aware attention, while CEG scores cross-image evidence during generation.

Policy Optimization for MLLMs. GRPO provides a general framework for group-relative policy optimization (Shao et al., 2024b). PAPO strengthens visual dependence (Wang et al., 2026), while Saliency-R1 rewards the alignment of token-level visual saliency with annotated regions (Gong et al., 2026). In IAD, JUDO, IAD-R1, and AD-FM incorporate tailored or fine-grained rewards for anomaly reasoning (Kang et al., 2026; Li et al., 2026; Liao et al., 2026). Our method further scores matched Query–Reference evidence at generation positions and aggregates these signals into a trajectory-level reward.

3 Method

3.1 Problem Formulation and Overview

Given a normal Reference image I^r, a Query image I^q and a textual instruction x, reference-guided anomaly reasoning generates a response $y = (y_1, \ldots, y_T)$ that identifies whether I^q is anomalous, localizes the defect when present, and explains the decision. Given $\mathcal{X} = (I^r, I^q, x)$, MLLM parameterized by θ generates $p_\theta(y \mid \mathcal{X}) = \prod_{t=1}^{T} p_\theta\left(y_t \mid I^r, I^q, x, y_{<t}\right)$. Our goal is to refine this generation to exploit visual evidence better for anomaly reasoning.

As illustrated in Fig. 2, our framework contains two stages. First, we introduce Comparative Attention Supervision (CAS) during supervised fine-tuning. CAS directly regularizes the model's visual attention so that its reasoning is grounded in the anomalous Query region while retaining the corresponding normal context from the Reference. Second, we introduce Comparative Evidence Policy Optimization (CEPO). Besides rewarding a trajectory on its final answer, CEPO also evaluates the Query–Reference evidence used during reasoning. Specifically, we construct Comparative Evidence Grounding (CEG) to identify Query regions that differ from their matched Reference patterns and to trace whether the reasoning step actually accesses such cross-image evidence. Next, the resulting comparative evidence is transferred into a trajectory-level reward for policy optimization.

Overall, CAS establishes where the model should ground its reasoning, while CEPO encourages the model to make effective use of Query–Reference evidence during generation.

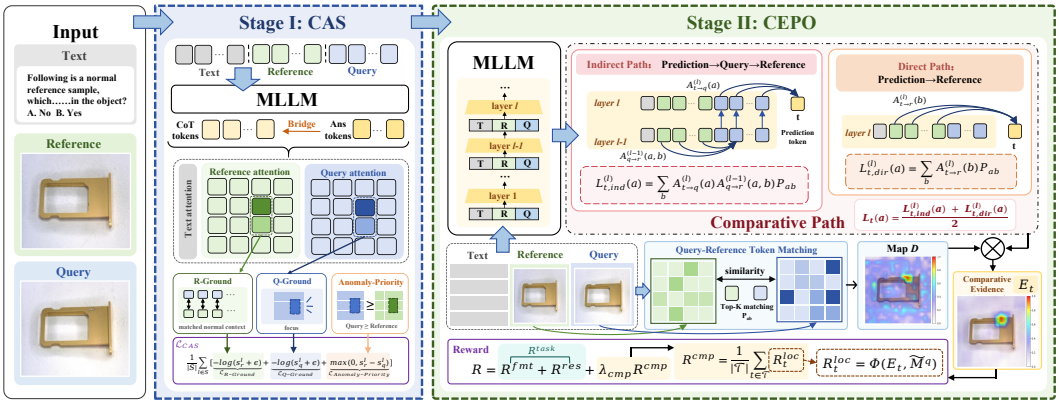


Figure 2: Overview of the proposed CEPO framework. Stage I: CAS builds anomaly-aware grounding by guiding the model to attend to anomalous Query regions and corresponding Reference regions. Stage II: CEPO further optimizes reasoning with comparative evidence rewards, encouraging predictions supported by proper Query–Reference comparison.

3.2 Comparative Attention Supervision

In this section, we introduce CAS, which uses anomaly masks to explicitly shape visual attention of the generated reasoning. The basic idea is intuitive: reasoning tokens should primarily attend to the anomalous region in I^q and its corresponding normal region in I^r.

From Anomaly Masks to Visual Tokens. For an anomalous Query, let $M^q \in \{0,1\}^{H\times W}$ denote its pixel-level anomaly mask. The Query and Reference are represented by visual-token grids of sizes $H_q \times W_q$ and $H_r \times W_r$, respectively. Under an approximate alignment of object position and layout, we use shared relative spatial coordinates to project the Query anomaly mask onto both visual-token grids via adaptive max pooling:

$$\widetilde{M}^q = \mathrm{AMP}_{H_q\times W_q}(M^q), \qquad \widetilde{M}^r = \mathrm{AMP}_{H_r\times W_r}(M^q). \tag{1}$$

Max pooling preserves the support of small defects at the visual-token resolution. Here, $\widetilde{M}^q$ identifies anomalous Query tokens, while $\widetilde{M}^r$ provides a coarse spatial prior for normal Reference context. This prior is most appropriate for image pairs with similar object layouts.

Reasoning-to-Image Attention. We next measure which visual tokens support the generated reasoning. Let $A^{l,h}_{t,j}$ denote the attention from generated token t to a preceding visual token j at decoder layer l and head h. Since we are interested in the relative use of the two images, we normalize the attention over all Query and Reference visual tokens:

$$\widehat{A}^{l,h}_{t,j} = \frac{A^{l,h}_{t,j}}{\sum_{k\in\mathcal{V}^q\cup\mathcal{V}^r} A^{l,h}_{t,k} + \epsilon}, \qquad j \in \mathcal{V}^q \cup \mathcal{V}^r. \tag{2}$$

We aggregate the attention of the generation positions $\mathcal{T}_e$ that carry reasoning evidence. For layer l, the resulting Query and Reference grounding maps are:

$$\bar{A}^{q,l}_j = \frac{1}{N_h\,|\mathcal{T}_e|}\sum_{h=1}^{N_h}\sum_{t\in\mathcal{T}_e}\widehat{A}^{l,h}_{t,j}, \qquad \bar{A}^{r,l}_k = \frac{1}{N_h\,|\mathcal{T}_e|}\sum_{h=1}^{N_h}\sum_{t\in\mathcal{T}_e}\widehat{A}^{l,h}_{t,k}. \tag{3}$$

These maps describe where the generated reasoning draws visual information from in the Query and Reference, respectively.

Anomaly-Aware Attention Supervision. Given the grounding maps, we measure how much attention falls inside the target regions:

$$s^l_q = \frac{\sum_{j\in\mathcal{V}^q}\bar{A}^{q,l}_j\widetilde{M}^q_j}{\sum_{j\in\mathcal{V}^q}\bar{A}^{q,l}_j + \epsilon}, \qquad s^l_r = \frac{\sum_{j\in\mathcal{V}^r}\bar{A}^{r,l}_j\widetilde{M}^r_j}{\sum_{j\in\mathcal{V}^r}\bar{A}^{r,l}_j + \epsilon}. \tag{4}$$

Here, s_q^l measures how strongly the reasoning is grounded in the actual anomaly, while s_r^l measures its access to the corresponding normal Reference context.

We encourage both scores to be high:

$$\mathcal{L}_{\text{CAS}} = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \left[\underbrace{-\log(s_q^l + \epsilon)}_{\mathcal{L}_{\text{Q-Ground}}} + \underbrace{-\log(s_r^l + \epsilon)}_{\mathcal{L}_{\text{R-Ground}}} + \underbrace{\max(0, s_r^l - s_q^l)}_{\mathcal{L}_{\text{Anomaly-Priority}}} \right]. \tag{5}$$

where $\mathcal{S}$ denotes the supervised decoder layers. The first two terms encourage the reasoning to access the relevant regions in both images. $\mathcal{L}_{\text{Anomaly-Priority}}$ prevents Reference attention from dominating Query attention, since the anomaly itself is observed in the Query and the Reference serves as normality evidence for comparison.

Connecting Grounded Reasoning to the Final Decision. Grounding the reasoning is useful only if the final prediction actually depends on it. Thus, we propose a lightweight bridge loss between the generated reasoning and answer tokens. Let $\mathcal{T}_{\text{cot}}$ and $\mathcal{T}_a$ denote the reasoning and answer positions. For each answer token a, we compute its attention to the preceding reasoning b_a and $\mathcal{L}_{\text{bridge}}$:

$$\mathcal{L}_{\text{bridge}} = -\frac{1}{|\mathcal{T}_a|} \sum_{a \in \mathcal{T}_a} \log(b_a + \epsilon), \quad b_a = \frac{1}{|\mathcal{S}| N_h} \sum_{l \in \mathcal{S}} \sum_{h=1}^{N_h} \sum_{t \in \mathcal{T}_{\text{cot}}} A_{a,t}^{l,h}. \tag{6}$$

With $\mathcal{L}_{\text{bridge}}$, we encourage the final answer to retain information from the visually grounded reasoning. Overall, Eq 7 is the loss function in SFT. $\gamma(\tau)$ linearly increases from 0 to 1 during the initial training phase. CAS introduces no additional learnable parameters. The mask-dependent CAS loss is applied only to anomalous queries with valid spatial annotations. For normal queries, we set $\mathcal{L}_{\text{CAS}} = 0$, while retaining their cross-entropy and bridge supervision.

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \gamma(\tau) \left(\lambda_{\text{CAS}} \mathcal{L}_{\text{CAS}} + \lambda_b \mathcal{L}_{\text{bridge}} \right). \tag{7}$$

3.3 Comparative Evidence Grounding

For policy optimization, we further need to determine whether a self-generated trajectory actually uses meaningful Query–Reference comparison. We characterize such comparative evidence through two questions: *(i)* "does a Query region differ from its corresponding normal pattern in I^r?", and *(ii)* "does the generated reasoning access this cross-image relation? ". CEG answers the first question with a discrepancy map and the second with a comparison-path map.

Query–Reference Discrepancy. We first estimate feature-based correspondences between the two images, allowing each Query token to match Reference tokens across spatial locations. Let q_a and r_b denote the representations of Query token a and Reference token b. Their similarity is

$$P_{ab} = \text{softmax}_b \left(\frac{\cos(q_a, r_b)}{\tau_{\text{corr}}} \right), q_a \in \mathcal{V}^q, r_b \in \mathcal{V}^r. \tag{8}$$

For robustness, we retain the top-K_{corr} Reference matches for each Query token and renormalize them. Based on P_{ab}, we reconstruct the normal Reference counterpart of Query token a:$\widehat{r}_a = \text{norm}\left(\sum_b P_{ab} r_b\right)$. The discrepancy between the Query token and its normal counterpart is then: $D_a = (1 - \cos(q_a, \widehat{r}_a))/2$. Hence, D_a is high when the local Query appearance cannot be well explained by its matched normal Reference evidence.

Tracing Comparative Access. A large discrepancy alone does not imply that the generated reasoning actually uses it. We therefore trace how each generated token accesses the matched Query–Reference evidence.

Let $A_{t \to q}^{(l)}(a)$ and $A_{t \to r}^{(l)}(b)$ denote the attention from generated token t to Query token a and Reference token b, respectively. We consider two ways in which Reference evidence can contribute to generation indirectly and directly:

$$L_{t,\text{ind}}^{(l)}(a) = \sum_b A_{t \to q}^{(l)}(a) A_{q \to r}^{(l-1)}(a, b) P_{ab}, \quad L_{t,\text{dir}}^{(l)}(a) = \sum_b A_{t \to r}^{(l)}(b) P_{ab}. \tag{9}$$

$L_{t,\mathrm{ind}}$ is an indirect path to capture Generation $\rightarrow$ Query $\rightarrow$ matched Reference, and $L_{t,\mathrm{dir}}$ is a direct path to measure whether generation directly accesses the Reference tokens matched to Query location a. Together, we combine them as $L_t(a) = (L^{(l)}_{t,\mathrm{ind}}(a) + L^{(l)}_{t,\mathrm{dir}}(a))/2$, called the comparison-path map, which indicates where the generated reasoning establishes Query–Reference interaction.

Comparative Evidence. A useful comparison should satisfy both conditions: the Query region should differ from its normal counterpart, and the reasoning should actually access the corresponding cross-image relation. We therefore define $E_t(a) = D_a L_t(a)$. The three maps are complementary: D indicates where the Query differs from normal, L_t indicates where the reasoning performs cross-image comparison, and E_t highlights regions where both occur simultaneously. We use these signals to construct the comparative reward in CEPO.

3.4 Comparative Evidence Policy Optimization

Building on anomaly-aware comparative grounding via CAS, we further introduce CEPO, which augments standard group-relative optimization with a comparative evidence reward derived from CEG. Rather than optimizing sampled completions solely according to outcomes, CEPO also evaluates whether their reasoning is supported by meaningful Query–Reference evidence.

Comparative Evidence Reward. A reliable comparative trajectory should ground its reasoning in visual evidence that is consistent with the true anomaly. As defined previously, for each generation position t in completion y, CEG provides three complementary spatial signals: the discrepancy map D, the comparison-path map L_t, and the comparative evidence map E_t. For anomalous queries, we evaluate the spatial alignment of comparative evidence at each generation position as

$$R^{\mathrm{loc}}_t = \Phi(E_t, \widetilde{M}^q), \tag{10}$$

where Φ measures the spatial agreement between the comparative evidence and the ground-truth anomaly region, jointly considering anomaly precision, anomaly coverage, and background leakage. The detailed formulation of Φ is provided in Appendix A.1.

We then aggregate the position-level evidence rewards over the generated trajectory:

$$R^{\mathrm{cmp}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} R^{\mathrm{loc}}_t, \tag{11}$$

where $\mathcal{T}$ denotes the set of generated positions in completion y. The resulting R^{cmp} measures the overall quality of comparative visual evidence supporting the generated reasoning.

Task Reward. For the sampled completion y, the task-level reward is given by

$$R^{\mathrm{task}} = R^{\mathrm{res}} + R^{\mathrm{fmt}}, \tag{12}$$

where R^{res}_i measures anomaly-decision correctness and R^{fmt}_i evaluates output validity. Due to limited page space, we provide the detailed reward definitions in Appendix A.1.

Policy Optimization. Finally, we have the overall reward function by combining task performance and comparative evidence as:

$$R = R^{\mathrm{task}} + \lambda_{\mathrm{cmp}} R^{\mathrm{cmp}}, \tag{13}$$

where λ_{cmp} controls the contribution of the comparative reward. For normal queries, we set $R^{\mathrm{cmp}} = 0$, and policy optimization uses the task reward.

In this way, CEPO retains task-level supervision while distinguishing otherwise similar trajectories by the visual evidence underlying their reasoning. It therefore favors generations that are not only correct, but are grounded in anomaly-relevant Query regions and supported by explicit Query–Reference comparison.

4 Experiments

Section 4.1 details our experimental setup, including the training process and the benchmarks we adopt to validate our method. Section 4.2 presents a comprehensive performance comparison on two IAD benchmarks. Finally, Section 4.3 provides in-depth ablation studies of our method.

Table 1: Performance comparison of proprietary and open-source MLLMs on the MMAD dataset in the standard 1-shot setting. Anomaly Discrimination uses the average accuracy of normal and abnormal categories. The best results are highlighted in **bold** and the second-best are underlined.

Model	Scale	Anomaly	Defect				Object		Average
		Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Random Chance	-	50.00	25.00	25.00	25.00	25.00	25.00	25.00	28.57
Human (expert)	-	95.24	75.00	92.31	83.33	94.20	86.11	80.37	86.65
Human (ordinary)	-	86.90	66.25	85.58	71.25	81.52	89.58	69.72	78.69
Claude-3.5-sonnet	-	60.14	60.14	48.81	67.13	79.11	85.19	79.83	68.36
Gemini-1.5-pro	-	68.63	60.12	58.56	70.38	82.46	89.20	82.25	73.09
GPT-4o-mini	-	64.33	48.58	38.75	63.68	80.40	88.56	79.74	66.29
GPT-4o	-	68.63	65.80	55.62	73.21	83.41	94.98	82.80	74.92
Qwen2.5VL	3B	51.00	41.07	44.87	61.43	75.99	86.54	79.58	62.94
Anomaly-R1 (Chao et al., 2025)	3B	60.20	63.50	70.10	80.40	85.20	82.40	86.10	75.41
AgentIAD (Miao et al., 2026)	3B	69.49	72.73	80.94	85.27	87.84	93.27	90.59	82.88
CEPO (Ours)	3B	**73.85** (+4.36)	**80.77** (+8.04)	**82.25** (+1.31)	**89.34** (+4.07)	**89.11** (+1.27)	**96.99** (+3.72)	**93.31** (+2.72)	**86.52** (+3.64)
Qwen2.5VL	7B	70.47	56.15	60.35	64.30	78.34	92.11	83.58	72.19
AnomalyGPT (Gu et al., 2024)	7B	65.57	27.49	27.97	36.86	32.11	29.84	35.82	36.52
OmniAD (Zhao et al., 2026)	7B	68.80	78.80	75.50	67.20	86.40	96.00	86.40	79.87
Saliency-R1 (Gong et al., 2026)	7B	81.22	78.67	80.06	78.90	80.73	82.30	81.63	80.50
JUDO (Kang et al., 2026)	7B	64.51	72.17	75.95	84.38	87.76	94.24	86.07	80.73
AD-FM (Liao et al., 2026)	7B	73.15	73.36	77.53	86.80	86.75	89.98	86.67	82.03
CEPO (Ours)	7B	**81.72** (+0.50)	**85.75** (+6.95)	**85.86** (+5.80)	**91.16** (+4.36)	**90.99** (+3.23)	**97.48** (+1.48)	**94.06** (+7.39)	**89.57** (+7.54)

4.1 SETUP

Implementation Details. We adopt Qwen2.5-VL-7B (Bai et al., 2025) as the foundation model and apply LoRA (Hu et al., 2022) with rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.05. For supervised fine-tuning, we train for 3 epochs with a learning rate of 5×10^{-5} and a batch size of 16. We apply CAS to the last 4 decoder layers, setting $\lambda_{\text{CAS}} = 0.05$ and $\lambda_b = 0.01$. We linearly ramp up the auxiliary objectives over the first 10% of training.

For CEPO, each rollout batch contains 32 prompts, with 8 responses sampled per prompt and a maximum generation length of 512 tokens. We use a learning rate of 1×10^{-5}, with the GRPO clipping parameter set to $\epsilon = 0.2$ and the KL coefficient set to $\beta = 0.001$. For CEG, we set the correspondence temperature to $\tau_{\text{corr}} = 0.07$ and retain the top-$K_{\text{corr}} = 1$ Reference match for each Query token. The comparative reward is weighted by $\lambda_{\text{cmp}} = 1.0$. At inference time, we adopt min-p sampling (Nguyen et al., 2025) with a temperature of 1.5 and a min-p threshold of 0.1.

Evaluation Benchmarks. We evaluate our method on two benchmarks, MMAD (Jiang et al., 2025) and MMR-AD (Yao et al., 2026). For MMAD, we follow the original question–answering protocol and report accuracy for each task and their average. We sample 20% of the data from each of the 38 categories for training and use the remaining 80% for evaluation. For MMR-AD, we follow the leave-one-dataset-out protocol of (Yao et al., 2026), where one subdataset is held out for evaluation while the remaining subdatasets are used for training. We repeat this procedure for four target datasets: MVTecAD (Bergmann et al., 2019), VisA (Zou et al., 2022), MVTec3D (Bergmann et al., 2022), and MPDD (Jezek et al., 2021). Further relevant details are provided in Appendix A.3.

4.2 MAIN RESULTS

Results on MMAD. Table 1 reports the results under the standard 1-shot MMAD setting. While general-purpose MLLMs show reasonable performance on object-level tasks, their performance remains limited on anomaly-centric tasks, and existing anomaly-oriented methods substantially narrow this gap. Nevertheless, their improvements are often uneven across different capabilities. In contrast, CEPO achieves consistent gains across all seven MMAD tasks with both model scales.

With Qwen2.5-VL-7B, CEPO achieves an average score of 89.57, improving the base model by 17.38 points and surpassing the strongest competing 7B method by 7.54 points. Notably, CEPO obtains the best performance across all evaluated tasks, with particularly clear gains in defect classification, localization, and object analysis, demonstrating improvements in both anomaly perception and higher-level reasoning. The same trend holds at the 3B scale, where CEPO raises the average score from 62.94 to 86.52 and outperforms the strongest 3B competitor by 3.64 points. These consistent improvements across model scales and task types support the effectiveness of explicitly grounding anomaly-relevant evidence and optimizing Query–Reference comparative reasoning.

Table 2: Performance evaluation of anomaly detection and localization for both commercial and open-source MLLMs in MMR-AD dataset.

Model	Scale	MVTecAD		VisA		MVTec3D		MPDD		Avg.	
		Accuracy	BBox	Accuracy	BBox	Accuracy	BBox	Accuracy	BBox	Accuracy	BBox
Commercial Model											
Gemini-2.5-pro	-	79.4	34.4	65.7	10.5	75.7	10.9	62.1	16.3	70.7	18.0
GPT-5	-	78.7	41.8	65.8	18.5	75.0	27.2	68.4	21.2	72.0	27.2
OpenAI-o4-mini	-	81.3	34.4	73.0	12.3	74.1	15.7	70.5	23.9	74.7	21.6
Open Source Model											
Llama4-Maverick	-	76.1	19.9	63.8	6.7	67.1	6.1	64.1	10.0	67.8	10.7
Gemma3	27B	74.4	4.0	58.5	0.5	75.1	0.7	59.0	3.5	66.8	2.2
InternVL3	8B	78.4	5.5	71.5	2.9	75.5	1.0	68.3	3.9	73.4	3.3
InternVL3.5	8B	76.5	5.4	67.8	2.2	65.8	1.1	64.3	4.1	68.6	3.2
Qwen2.5-VL	7B	75.0	8.9	65.9	2.2	59.9	1.2	63.6	7.2	66.1	4.9
Task-Specific Anomaly Detection Model											
AnomalyGPT (Gu et al., 2024)	7B	85.7	-	72.8	-	64.7	-	67.6	-	72.7	-
Anomaly-R1 (Chao et al., 2025)	7B	88.8	66.3	74.9	36.4	**74.6**	**48.5**	70.8	35.9	77.3	46.8
CEPO (Ours)	7B	**90.8**	**80.8**	**75.5**	**39.1**	71.9	36.1	**77.5**	**61.6**	**78.9**	**54.4**

Results on MMR-AD. We further evaluate cross-dataset generalization on MMR-AD across four industrial anomaly benchmarks. For anomaly localization, we report the results under 0.1 IoU (i.e., if $IoU \geq 0.1$, the predicted box is considered correct). We include baselines that can be reliably reproduced under the same leave-one-dataset-out protocol. JUDO(Kang et al., 2026), AD-Copilot(Jiang et al., 2026), and AD-FM(Liao et al., 2026) are not included because their training data, model checkpoints, or complete implementations are not publicly available, preventing a fair comparison under our protocol. As shown in Table 2, CEPO achieves an average accuracy of 78.9% and a BBox score of 54.4%, substantially improving the Qwen2.5-VL-7B backbone from 66.1% and 4.9%, respectively. The particularly large gain in localization indicates that post-training not only improves anomaly recognition but also enables the model to more reliably identify spatially relevant anomaly evidence across unseen datasets.

Compared with the Anomaly-R1, CEPO improves the average accuracy from 77.3% to 78.9% and the BBox score from 46.8% to 54.4%. CEPO achieves the best detection and localization results on MVTecAD, VisA, and MPDD. The advantage is particularly pronounced in localization, reaching 80.8% on MVTecAD and 61.6% on MPDD, outperforming Anomaly-R1 by 14.5 and 25.7 points, respectively. CEPO also surpasses all evaluated proprietary and open-source general MLLMs in average detection accuracy, while demonstrating a substantially larger advantage in spatial localization. Overall, these results establish strong cross-dataset generalization of CEPO across diverse industrial anomaly scenarios, particularly in localizing anomaly-relevant visual evidence.

Table 3: Ablation studies of the proposed key components on MVTecAD. (a) Effects of CAS-SFT and CEPO. CE-SFT means SFT with vanilla CE. (b) Contributions of each components in Eq. 5.

(a) CAS and CEPO.

No.	CE-SFT	CAS-SFT	CEPO	Acc.	Δ	BBox	Δ
1	✓			84.01	-	55.51	-
2	✓	✓		85.05	+1.04	69.57	+14.06
3	✓	✓	✓	**90.84**	+5.79	**80.83**	+11.26

(b) CAS components.

No.	Q-Ground	R-Ground	Anomaly-Priority	Acc.	BBox
1				84.01	55.51
2	✓			84.42	58.45
3	✓	✓		85.01	60.10
4	✓	✓	✓	**85.05**	**69.57**

4.3 Ablation Studies and Analysis

Without loss of generality, we conduct ablation studies on MVTecAD of MMR-AD to investigate the contributions of CAS and CEPO, and the key components in SFT and reinforcement-learning stages. Unless otherwise specified, all variants use Qwen2.5-VL-7B and the same training configuration.

Effect of Key Components. We first examine the contribution of the two training stages in Table 3a. Incorporating CAS loss with CE loss improves accuracy to 85.05% and BBox to 69.57%, demonstrating the benefit of explicitly grounding the generated reasoning in anomaly-relevant Query regions and their corresponding normal Reference context. Building on this grounding, CEPO boosts the performance to 90.84% accuracy and 80.83% BBox, yielding additional gains of 5.79 and 11.26 points, respectively. These progressive improvements validate the complementary roles of the two components: CAS establishes anomaly-aware visual grounding, while CEPO further improves reasoning by optimizing the use of Query–Reference comparative evidence.

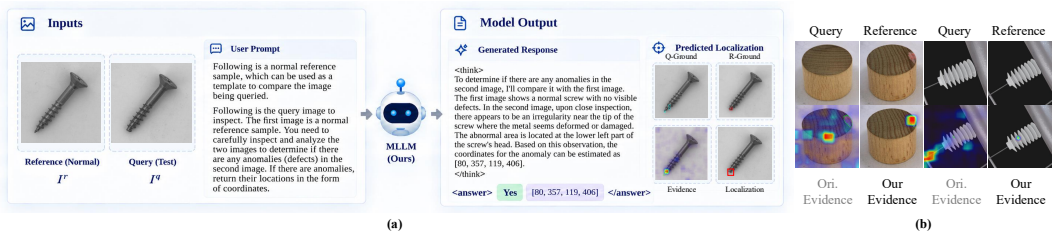


Figure 3: Qualitative results of visual grounding and anomaly localization. (a) An example showing the input image pair, generated response, Query and Reference grounding maps, comparative evidence, and predicted bounding box. (b) Comparison of evidence maps from the baseline and our method. Green contours indicate ground-truth anomaly regions.

Ablation of CAS. We further examine the contribution of each CAS component in Table 3b. Q-Ground improves both anomaly recognition and localization, confirming the importance of grounding generated reasoning in anomaly-relevant Query regions. Building on this, R-Ground further improves performance by incorporating the corresponding normal Reference evidence, extending visual grounding from Query-only supervision to explicit Query–Reference comparative grounding. This progression demonstrates that effective anomaly reasoning relies not only on identifying relevant evidence in the Query, but also on associating it with its normal counterpart in the Reference. Adding Anomaly-Priority further refines the grounding objective and leads to the full CAS performance of 85.05% accuracy and 69.57% BBox.

Ablation of RL Rewards. Finally, we ablate the reward design of the proposed CEPO in Table 4. On the CAS-SFT checkpoint, policy optimization with standard format and decision-correctness rewards improves accuracy from 85.05% to 88.15% and BBox accuracy from 69.57% to 77.57%, confirming the benefit of outcome-level optimization. Incorporating the proposed comparative evidence reward further raises accuracy to 90.84% and BBox accuracy to 80.83%, bringing additional gains of 2.69 and 3.26 points, respectively. Because the two policy-optimization settings share the same CAS initialization and differ only in the comparative reward, these gains are directly attributable to explicitly incorporating Query–Reference evidence into policy optimization, beyond conventional outcome supervision.

Table 4: Ablation study of the reward design during the RL stage on MMR-AD.

No.	Format	Result	Cmp.	Acc.	Δ	BBox	Δ
1				85.05	-	69.57	-
2	✓	✓		88.15	+3.10	77.57	+8.00
3	✓	✓	✓	**90.84**	+2.69	**80.83**	+3.26

Qualitative Analysis. We further examine visual grounding through qualitative examples in Figure 3. As illustrated in Figure 3(b), our method focuses on the actual anomalous regions, whereas the baseline attends to irrelevant areas, consistent with the localization gains reported above. Figure 3(a) presents a complete example: the grounding maps highlight the defective Query region and its corresponding normal Reference context, while the comparative evidence and predicted bounding box both focus on the screw tip. These visualizations illustrate improved anomaly grounding and spatial agreement between comparative evidence and predicted localization.

5 Conclusion

In this work, we study evidence-grounded comparative reasoning for reference-guided IAD. We show that reliable anomaly reasoning requires not only identifying anomaly-relevant evidence, but also comparing it with the corresponding normal reference. To this end, we introduce CAS for anomaly-aware visual grounding, together with CEG and CEPO to model and optimize Query–Reference comparative evidence. Experiments on MMAD and MMR-AD demonstrate consistent improvements in anomaly reasoning, localization, and cross-dataset generalization, while ablations verify the contributions of grounding and comparative evidence optimization.

AI USE STATEMENT

We used generative AI tools to assist with literature exploration, manuscript drafting, language refinement and experimental planning. All AI-assisted materials were reviewed and revised by the authors, who take full responsibility for the final manuscript, methodological claims, and reported results.

REPRODUCIBILITY STATEMENT

Sec. 3 provides the detailed network architectures and training objectives. Algorithmic details are displayed in Appendix A.1 and Algorithm 1. Appendix A.3 specifies the dataset splits, preprocessing and evaluation settings.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600. IEEE, 2019.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 202–213. SCITEPRESS - Science and Technology Publications, 2022.

Ankan Bhunia, Changjian Li, and Hakan Bilen. Odd-one-out: Anomaly detection by comparing with neighbors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20395–20404, 2025.

Mingxiu Cai, Zhe Zhang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. RAID: Retrieval-augmented anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21367–21378, 2026.

Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. AnomalyR1: A GRPO-based end-to-end MLLM for industrial anomaly detection. *arXiv preprint arXiv:2504.11914*, 2025. URL https://arxiv.org/abs/2504.11914.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9737–9746, 2022.

Shizhan Gong, Minda Hu, Qiyuan Zhang, Chen Ma, and Qi Dou. Saliency-R1: Enforcing interpretable and faithful vision-language reasoning via saliency-map alignment reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24095–24106, June 2026.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1932–1940, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 66–71, 2021. doi: 10.1109/ICUMT54235.2021.9631567.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *International Conference on Learning Representations*, 2025.

Xi Jiang, Yue Guo, Jian Li, Yong Liu, Bin-Bin Gao, Hanqiu Deng, Jun Liu, Heng Zhao, Chengjie Wang, and Feng Zheng. AD-Copilot: A vision-language assistant for industrial anomaly detection via visual in-context comparison, 2026. URL https://arxiv.org/abs/2603.13779.

Hyunju Kang, Woohyun Lee, Jaewon Kim, and Hogun Park. JUDO: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa. In *International Conference on Learning Representations*, 2026.

Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. IAD-R1: Reinforcing consistent reasoning in industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6583–6591, 2026.

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

Jingyi Liao, Yongyi Su, Rong-Cheng Tu, Zhao Jin, Wenhao Sun, Yiting Li, Xun Xu, Dacheng Tao, and Xulei 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*, volume 40, pp. 15234–15242, 2026. doi: 10.1609/aaai.v40i18.38548.

Hongchi Ma, Guanglei Yang, Debin Zhao, Yanli Ji, and Wangmeng Zuo. ReMP-AD: Retrieval-enhanced multi-modal prompt fusion for few-shot industrial visual anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20425–20434, 2025.

Junwen Miao, Penghui Du, Yingying Fan, Yi Liu, Yu Wang, Runze He, Lida Huang, and Yan Wang. AgentIAD: Agentic industrial anomaly detection via adaptive memory augmentation, 2026. URL https://arxiv.org/abs/2512.13671.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *International Conference on Learning Representations*, 2025.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024a.

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

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. MRAD: Zero-shot anomaly detection with memory-driven retrieval. In *International Conference on Learning Representations*, 2026.

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

Xincheng Yao, Zefeng Qian, Chao Shi, Jiayang Song, and Chongyang Zhang. MMR-AD: A large-scale multimodal dataset for benchmarking general anomaly detection with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43072–43082, June 2026.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8330–8339, 2021.

Shifang Zhao, Yiheng Lin, Lu Han, Yao Zhao, and Yunchao Wei. OmniAD: Detect and understand industrial anomaly via multimodal reasoning, 2026. URL https://arxiv.org/abs/2505.22039.

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

A Supplementary Material

A.1 Additional Implementation Details

Task rewards in Eq. 12. The task reward combines output-format validity and task performance:

$$R^{\text{task}} = R^{\text{fmt}} + R^{\text{res}}. \tag{14}$$

The format reward checks whether the completion contains properly ordered and parsable <think>...</think> and <answer>...</answer> spans and satisfies the task-specific answer schema. Let $\mathcal{F}$ denote the set of valid completions. We define

$$R^{\text{fmt}} = 0.5 \times \mathbf{1}[y \in \mathcal{F}]. \tag{15}$$

For structured anomaly detection, the result reward jointly evaluates decision correctness and defect coverage. Let $z \in \{0, 1\}$ denote the ground-truth anomaly label and $\hat{z}$ the predicted Yes/No decision, where $z = 1$ indicates an anomalous sample. For anomalous samples, predicted and ground-truth boxes are greedily matched using an IoU threshold of > 0.5. Each box participates in at most one match, preventing duplicate predictions from being reused to cover multiple defects. Let K_{miss} denote the number of unmatched ground-truth boxes. The result reward is

$$R^{\text{res}} = \underbrace{\mathbf{1}[\hat{z} = z]}_{\text{Decision correctness}} \underbrace{- 0.2 \times \mathbf{1}[z = 1] K_{\text{miss}}}_{\text{Missed-defect penalty}}, \tag{16}$$

where $K_{\text{miss}} = 0$ for normal samples. The correctness term is zero for an incorrect or unparseable decision. This reward encourages correct anomaly decisions while penalizing incomplete defect localization. For multiple-choice tasks, R^{res} reduces to an indicator of whether the predicted option matches the ground-truth answer.

The definition of Φ in Eq. 10. For an anomalous query, let $E_t(a)$ denote the comparative evidence at Query visual token a for generation position t, and let $m_a = \widetilde{M}_a^q \in \{0, 1\}$ denote its anomaly label. Before spatial evaluation, the evidence map is smoothed using a 3×3 average filter on the visual-token grid. For simplicity, E_t below denotes the smoothed map.

We evaluate evidence precision and anomaly coverage as

$$p_t = \frac{\sum_a m_a E_t(a)}{\sum_a E_t(a) + \epsilon}, \qquad c_t = \frac{\sum_a m_a E_t(a)}{\sum_a m_a + \epsilon}, \tag{17}$$

where ϵ ensures numerical stability. Precision measures the fraction of evidence inside the anomaly region, while coverage measures the normalized evidence strength across that region.

We combine precision and coverage using

$$F_{\beta_{\text{recall}},t} = \frac{(1 + \beta_{\text{recall}}^2) p_t c_t}{\beta_{\text{recall}}^2 p_t + c_t + \epsilon}, \qquad \beta_{\text{recall}} = 0.5. \tag{18}$$

The spatial evidence score further penalizes background evidence:

$$\Phi(E_t, \widetilde{M}^q) = F_{\beta_{\text{recall}},t} - \lambda_{\text{bg}} \frac{\sum_a (1 - m_a) E_t(a)}{\sum_a E_t(a) + \epsilon}, \qquad \lambda_{\text{bg}} = 0.5. \tag{19}$$

This score favors evidence concentrated within the annotated anomaly while encouraging support across the anomalous region. The mask-dependent score Φ is not evaluated for normal queries.

Experimental hyperparameter configurations. Table 5 summarizes the main training and inference settings. RL starts from the merged CAS-SFT checkpoint and updates LoRA parameters while keeping the visual encoder frozen. The frozen CAS-SFT policy serves as the KL reference, and truncated completions are excluded from the optimization loss. Ground-truth masks and boxes are used for training stage only. Inference requires neither spatial annotations nor additional reward computation.

Table 5: Main hyperparameters for the 7B model. RL settings correspond to the MVTecAD-held-out experiment.

Hyperparameter	Value
Backbone	Qwen2.5-VL-7B
LoRA rank / alpha / dropout	8 / 16 / 0.05
Supervised fine-tuning	
Epochs / batch size	3 / 16
Learning rate	5×10^{-5}
CAS layers	Last 4 layers
CAS / bridge weights	0.05 / 0.01
Auxiliary-loss ramp-up	First 10%
Reinforcement learning	
Samples / epoch	8,000 / 1
Learning rate	1×10^{-5}
Scheduler / precision	Cosine / BF16
Completions per prompt	8
Context / completion limit	4,096 / 512 tokens
Clipping / KL coefficient	0.2 / 0.001
Comparative-reward weight	1.0
Rollout and inference	
Temperature / min-p	1.5 / 0.1

A.2 Pseudo Code

The two-stage training procedure is summarized in Algorithm 1. CAS-SFT first establishes anomaly-aware visual grounding. CEPO then evaluates self-generated completions using task rewards and comparative evidence. Position-level comparative scores are aggregated into a trajectory reward before group-relative advantage computation and policy optimization.

A.3 Dataset Details

MMAD. MMAD evaluates multimodal industrial anomaly understanding through seven question-answering tasks: anomaly discrimination, defect classification, defect localization, defect description, defect analysis, object classification, and object analysis. Following the data-splitting protocol of prior work, we use 20% of the data from each of the 38 categories for training and the remaining 80% for evaluation, with a 1:1 ratio of normal to anomalous samples in the training set. Under the 1-shot setting, each Query is accompanied by one normal Reference image. We report the score for each task and their unweighted average. For anomaly discrimination, the reported score averages the accuracies on normal and anomalous samples.

Algorithm 1 Comparative Evidence Policy Optimization

Require: Training set $\mathcal{D}$, policy π_θ, rollout size G
Ensure: Optimized policy π_θ
Stage 1: Anomaly-Aware Grounding with CAS
for each $(I^r, I^q, x, y, M^q) \in \mathcal{D}$ **do**
  Project M^q onto Query and Reference token grids
  Extract Query–Reference grounding maps $\bar{A}^q$, $\bar{A}^r$
  Compute $\mathcal{L}_{\mathrm{CAS}}$ and $\mathcal{L}_{\mathrm{bridge}}$
  Update π_θ with $\mathcal{L}_{\mathrm{SFT}}$
end for
Stage 2: Comparative Evidence Policy Optimization
for each $(I^r, I^q, x) \in \mathcal{D}$ **do**
  Sample G responses $\{y_i\}_{i=1}^{G}$ from $\pi_{\theta_{\mathrm{old}}}$
  for $i = 1, \ldots, G$ **do**
    Evaluate task reward R_i^{task}
    Compute correspondence P, discrepancy D, and comparative paths L
    for $t \in \mathcal{T}_i$ **do**
      Construct evidence $E_t \leftarrow D \odot L_t$
      Evaluate D, L, and E to obtain R_t^{loc}
    end for
    Aggregate $\{R_t^{\mathrm{loc}}\}$ into R^{cmp}
    $R \leftarrow R^{\mathrm{task}} + \lambda_{\mathrm{cmp}} R^{\mathrm{cmp}}$
  end for
  $A \leftarrow (R - \mu_g)/(\sigma_g + \epsilon)$
  Update π_θ using $\{A\}_{i=1}^{G}$
end for
return π_θ

MMR-AD. MMR-AD evaluates anomaly detection and localization through structured model responses. We adopt leave-one-dataset-out evaluation, holding out MVTecAD, VisA, MVTec3D, and MPDD in turn. For each run, the target dataset is reserved for evaluation and the remaining source datasets provide training data. We report detection accuracy and the BBox localization score for each target dataset, together with their arithmetic averages across the four targets. The detailed RL configuration above uses 8,000 training samples for the MVTecAD-held-out run.

A.4 Qualitative Examples

Figure 4 presents qualitative examples of our model on diverse anomaly categories. For each case, we visualize the normal Reference and Query images, the generated reasoning and prediction, the Query and Reference grounding maps, the resulting comparative evidence, and the predicted anomaly localization. Across different object types and defect appearances, the model consistently concentrates Query grounding on anomaly-relevant regions while attending to the corresponding normal context in the Reference. The resulting comparative evidence is spatially aligned with the predicted anomalous region and supports accurate localization across both salient and subtle defects. These examples further illustrate that our method does not rely solely on Query-side anomaly cues, but explicitly incorporates corresponding Reference evidence during comparative reasoning. Overall, the visualizations qualitatively demonstrate the effectiveness of our approach in establishing anomaly-aware Query–Reference grounding and producing spatially coherent comparative evidence for anomaly reasoning.

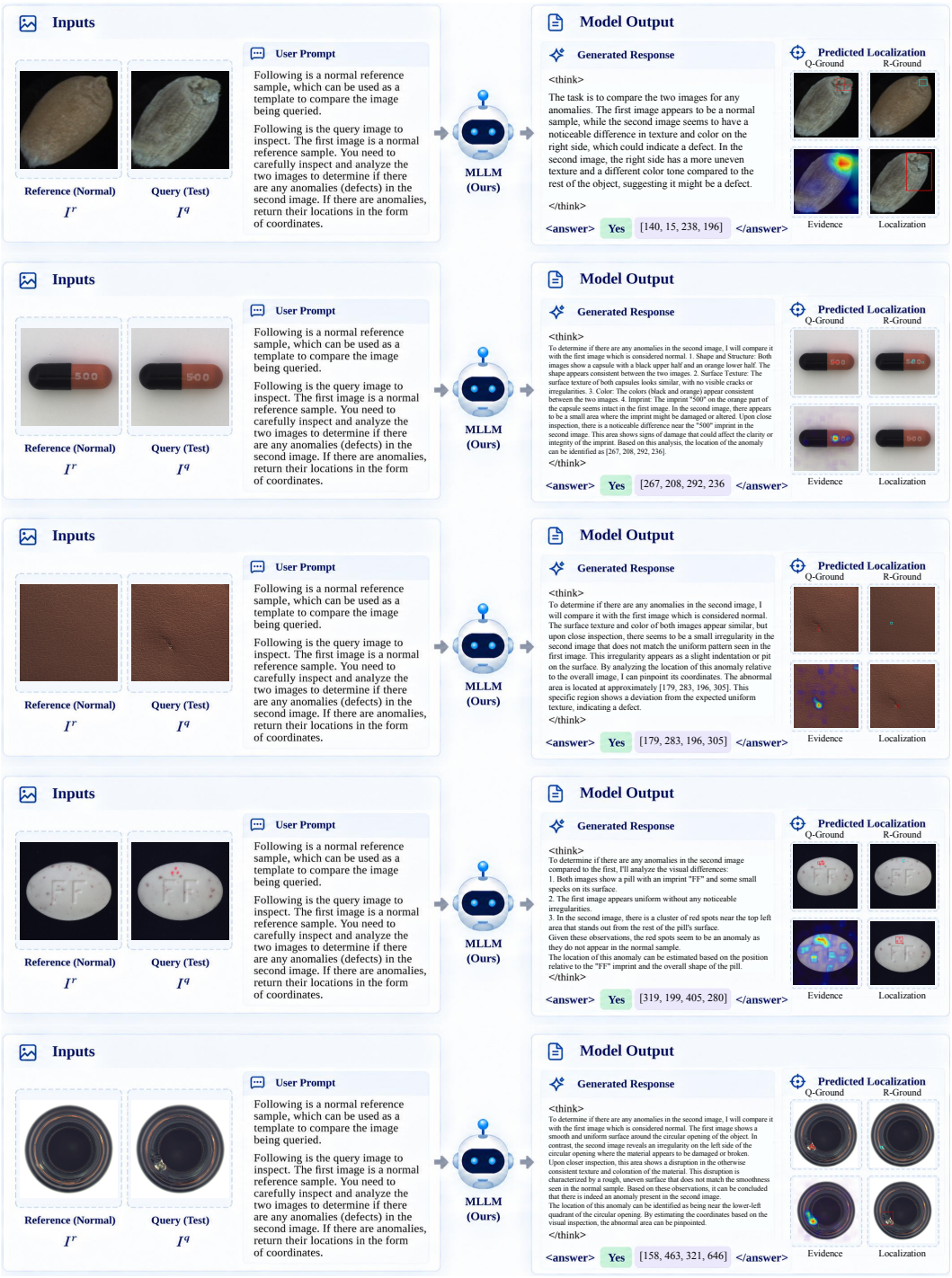


Figure 4: Qualitative visualization of reference-guided anomaly reasoning across representative examples. For each sample, we show the normal Reference and Query images, the generated response, Query grounding (Q-Ground), Reference grounding (R-Ground), comparative evidence, and the predicted anomaly localization. Across diverse defect types, our model grounds anomalous Query regions while attending to the corresponding normal Reference context, resulting in comparative evidence that is spatially consistent with the predicted localization.