EGVLR: Evidence-Grounded Vision–Language Reinforcement for Anomaly Reasoning

Anonymous ECCV 2026 Submission

Paper ID #7001

Abstract. Multimodal large language models (MLLMs) have recently shown promise for industrial anomaly understanding, where models are expected not only to determine whether an object is defective, but also to localize visual evidence and explain defect-related decisions. However, existing MLLMs often rely on language priors or shortcut reasoning when facing fine-grained industrial defects, resulting in inconsistent outputs such as correct answers with invalid locations, hallucinated defect boxes on normal samples, or rationales that contradict the final prediction. To address these issues, we propose **EGVLR**, an evidence-grounded vision–language reinforcement framework for anomaly reasoning. EGVLR follows a unified Evidence-Driven Diagnostic Protocol (EDDP), where each response explicitly contains visual evidence, diagnostic logic, spatial location, and the final answer. The framework consists of four stages: Progressive Visual-Evidential Fine-Tuning (PVE-FT), Knowledge-Grounded Instruction Tuning (KG-IT), Geometry-Semantic Decoupled Preference Optimization (GS-DPO), and Box-Guided Segmentation Rendering (BGSR). PVE-FT first aligns the model with localized visual evidence through synthetic anomalies, 3×3 grid grounding, local verification decoys, and null-hypothesis calibration. KG-IT then internalizes industrial-domain QA behavior while preserving the same output schema. GS-DPO further optimizes answer correctness, geometric localization, BGE-based rationale semantics, and answer–location–rationale consistency. Finally, BGSR converts predicted bounding boxes into dense masks using an off-the-shelf segmentation backend. Experiments on MMAD demonstrate that EGVLR improves evidence-grounded industrial anomaly reasoning, especially on defect localization, reasoning-dependent QA, and spatial false-positive control.

Keywords: Industrial Anomaly Reasoning · Multimodal Large Language Models · Reinforcement Learning · Visual Grounding · Preference Optimization

1 Introduction

Industrial anomaly inspection is moving from closed-set defect scoring toward interactive anomaly understanding. In a conventional pipeline, an anomaly detector returns an image-level score or a pixel-level heatmap; a human inspector must still decide what the defect is, whether the highlighted region is meaningful, and how the query differs from a normal product. Recent multimodal

large language models (MLLMs) make a richer interface possible: they can receive inspection images and natural-language questions, compare a query with a normal reference, describe visible deviations, and return structured answers. Benchmarks such as MMAD [**?**] therefore evaluate not only whether a model detects an anomaly, but also whether it can reason about defect category, location, description, and analysis.

This setting exposes a gap between general visual-language capability and industrial inspection reliability. Industrial defects are usually sparse, local, and visually subtle. A small crack, contaminant, missing component, solder defect, or abnormal texture may occupy only a tiny region, while the rest of the image remains visually normal. Moreover, the meaning of a deviation is category-dependent: a texture change that is harmless for one product may be a defect for another. General MLLMs are mostly trained on broad web image–text data, where such localized industrial evidence is rare and where answers are not tied to precise spatial evidence. As a result, they may answer inspection questions from language priors instead of image evidence. A model can describe a plausible scratch because the prompt asks about defects, or output a non-empty location even when the query is visually consistent with the normal reference.

For industrial anomaly reasoning, the failure is often not limited to the final answer. A response can be unusable even when the selected option is correct: the predicted box may not cover the defect, the rationale may mention a visual cue that is absent, or a normal sample may receive a hallucinated defect location. We refer to this as evidence inconsistency among answer, location, and rationale. Such inconsistency is problematic for deployment because industrial users require inspectable decision traces, not merely a multiple-choice answer. In particular, false-positive boxes on normal samples increase manual verification cost and weaken trust in the system.

A straightforward approach is to fine-tune an MLLM on anomaly QA data, but supervised fine-tuning alone mainly encourages imitation of target responses. It does not directly discourage contradictory traces such as a normal answer paired with abnormal evidence, or a correct answer paired with an invalid box. Preference or reinforcement learning can optimize such structured properties, but applying it directly to a base MLLM is unstable: the model may not follow the required output format, spatial rewards become ambiguous for normal or non-spatial questions, and answer-only rewards may reinforce shortcut decisions. These observations suggest that anomaly reasoning should be aligned progressively: first establish visual evidence grounding, then introduce domain-oriented instructions, and finally optimize explicit consistency rewards.

We propose **EGVLR**, an Evidence-Grounded Vision–Language Reinforcement framework for industrial anomaly reasoning. EGVLR is organized around a unified Evidence-Driven Diagnostic Protocol (EDDP). Every trainable stage produces the same structured output containing `<evidence>`, `<logic>`, `<location>`, and `<answer>`. The location field is always a list of normalized bounding boxes, and defect-free, negative-region, or non-spatial samples use `<location>[]</location>`. This simple constraint removes ambiguity between grid labels, boxes, and nor-

mal cases, allowing the same parser and reward rules to be used throughout training and rendering.

EGVLR has four stages. **Progressive Visual-Evidential Fine-Tuning (PVE-FT)** first trains visual evidence grounding with synthetic localized anomalies, 3×3 grid supervision, local verification decoys, and null-hypothesis calibration. **Knowledge-Grounded Instruction Tuning (KG-IT)** then teaches industrial-domain QA behavior using domain questions, one-normal visual QA, and comparative inspection QA while preserving the EDDP schema. **Geometry-Semantic Decoupled Preference Optimization (GS-DPO)** further aligns the model by separately rewarding format validity, answer correctness, box geometry, BGE-based rationale semantics, and answer–location–rationale coupling. Finally, **Box-Guided Segmentation Rendering (BGSR)** maps predicted boxes to dense masks using an external segmentation backend; it is a rendering step and does not change the MLLM answer.

We evaluate EGVLR on MMAD under the industrial anomaly QA setting. The results show that the proposed staged alignment improves evidence-grounded reasoning and reduces spatial false positives compared with general MLLMs and anomaly-oriented baselines. Ablations demonstrate that visual pre-alignment, knowledge-grounded tuning, geometry-semantic preference optimization, and box-guided rendering address different failure modes and are complementary.

Contributions. The main contributions are summarized as follows:

- We introduce a unified Evidence-Driven Diagnostic Protocol (EDDP) that connects visual evidence, diagnostic logic, normalized bounding-box localization, and final answers across all trainable stages.
- We propose Progressive Visual-Evidential Fine-Tuning (PVE-FT), a vision-only pre-alignment stage that teaches localized evidence grounding through synthetic anomalies, grid localization, local decoy verification, and null-hypothesis calibration.
- We design Geometry-Semantic Decoupled Preference Optimization (GS-DPO), which combines answer reward, box reward, BGE-based rationale semantic reward, and semantic coupling reward to reduce inconsistent reasoning and spatial reward hacking.
- We conduct MMAD-focused experiments and ablations that analyze stage contributions, reward components, normal false-positive behavior, and box-guided segmentation rendering.

2 Related Work

2.1 Industrial Anomaly Detection and Foundation Models

Industrial anomaly detection (IAD) has been extensively studied as image-level anomaly classification and pixel-level anomaly localization. Representative benchmarks such as MVTec AD, VisA, and Real-IAD [**?**,**?**,**?**] have supported methods

based on reconstruction, feature modeling, patch memory, and discriminative segmentation [?,?]. More recent work explores multi-class, cross-class, zero-shot, and few-shot settings by leveraging pretrained visual or vision-language representations [?,?]. These methods improve transferability and localization, but their outputs are usually anomaly scores, heatmaps, or masks. They are not designed to answer natural-language questions or provide explicit diagnostic rationales.

EGVLR targets a complementary problem. Instead of replacing image-only anomaly detectors, it studies how an MLLM can produce an inspectable reasoning trace: what evidence is observed, how the evidence supports the decision, where the abnormal region is, and what final answer should be selected. This requires both visual grounding and language-level consistency, which are not directly optimized in most conventional IAD frameworks.

2.2 Multimodal Anomaly Data and Benchmarks

The rise of MLLMs has motivated multimodal anomaly benchmarks that evaluate language-conditioned inspection. MMAD [?] formulates industrial anomaly understanding as multiple-choice QA over object recognition, defect discrimination, localization, description, and analysis. Anomaly instruction datasets such as Anomaly-Instruct-125K [?] and broader multimodal anomaly resources such as MMR-AD [?] further indicate that anomaly-specific post-training data is important for adapting MLLMs to fine-grained industrial evidence. These resources show that general MLLMs are not automatically reliable anomaly reasoners: they often miss small defects, misinterpret benign variations, or localize imprecisely.

Our work follows this line but focuses on the reliability of the reasoning trace rather than only the availability of multimodal anomaly data. We use MMAD as the main evaluation benchmark and design training stages that explicitly address evidence grounding, normal-case calibration, and answer–location–rationale consistency. In this sense, EGVLR complements existing datasets by providing a post-training and reward-design protocol for structured anomaly reasoning.

2.3 MLLMs for Industrial Anomaly Understanding

Several recent methods adapt MLLMs to industrial inspection through prompting, instruction tuning, or task-specific visual modules. AnomalyGPT [?], Myriad [?], FabGPT [?], VMAD [?], and Anomaly-OV [?] demonstrate that language supervision can improve anomaly recognition and explanation. Other works, including OmniAD [?], JUDO [?], AnomalyR1 [?], and IAD-R1 [?], further explore reference-guided reasoning, domain knowledge, and reinforcement-style training for industrial anomaly QA.

These studies establish that anomaly-oriented post-training is necessary, but several reliability questions remain open. First, a response format may contain a rationale without forcing the rationale to match the final decision. Second, spatial predictions may be evaluated only on abnormal samples, leaving normal-case box hallucination under-specified. Third, domain knowledge and visual grounding are often learned together, making it difficult to tell whether gains come from better

visual evidence localization or from stronger textual priors. EGVLR separates these roles: PVE-FT is a vision-only pre-alignment stage, KG-IT injects domain-oriented QA behavior, and GS-DPO explicitly optimizes the consistency among answer, location, and rationale.

2.4 Instruction Tuning and Preference Optimization

Instruction tuning is widely used to adapt MLLMs to downstream tasks [?, ?]. For anomaly reasoning, it can teach prompt following, option selection, domain terminology, and structured response style. However, because SFT is token-level imitation, it does not directly reward the global validity of a generated trace. A model may learn to output the correct tags but still produce a box inconsistent with the answer or a rationale inconsistent with the visual evidence.

Preference and reinforcement learning methods offer a way to align generated responses with task-level rewards. Recent rule-based reasoning methods and GRPO-style training [?, ?] reduce the need for manually annotated preference pairs by computing rewards from verifiable output properties. For anomaly reasoning, reward design must handle both language and geometry. Answer-only rewards can encourage shortcut prediction, whereas geometry-only rewards can be exploited through oversized or redundant boxes. EGVLR therefore decomposes GS-DPO into format, answer, box, semantic rationale, and coupling channels. The semantic channel uses BGE embeddings [?] to compare rationales with normal, abnormal, and domain-reasoning prototypes, while the coupling channel checks whether the predicted answer, location, and rationale direction agree.

2.5 Reference-Guided Reasoning and Box-Prompted Rendering

Industrial defects are often defined relative to normal appearance, so reference-guided comparison is an important cue for anomaly reasoning. A normal reference helps distinguish true defect evidence from benign variation in pose, lighting, texture, or product layout. EGVLR uses a fixed paired-image convention: the first image is the query/test image and the second image is the normal reference. This convention is used consistently in PVE-FT, KG-IT, and GS-DPO to reduce ambiguity in comparative reasoning.

For dense visualization, segmentation foundation models such as SAM variants [?, ?] can convert spatial prompts into masks. EGVLR treats this as a rendering problem rather than a reasoning problem. The MLLM predicts normalized boxes under the EDDP schema; BGSR maps non-empty boxes to pixel coordinates and sends them to an off-the-shelf segmentation backend. If the location is empty, no mask is produced. This separation avoids attributing the segmentation backend's mask quality to the MLLM's reasoning ability and keeps the training-time contribution distinct from the visualization module.

3 Method

3.1 Overview and Unified EDDP Schema

EGVLR follows a four-stage pipeline for industrial anomaly reasoning: PVE-FT, KG-IT, GS-DPO, and BGSR. The central design is a unified Evidence-Driven Diagnostic Protocol (EDDP) used by all trainable stages:

$$y = \texttt{<think>}\ \texttt{<evidence>}\ e\ \texttt{</evidence>}\ \texttt{<logic>}\ l\ \texttt{</logic>}\ \texttt{</think>}\ \texttt{<location>}\ B\ \texttt{</location>}\ \texttt{<answer>}\ a\ \texttt{</answer>}. \tag{1}$$

Here, e is the visual or domain evidence, l is the diagnostic logic, a is the final answer, and B is a list of normalized bounding boxes. Each box is represented as $[x_1, y_1, x_2, y_2]$ in the $[0, 1000]$ coordinate system. For normal, negative-region, or non-spatial QA samples, the correct location is the empty list, implemented as `<location>[]</location>`. The `<location>` field never stores a grid ID; grid IDs are used only as textual spatial evidence and metadata.

For paired inputs, the image order is fixed throughout the whole framework:

$$\text{first image} = \text{query/test image}, \quad \text{second image} = \text{normal reference image}. \tag{2}$$

This convention removes ambiguity between comparative training, reward parsing, and mask rendering. Figure 1 summarizes the complete pipeline and highlights how the same EDDP schema connects visual pre-alignment, knowledge-grounded tuning, preference optimization, and box-guided rendering.

3.2 Stage I: Progressive Visual-Evidential Fine-Tuning

PVE-FT teaches the MLLM to ground its decision in localized visual evidence before introducing domain-level reasoning. Given a normal image I, we synthesize a query image $\tilde{I}$ by inserting a localized anomaly with CutPaste-style patch replacement or DTD texture insertion. The generator records the anomaly mask M, normalized box $B^\star$, 3×3 grid ID $g^\star$, and optional subregion label $s^\star$. The target rationale is then generated programmatically from this controlled metadata. It only verbalizes observable visual evidence and does not introduce external domain knowledge.

PVE-FT contains five visual-only QA families: synthetic anomaly detection, 3×3 grid localization, fine subregion localization, local verification with positive and decoy regions, and normal-reference null-hypothesis calibration. For all grid-based questions, only the query image is divided into a 3×3 grid over the displayed image canvas. The full grid convention and prompt templates are provided in the supplementary material.

For abnormal synthetic samples, the target `<location>` contains the generated box. For normal samples and decoy-region samples, the target is `<location>[]</location>`. PVE-FT is optimized with autoregressive cross-entropy:

$$\mathcal{L}_{\mathrm{PVE}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{PVE}}} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{3}$$

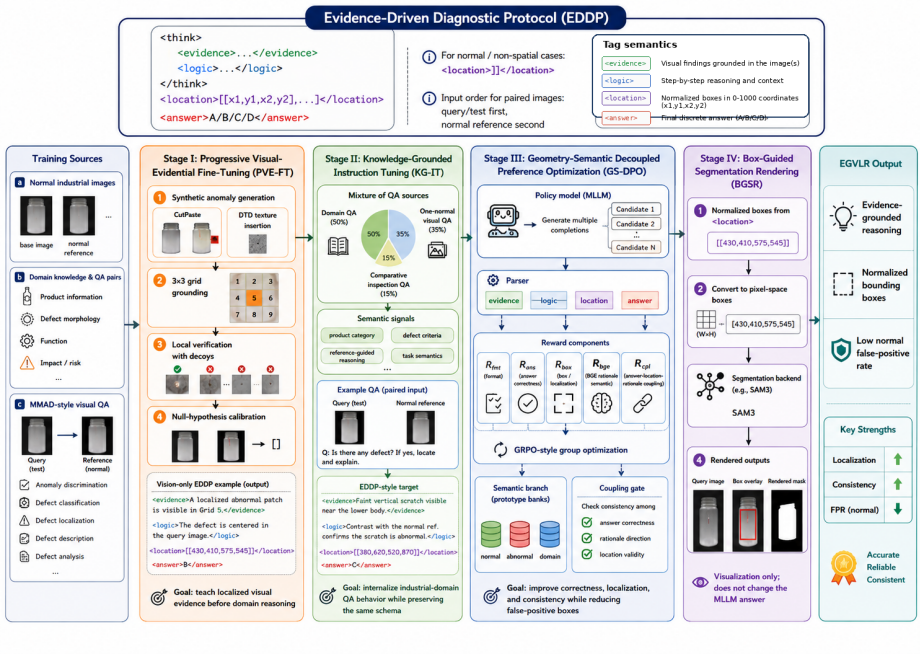


Fig. 1: Overview of EGVLR. The framework uses one unified Evidence-Driven Diagnostic Protocol (EDDP) across all trainable stages. PVE-FT teaches localized visual evidence with synthetic anomalies, grid grounding, local decoys, and null-hypothesis calibration. KG-IT injects domain-oriented and reference-guided QA behavior. GS-DPO aligns answer correctness, box validity, rationale semantics, and answer–location–rationale coupling. BGSR converts non-empty predicted boxes into rendered masks using an external segmentation backend; it does not change the MLLM answer.

3.3 Stage II: Knowledge-Grounded Instruction Tuning

After visual-evidential pre-alignment, KG-IT teaches the model task semantics, object-level inspection knowledge, and reference-guided anomaly QA behavior. In our MMAD instantiation, KG-IT uses three sources: domain QA generated from domain snippets, one-normal visual QA, and comparative inspection QA. Unless otherwise specified, we use a 50:35:15 mixture ratio for domain QA, one-normal visual QA, and comparative QA.

KG-IT preserves the same EDDP schema. For samples without reliable box supervision, we still keep the `<location>` field and set it to `[]`, while marking these samples as not requiring spatial reward during preference optimization. This avoids format drift between PVE-FT, KG-IT, and GS-DPO. KG-IT is optimized with:

$$\mathcal{L}_{\mathrm{KG}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{KG}}} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{4}$$

3.4 Stage III: Geometry-Semantic Decoupled Preference Optimization

Instruction tuning teaches the model to follow the EDDP schema, but token-level imitation alone cannot explicitly penalize inconsistent outputs. GS-DPO therefore decouples the reward into format, answer, geometry, semantic rationale, and coupling components. We implement this stage with a GRPO-style objective. For each prompt q, the policy samples a group of G completions $\{o_i\}_{i=1}^{G}$ and assigns reward R_i to each completion. The group-normalized advantage is:

$$\hat{A}_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \quad \mu_R = \frac{1}{G}\sum_{i=1}^{G} R_i. \tag{5}$$

When the group reward variance is zero, we set the normalized advantages to zero to avoid unstable updates. The policy objective is:

$$\mathcal{J}_{\mathrm{GS}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i(\theta)\hat{A}_i, \mathrm{clip}(\rho_i(\theta), 1-\epsilon_c, 1+\epsilon_c)\hat{A}_i\right) - \beta D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})\right], \tag{6}$$

where $\rho_i(\theta) = \pi_\theta(o_i \mid q)/\pi_{\theta_{\mathrm{old}}}(o_i \mid q)$.

Decoupled reward. The total reward is:

$$R = \lambda_f R_{\mathrm{fmt}} + \lambda_a R_{\mathrm{ans}} + \lambda_b R_{\mathrm{box}} + \lambda_r R_{\mathrm{bge}} + \lambda_c R_{\mathrm{cpl}}. \tag{7}$$

R_{fmt} checks parseable EDDP fields, and $R_{\mathrm{ans}} = \mathbb{I}(a_{\mathrm{pred}} = a_{\mathrm{gt}})$. For abnormal samples with spatial supervision, R_{box} matches predicted and ground-truth boxes using IoU or DIoU with a count penalty for redundant boxes. For normal or negative-region samples, $R_{\mathrm{box}} = 1$ if the predicted location is empty and -1 otherwise. For non-spatial QA marked as `require_location=false`, $R_{\mathrm{box}} = 0$.

For rationale semantics, we encode the generated evidence and logic using `BAAI/bge-small-en-v1.5`. Let s_N, s_A, and s_O be the maximum cosine similarities between the generated rationale and normal, abnormal, and domain-reasoning prototype banks, respectively. For abnormal visual samples, $R_{\mathrm{bge}} = \tanh(\gamma(s_A - s_N))$; for normal or negative-region samples, $R_{\mathrm{bge}} = \tanh(\gamma(s_N - s_A))$; and for domain-only QA, $R_{\mathrm{bge}} = \tanh(\gamma(s_O - \max(s_N, s_A)))$.

The semantic coupling reward enforces answer–location–rationale consistency:

$$R_{\mathrm{cpl}} = \mathbb{I}(a_{\mathrm{pred}} = a_{\mathrm{gt}}) \cdot \mathbb{I}(\mathrm{LocValid}) \cdot \mathbb{I}(\mathrm{SemValid}). \tag{8}$$

For normal samples, location validity means an empty location and semantic validity means a normal-supporting rationale. For abnormal samples, location validity means a valid matched box and semantic validity means an abnormal-supporting rationale. This reward design penalizes redundant boxes, non-empty boxes on normal images, answer-correct but location-invalid completions, and rationales that contradict the final answer.

3.5 Stage IV: Box-Guided Segmentation Rendering

BGSR does not further train or modify the MLLM. It only converts the predicted `<location>` field into dense visual masks. If the predicted location is empty, no mask is generated. Otherwise, each normalized box is mapped back to the original image resolution and used as a box prompt for an off-the-shelf segmentation backend. In our implementation, we use SAM3. This stage decouples cognitive anomaly reasoning from dense mask rendering: the MLLM is responsible for evidence-grounded reasoning and box localization, while the segmentation backend only renders pixel-level masks from verified boxes.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on MMAD [?], a multimodal benchmark for industrial anomaly understanding. MMAD evaluates anomaly discrimination and reasoning-dependent defect QA, including defect classification, localization, description, and analysis. All medical-domain experiments are omitted from the main paper to keep the evaluation focused on industrial anomaly reasoning.

Backbone and training. We instantiate EGVLR with Qwen3-VL-8B-Instruct as the backbone. For paired-image inputs, the first image is the query/test image and the second image is a normal reference image from the same object category. PVE-FT uses synthetic visual-evidential QA generated from normal industrial images. KG-IT uses the 50:35:15 mixture described in Sec. 3.3. GS-DPO is initialized from the KG-IT checkpoint and optimized with the reward in Eq. 7. BGSR is applied only for visualization and does not modify model answers.

Baselines and metrics. We compare against general-purpose MLLMs, open-source VLMs, and anomaly-specialized baselines, including Qwen-VL variants, LLaVA-style models, AnomalyGPT, OmniAD, AnomalyR1, IAD-R1, and JUDO when available. For QA tasks, we report accuracy. For spatial reliability, we additionally report localization accuracy, format validity, image-level false-positive rate FPR_{img}, and average false-positive box count FPBox on normal samples:

$$\mathrm{FPR}_{img} = \frac{\#\{x \in \mathcal{N} : \hat{B}(x) \neq \emptyset\}}{|\mathcal{N}|}, \quad \mathrm{FPBox} = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} |\hat{B}(x)|. \tag{9}$$

4.2 Main Results on MMAD

Table 1 reports the main comparison on MMAD. EGVLR improves reasoning-dependent tasks such as defect localization, defect description, and defect analysis, indicating that simply scaling a general MLLM is insufficient for industrial anomaly reasoning. The gains are most visible when the task requires the model to connect localized evidence with a final answer. Compared with anomaly-specialized baselines, EGVLR further improves spatial reliability by reducing non-empty boxes on normal samples.

Table 1: Comprehensive performance comparison on the MMAD industrial benchmark. All models are evaluated under the 1-shot (anchor-guided) setting. The best results are highlighted in **bold** and the second-best are underlined.

Model	Scale	Anomaly Discrim.	Defect Class.	Defect Local.	Defect Desc.	Defect Analysis	Object Class.	Object Analysis	Average
Commercial Models									
GPT-4o	-	68.63	65.80	55.62	73.21	83.41	94.98	82.80	74.92
Gemini 2.5 Pro	-	**83.07**	73.86	67.20	79.97	86.27	94.88	83.08	81.19
Open-Source & Specialized									
Qwen2.5-VL-Instruct	7B	71.39	54.35	61.17	65.81	79.32	91.44	84.43	72.56
LLaVA-OneVision-1.5	8B	60.85	56.29	55.25	74.73	83.21	91.03	**89.14**	72.93
Qwen3-VL-Instruct (Base)	8B	69.75	59.34	59.08	73.54	81.29	90.30	89.07	74.62
AnomalyR1	7B	60.93	64.81	70.72	79.06	85.52	93.12	86.91	77.29
OmniAD	7B	68.80	**78.80**	75.50	67.20	86.40	**96.00**	86.40	79.90
JUDO	7B	64.51	72.17	75.95	84.38	87.76	94.24	86.07	80.73
EGVLR (Ours)	**8B**	71.50	78.75	**76.85**	**84.42**	**87.82**	93.30	85.21	**82.55**

Backbone effect. We also compare our Qwen2.5-VL-7B and Qwen3-VL-8B EGVLR variants to separate backbone scaling from evidence-grounded alignment. Reported open-source MMAD one-shot baselines show that Qwen3-VL-8B improves the average score over Qwen2.5-VL-7B, but its native defect localization score is lower. In contrast, after applying EGVLR, the Qwen3-based variant slightly surpasses the Qwen2.5-based variant in defect localization. This suggests that PVE-FT and GS-DPO compensate for Qwen3's native localization weakness by explicitly aligning localized evidence, normalized boxes, and rationale semantics; the spatial gains are therefore mainly attributed to the proposed evidence-grounded alignment pipeline rather than backbone scaling alone.

4.3 Stage and Reward Ablation

We ablate the contribution of each stage and reward component in Tables 2 and 3. PVE-FT improves format validity and coarse localization because it teaches the EDDP schema and visual grounding before domain QA. KG-IT improves task-level reasoning by injecting domain-oriented instruction behavior. GS-DPO further improves consistency-sensitive metrics, especially localization and false-positive control. In reward ablation, answer-only optimization improves final-choice accuracy but is insufficient for reliable anomaly reasoning. Removing the box reward weakens localization; removing the BGE rationale reward weakens semantic alignment; removing the coupling reward increases answer–location–rationale contradictions.

4.4 Diagnostic Analysis

Table 4 analyzes visual pre-alignment and false-positive control. Removing grid grounding weakens localization, while removing local verification decoys or null-hypothesis calibration increases non-empty box predictions on normal samples.

Table 2: Stage ablation on MMAD. FPR_{img} is reported in percentage. BGSR only affects mask rendering and does not modify QA accuracy.

Variant	Avg. Acc.	Loc. Acc.	Format	$FPR_{img}\downarrow$	FPBox↓
Base MLLM	74.62	59.08	58.7	42.6	0.78
+ PVE-FT	77.84	71.90	94.6	29.4	0.47
+ PVE-FT + KG-IT	80.63	75.42	97.3	23.8	0.34
+ PVE-FT + KG-IT + GS-DPO	**82.55**	**76.85**	**98.6**	**13.9**	**0.18**

Table 3: Reward ablation for GS-DPO. FPR_{img} is reported in percentage. The full reward combines format, answer, box, BGE rationale, and semantic coupling rewards.

Variant	Avg. Acc.	Loc. Acc.	Format	$FPR_{img}\downarrow$	FPBox↓
Answer-only GRPO	80.73	72.34	94.8	31.7	0.53
Standard GRPO	81.62	75.02	97.5	24.6	0.37
w/o box reward	80.41	71.72	98.2	30.5	0.49
w/o BGE rationale reward	81.86	75.91	98.4	20.7	0.31
w/o semantic coupling	81.94	76.18	98.3	19.3	0.28
Full GS-DPO	**82.55**	**76.85**	**98.6**	**13.9**	**0.18**

GS-DPO further reduces both image-level false positives and average false-positive boxes, showing that semantic coupling helps suppress spatial hallucination. BGSR is used only for visualization and does not affect QA accuracy; detailed mask rendering results are provided in the supplementary material.

4.5 Qualitative Analysis

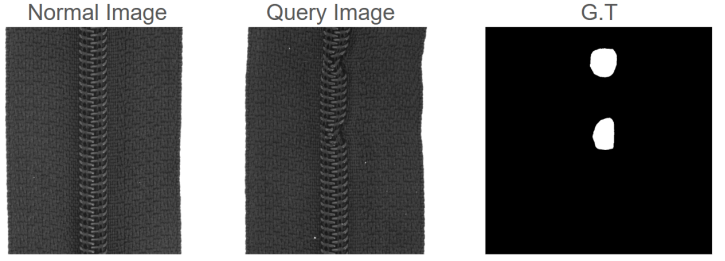
Figure 2 shows qualitative examples on MMAD. General-purpose MLLMs often produce plausible explanations without reliable spatial grounding. In contrast, EGVLR first identifies localized evidence, then predicts a bounding box, and finally provides the answer. On normal samples, EGVLR tends to retain the null hypothesis and outputs an empty location. This behavior is consistent with the design of PVE-FT and GS-DPO.

4.6 Failure Cases

Although EGVLR improves evidence-grounded anomaly reasoning, several failure cases remain. Extremely small or low-contrast defects may still be missed, especially when the defect resembles normal texture variation. When multiple defects appear in one image, the model may localize only the most salient region and miss secondary defects. Finally, BGE-based semantic scoring is more robust than keyword matching but is not a full natural-language inference model and may still be affected by subtle negation or ambiguous rationale wording.

Table 4: Diagnostic analysis on spatial grounding and false-positive control. FPR_{img} is reported in percentage.

Setting	Avg. Acc.	Loc. Acc.	$\text{FPR}_{img}\downarrow$	FPBox↓
w/o grid convention	81.76	74.38	15.6	0.21
w/o local decoys	81.61	75.72	21.8	0.34
w/o null calibration	81.42	75.88	25.4	0.41
SFT only	80.63	75.42	23.8	0.34
Full GS-DPO	**82.55**	**76.85**	**13.9**	**0.18**



Model	Diagnostic reasoning trace	Prediction
LLaVA-OneVision (8B)	The object is identified as a zipper, but the response attributes the defect to missing teeth and describes non-existent gaps along the track. This is a language-prior hallucination: the explanation is plausible for a zipper, but it is not supported by the visual evidence.	× Missing teeth
Qwen3-VL-Instruct (8B)	The response predicts a defect but gives a generic description about irregular coloration and border structure. It fails to isolate the true morphological cue on the zipper teeth and does not provide a reliable localized decision trace.	× Broken border
EGVLR (Ours) (8B)	<think> <evidence> Compared with the normal reference, the fabric border remains intact, while two localized regions on the center zipper track show teeth pushed closer together without clear missing gaps. </evidence> <logic> The abnormal cue is a spacing deformation of the teeth rather than missing parts or border damage; therefore the defect corresponds to squeezed zipper teeth. </logic> </think> <location>[[430,90,570,230],[450,380,560,540]]</ <answer>C</answer>	✓ **Teeth pushed closer together**

Fig. 2: Qualitative comparison on a zipper anomaly example. The visual pair contains a normal reference and a query image whose teeth are pushed closer together. Baselines produce plausible but weakly grounded defect descriptions, whereas EGVLR follows the unified EDDP schema, localizes the multi-instance defect with normalized 0–1000 boxes, and predicts the correct answer.

5 Conclusion

We presented EGVLR, an evidence-grounded vision–language reinforcement framework for industrial anomaly reasoning on MMAD. EGVLR uses a unified EDDP schema to connect visual evidence, diagnostic logic, bounding-box localization, and final answers across all trainable stages. PVE-FT teaches the model to localize visual evidence and retain the null hypothesis for normal samples. KG-IT injects domain-oriented QA behavior while preserving the same output schema. GS-DPO further aligns answer correctness, spatial validity, BGE-based rationale semantics, and answer–location–rationale consistency. Finally, BGSR converts predicted boxes into dense masks using an external segmentation backend. Experiments and ablations on MMAD show that EGVLR improves reasoning-dependent anomaly QA and reduces spatial false positives. Future work includes stronger multi-instance localization rewards, improved contradiction-aware rationale scoring, and interactive inspection workflows.