Beyond Common Sense: Grounding Logical Anomaly Detection in Inspection Criteria

Yuanze Li[1,2], Shihao Yuan[1,2], Zimeng Zhu[1], Ming Liu[1,2](✉), Wangmeng Zuo[1,2], and Guangming Shi[2,3]

{sqleopop, csshihao, amdreamengz, csmliu}@outlook.com, cswmzuo@gmail.com, gmshi@xidian.edu.cn

[1] Faculty of Computing, Harbin Institute of Technology, Harbin, China, 150001
[2] Pazhou Lab, Huangpu, China
[3] Xidian University, China

Abstract. Multimodal large language models have significantly advanced zero-shot industrial anomaly detection, yet they remain highly constrained when identifying logical anomalies. Fundamentally, logical anomalies arise from discrepancies between inspected products and their predefined constraints. Current detection paradigms typically operate without access to these explicit standards, forcing models to rely on pretrained common sense rather than strict logical rules. To address this limitation, we introduce **SCAN**, a **S**ystematic **C**riteria-driven **AN**omaly inspection framework optimized through a two-stage post-training pipeline. We first employ supervised finetuning based on rule-by-rule Chain-of-Thought (CoT) reasoning to systematically verify explicit inspection criteria. We then apply reinforcement learning equipped with a rule-level credit assignment mechanism to enforce precise anomaly attribution and overcome severe errors caused by sparse feedback. However, such criteria-driven training demands large-scale logical anomaly data paired with explicit inspection rules, yet existing datasets remain largely skewed toward structural defects. To bridge this gap, we propose **FLAW**, a **F**ine-grained **L**ogical **A**nomaly dataset **W**ith criteria. This dataset leverages image synthesis and logical data retrieval to endow images with rich object-level metadata, enabling the construction of diverse inspection standards and logical violations. Extensive experiments on MVTec LOCO demonstrate that our 8B-parameter SCAN model improves upon its base model by 7.8% and establishes a new state of the art, outperforming GPT5.2-Chat and Gemini-3.1-Pro by 4.8% and 1.0%, respectively. Source code and data will be publicly available at https://github.com/tzjtatata/SCAN.

Keywords: Industrial Anomaly Detection · Logical Anomaly Detection · Large Multimodal Model · Multimodal Reasoning · Reinforcement Learning

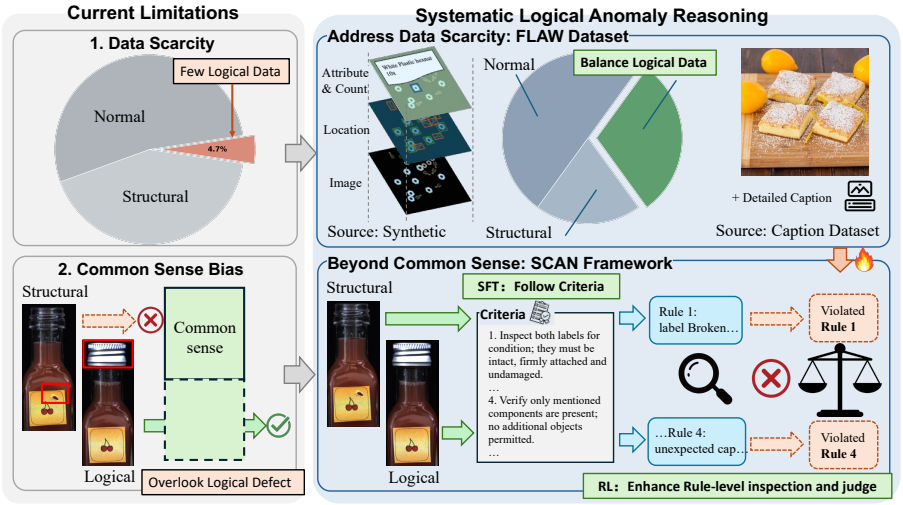

Fig. 1: Overview of our systematic logical anomaly reasoning approach. To address the current limitations of **data scarcity** (top left) and **common sense bias** (bottom left), we propose the **FLAW dataset** to balance logical anomaly distributions (top right), enabling the **SCAN** framework (bottom right). SCAN shifts from implicit common-sense judgments to explicit, rule-level inspection and attribution via SFT and RL, enhancing fine-grained perception.

1 Introduction

Modern industrial production increasingly demands that products satisfy not only structural integrity but also complex logical constraints that govern how components should be combined, configured, and arranged within a product. Violations of these constraints, termed logical anomalies [3], pose a critical yet underexplored challenge for automated inspection. Although large multimodal models (LMMs) have achieved remarkable progress in industrial anomaly detection [12, 22, 23, 44] through their broad visual understanding [16, 17, 27, 30, 31], they exhibit significant limitations when confronted with logical inconsistencies. As illustrated in Fig. 1, we trace these limitations to two primary causes.

First, prevailing detection approaches do not incorporate explicit inspection standards, leaving models to fall back on pretrained common sense as the sole basis for judgment. However, pretrained common sense is inherently misaligned with the stringent and context-specific requirements of industrial manufacturing. As illustrated in Fig. 1, while a fully sealed juice bottle is universally recognized as normal under general assumptions, it constitutes a severe logical defect if it appears before the capping stage. Consequently, relying on common sense alone inevitably yields elevated false-positive and missed-detection rates. To shift the detection paradigm from implicit common sense inference to explicit criteria verification, we introduce **SCAN**, a systematic criteria-driven anomaly inspection framework that formalizes detection as structured verification against explicit inspection criteria. We establish a unified training pipeline that begins with

supervised fine-tuning to train the model to perform systematic, rule-by-rule verification against detailed inspection criteria. We then apply reinforcement learning to further sharpen its reasoning fidelity. To ensure precise anomaly attribution across multiple rules, we develop a rule-level reward that provides fine-grained supervisory signals. However, a naive application of these rewards at the sequence level introduces severe credit assignment errors [15,19], as the model cannot distinguish which reasoning steps are responsible for a particular rule violation. We address this by distributing rule-level rewards at token granularity, enabling accurate credit assignment that substantially improves reasoning fidelity.

Second, deriving criteria-driven inspection at scale exposes a critical data bottleneck. In the most widely used anomaly detection datasets, logical anomalies account for merely 4.7% of MVTec AD and 8.2% of VisA, leaving models trained on these distributions heavily biased toward structural defect patterns. To overcome this limitation, we introduce FLAW, a large-scale dataset constructed through a novel pipeline that explicitly provides the fine-grained relational and logical metadata necessary for this task. FLAW combines two complementary data sourcing strategies, controlled image synthesis and targeted retrieval from vision-language corpora, to produce images equipped with rich object-level metadata such as counts, categories, attributes, and spatial relations. This metadata serves as the foundation for automatically generating diverse inspection criteria and constructing a broad spectrum of logical violations, substantially alleviating the scarcity of logical anomaly data that currently limits progress in this direction.

Extensive experiments on MVTec LOCO [3] demonstrate that SCAN establishes a new state of the art in zero-shot logical anomaly detection, improving upon its base model by 7.8% and surpassing GPT-5.2-Chat [33] and Gemini-3.1-Pro [11] by 4.8% and 1.0%, respectively. The framework generalizes consistently across diverse open-source architectures ranging from 3B to 8B parameters, equipping compact models with capabilities that match or exceed those of the strongest proprietary systems. Ablation studies validate each component of our framework, confirming the complementary contributions of the data sourcing strategies, explicit criteria, and each training stage. Qualitative analyses further show that SCAN yields better visual perception and anomaly attribution, producing reasoning traces that can support downstream industrial applications such as automated defect triage.

In summary, our contributions are listed as follows:

- We propose SCAN, a systematic criteria-driven inspection framework that shifts logical anomaly detection from implicit common sense inference to explicit rule-level verification.
- We introduce FLAW, a large-scale logical anomaly dataset explicitly paired with inspection criteria.
- Comprehensive experiments on MVTec LOCO demonstrate that SCAN establishes a new state of the art in zero-shot logical anomaly detection while delivering precise defect reasoning.

2 Related Works

Industrial anomaly detection datasets. MVTec-AD [4] and VisA [48] are the most widely adopted datasets, providing diverse product categories with anomaly labels and pixel-level masks. PCB-Bank [43] specifically targets the domain of complex printed circuit boards. Recent efforts such as RealIAD [37], MANTA [10], RealIAD Variety [47], IMDD-1M [32], and ReinAD [38] have considerably broadened data scale and diversity. With the rise of multimodal large language models, datasets with fine-grained textual annotations have also emerged. InstructIAD [23] augments VisA with human-written captions. JUDO [18] constructs visual QA data for domain knowledge injection. Anomaly-Instruct-125K [41] provides instruction-tuning data with anomaly reasoning traces. Despite these advances, existing benchmarks still suffer from a scarcity of logical anomalies and lack fine-grained annotations over entire images, which hinders models from systematically detecting violations of semantic or structural rules. Our FLAW dataset addresses both limitations by providing holistic annotations with diverse images that cover logical anomalies alongside structural defects.

Large Multimodal Models for Anomaly Detection. Recent large multimodal models (LMMs) have brought strong visual generalization to industrial anomaly detection. A common paradigm is to integrate domain-specific visual experts, such as CLIP-based detectors in AnomalyGPT [12] and broader expert ensembles in Myriad [22], to supply fine-grained cues to the LMM. Triad [23] further incorporates production-process priors, while Anomaly-OV [41] and SAGE [44] strengthen anomaly perception through anomaly-aware feature selection and preference optimization, respectively. However, these perception-centric approaches work well on locally salient defects while easily missing logical anomalies that violate rules without exhibiting visible damage. Our work addresses this gap with a criteria-driven inspection mechanism that guides the model to attend to holistic image features, enabling more reliable detection of logical anomalies.

Reinforcement Learning for Anomaly Detection. A growing line of work applies reinforcement learning to sharpen LMMs for anomaly detection. AnomalyR1 [5] proposes task-specific rewards for RLVR. OmniAD [45] combines SFT with GRPO [35]. EMIT [13] introduces difficulty-aware rollout filtering and IAD-R1 [21] enriches supervision with location and type prediction rewards. AgentIAD [28] equips the model with agent capabilities for interactive inspection. While these methods improve overall accuracy, they treat the reasoning trace as a monolithic output. In contrast, our approach introduces rule-based fine-grained credit assignment on top of RL, providing step-level supervision over the thinking process and achieving more accurate attribution of logical anomalies.

Rubric-based Reinforcement Learning. Rubric-based reinforcement learning has attracted considerable interest for automating the evaluation of model outputs. Existing methods achieve improved performance by generating rubrics for reliable output quality assessment [26,39] and enabling fine-grained credit assignment [9,40]. We argue that rubrics are equally essential for logical anomaly detection. We therefore incorporate quality inspection rules as naturally occurring rubrics in real-world scenarios to achieve cost-effective fine-grained credit

assignment and enhance the reliability of anomaly detection by large language models.

3 Methods

SCAN formulates zero-shot logical anomaly detection as a criteria-driven inspection task. Given an image and a set of K inspection rules encoding product-specific manufacturing requirements, the model evaluates each rule against the image and produces a structured reasoning trace with per-rule judgments and a final anomaly determination.

We first teach the model to perform systematic rule-by-rule inspection through supervised fine-tuning (§3.1). Reinforcement learning then refines per-rule accuracy by first decomposing image-level feedback into per-rule rewards (§3.2), which a granular credit assignment mechanism then distributes directly to each rule's corresponding tokens to provide precise, token-level supervision (§3.3). Both training stages require large-scale data with fine-grained object-level metadata from which diverse inspection rules and logical violations can be systematically constructed. We address this data requirement through the FLAW dataset (§3.4), built by fusing image synthesis with logical data search.

3.1 Structured Rule Inspection via Supervised Fine-Tuning

The first training stage uses supervised fine-tuning to teach the model to inspect each rule against the image and produce structured reasoning. Each training example pairs an image and a set of inspection rules with a chain-of-thought trace that evaluates every rule in sequence, examining the visual evidence for each rule and concluding whether the specified condition is satisfied or violated.

Directly applying SFT on these examples produces a degeneration pattern we term *prefix interference*. Because rules are evaluated sequentially and each rule's textual description appears in the prompt, the model gradually shifts attention from the image to the rule text itself as the sequence grows. Rather than grounding its reasoning in visual evidence, the model restates the current rule's specification as though it were an observation, effectively confirming every rule by default and causing severe missed detections. This effect intensifies for rules that appear later in the sequence, where the longer preceding context amplifies the model's reliance on textual cues over visual inspection.

We address prefix interference by augmenting the training data with order-varied pairs. For each training example, we fix the anomalous rule at its original position and permute the normal rules that precede it, producing multiple variants that share the same anomaly judgment but differ in prefix context. Since the anomaly judgment remains invariant across permutations while the prefix changes, the model learns to ground each rule evaluation in the image rather than in the preceding output tokens. After this stage, the model reliably produces structured rule-by-rule reasoning, but its per-rule judgments are not always accurate, motivating the reinforcement learning mechanisms described in Sec. 3.2.

3.2 Rule-Level Reward

A natural approach to refining per-rule accuracy is reinforcement learning with verifiable rewards (RLVR [14]). Standard RLVR assigns a single scalar reward based on the final prompt-level prediction. In our multi-rule setting, this creates reward entanglement. The model can flag the wrong rules as violated yet still reach the correct prompt-level conclusion, receiving a positive reward for flawed rule-level reasoning. Conversely, a rollout with largely correct per-rule judgments may receive a negative reward if a single mistake flips the final prediction. The image-level signal conflates the contributions of individual rules, preventing the training signal from correcting specific misjudgments.

We resolve this by decomposing the reward from the image level to the rule level. Because the supervised fine-tuning stage trains the model to produce structured output with a dedicated reasoning segment for each rule, we can extract per-rule judgments from every rollout via simple regular-expression matching. Let $\hat{y}_i^k$ denote the judgment extracted from rollout $\mathbf{o}_i$ for rule c_k, and let y^k denote the ground-truth label. The rule-level reward is defined as

$$r_i^k = \mathbb{1}[\hat{y}_i^k = y^k] \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function. Each rule in each rollout now receives an independent binary reward, disentangling the per-rule feedback from the image-level signal. This decomposition is necessary but not sufficient on its own. In standard GRPO, advantages are still computed per rollout and applied uniformly to all tokens, so the per-rule structure of the reward is lost when it reaches the policy gradient [25]. The following section addresses this bottleneck.

3.3 Rule-level Credit Assignment

The rule-level rewards from Eq. (1) identify which rules are judged incorrectly in each rollout, but standard GRPO computes a single advantage per rollout and broadcasts it to all tokens uniformly. Applying rule-specific rewards at the sequence level therefore still leaves the model unable to locate which reasoning steps caused a particular rule's failure. To translate the per-rule reward decomposition into effective training gradients, each rule's feedback must reach only the tokens that produced that rule's judgment. We achieve this by restructuring the advantage computation in GRPO so that each token receives an advantage determined solely by the rule segment it belongs to. The complete procedure is given in Algorithm 1.

For each prompt, we sample G rollouts from the current policy. From each rollout, we extract per-rule judgments and identify segment boundaries using regular-expression matching on the structured output produced by the SFT stage. Each judgment is independently scored against its ground-truth label to obtain the per-rule reward r_i^k defined in Eq. (1). We then normalize rewards at the rule level rather than the rollout level. For each rule k, the G rewards $\{r_1^k, \ldots, r_G^k\}$ across rollouts form a single normalization group, yielding a per-rule

Algorithm 1 Granular Credit Assignment for GRPO

Require: Prompt q with K rules $\{c_k\}_{k=1}^{K}$, ground-truth labels $\{y^k\}_{k=1}^{K}$
Sample G rollouts $\{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$ from $\pi_{\theta_{\text{old}}}(\cdot \mid q)$
for each rollout $\mathbf{o}_i$, $i = 1, \ldots, G$ **do**
    Extract per-rule judgments $\{\hat{y}_i^1, \ldots, \hat{y}_i^K\}$ and segment boundaries $\{S_i^1, \ldots, S_i^K\}$
    for each rule $k = 1, \ldots, K$ **do**
        $r_i^k \leftarrow \mathbb{1}[\hat{y}_i^k = y^k]$ ▷ Per-rule reward
    end for
end for
for each rule $k = 1, \ldots, K$ **do**
    $\mu^k \leftarrow \frac{1}{G}\sum_{j=1}^{G} r_j^k, \quad \sigma^k \leftarrow \text{std}(\{r_j^k\}_{j=1}^{G})$
    for each rollout $i = 1, \ldots, G$ **do**
        $\hat{A}_i^k \leftarrow (r_i^k - \mu^k) \,/\, \sigma^k$ ▷ Rule-level advantage
    end for
end for
for each rollout $\mathbf{o}_i$ **do**
    for each rule $k = 1, \ldots, K$ **do**
        for each token $t \in S_i^k$ **do**
            $\hat{A}_{i,t} \leftarrow \hat{A}_i^k$ ▷ Token-level assignment
        end for
    end for
end for
Update π_θ with GRPO using rule-varying advantage tensor $\hat{A}_{i,t}$

advantage

$$\hat{A}_i^k = \frac{r_i^k - \text{mean}(r_j^k)}{\text{std}(r_j^k)}, \tag{2}$$

for each rollout. Finally, each token is assigned the advantage of the rule segment it belongs to, producing a token-varying advantage $\hat{A}_{i,t}$ that replaces the uniform per-rollout advantage in standard GRPO. A rollout that correctly judges rule k receives a positive advantage for that rule's tokens regardless of errors elsewhere, while an incorrect judgment receives a negative advantage regardless of other successes. The policy gradient for each segment is therefore driven solely by the correctness of that segment's own conclusion.

3.4 FLAW Dataset

The training stages above require large-scale data covering diverse logical anomaly types, yet existing datasets are heavily skewed toward structural defects and lack the fine-grained relational metadata needed for automated rule construction. We address this gap with FLAW, a large-scale logical anomaly dataset explicitly paired with inspection criteria. FLAW combines controlled image synthesis and targeted retrieval from vision-language corpora to produce images equipped with rich object-level metadata, from which diverse inspection rules and logical violations are systematically generated.

Synthesis pipeline. We extract product instances via segmentation masks and compose them onto diverse backgrounds, obtaining complete knowledge of every object in the scene, including its category, count, visual appearance, and bounding-box location. With this metadata, we prompt Qwen3-VL-32B [1] with the full annotation alone to generate K inspection rules per image, each targeting a specific type such as quantity constraints, object-type verification, foreign-object detection, or structural consistency. Among the K rules, N are constructed to deliberately contradict the ground truth while the rest remain consistent with it. Because the metadata is known by construction, every label is unambiguously determined. Chain-of-thought reasoning is then generated by prompting Qwen3-VL-32B with both the image and the annotation, and verified using Qwen3-30B-A3B [42] against ground-truth labels, with disagreements discarded. We apply this pipeline to MANTA [10] and RealIAD [37], retaining products with structural anomalies so that the subset covers both defect types.
Collection pipeline. Cut-and-paste synthesis cannot produce anomalies involving attribute variations or component relationships. To cover these patterns, we retrieve images from ShareGPT4V [7], ALLaVA [6], and Denseworld-1m [20], whose detailed captions encode object attributes, spatial arrangements, and part-whole relationships sufficient for more diverse logical data. We filter candidates with Qwen3-30B-A3B, retaining only images whose captions describe object attributes, component relationships, or structural irregularities. For each selected image, Qwen3-VL-30B generates multiple inspection rules paired with binary labels, followed by chain-of-thought reasoning that evaluates each rule in sequence. From approximately 100K images, this process yields roughly 10K training examples covering anomaly types beyond the reach of synthesis.
Dataset composition. FLAW also includes a third subset from real anomaly datasets in InstructIAD [23,24] and a three-class subset of MVTec-AD [4], for which we manually compose inspection rules encoding domain-specific criteria and generate chain-of-thought reasoning following the same verification procedure.

Together, the three subsets provide balanced coverage of structural and logical anomalies from both synthetic and real-world sources, with every example grounded in rule-by-rule reasoning traces. Please refer to Appendix A for statistics on the FLAW datasets.

4 Experiments

4.1 Implementation details

Training details. We conduct experiments on three vision-language models of varying scales, namely Qwen2.5-VL-3B, Qwen2.5-VL-7B [2], and Qwen3-VL-8B [1]. Our training pipeline comprises two stages: supervised fine-tuning followed by a successive reinforcement learning stage. In SFT, we perform supervised fine-tuning using the LLaMA-Factory [46] framework. We freeze the vision encoder and update all remaining parameters. The maximum image resolution

is set to 1024×1024. We adopt the AdamW optimizer with a learning rate of 1e-5 and a warmup ratio of 0.03. The batch size is set to 32. The 3B model is trained on 2 NVIDIA A100 GPUs, while the 7B and 8B models are trained on 4 A100 GPUs, each taking approximately 6 hours. In the RL stage, we apply reinforcement learning based on VeRL [36]. We use a batch size of 16 with a rollout size of 8, a learning rate of 1.0e-6, and a KL penalty coefficient of 1e-2. For rollout generation, we set the temperature to 1.0 and top-p to 1.0, with top-k sampling disabled. This stage is trained on 6,400 samples for 1 epoch. The RL stage takes approximately 8 hours on the same hardware configuration as SFT.
Evaluation details. We conduct experiments on the MVTec-LOCO [3] dataset following the zero-shot setting that is standard in industrial anomaly detection, where the model has no prior exposure to any of the product categories encountered at test time. During inference, we set the temperature to 0.01 and top-k to 1 to produce near-deterministic outputs, while all baseline methods are evaluated with their officially recommended hyperparameters for a fair comparison. Beyond simply checking whether the predicted anomaly label is correct, we further require the model to provide a faithful explanation for its prediction. To this end, we design an evaluation protocol comprising two distinct metrics: binary metrics and scoring metrics. The binary metrics solely evaluate the correctness of the final detection outcome. In contrast, the scoring metrics employ an LLM-as-judge setup to assess the model's reasoning capabilities. Concretely, we supply the ground-truth annotations from MVTec-LOCO to Qwen3-30B-A3B [42], which evaluates the responses. Under the scoring metrics, a prediction is counted as correct only if the model accurately identifies the anomaly and provides valid reasoning, including the specific cause of the anomaly. This strict evaluation penalizes models that happen to arrive at the right answer through spurious reasoning, offering a more reliable measure of genuine anomaly understanding.
Baselines. We evaluate our method by training on three base models of varying scales, namely Qwen2.5-VL-3B, Qwen2.5-VL-7B, and Qwen3-VL-8B, and compare the resulting models against their respective pretrained counterparts to demonstrate the effectiveness of our training recipe. In addition, we include several domain-specific large multimodal models for industrial anomaly detection, including AnomalyGPT-7B [12], Myriad-7B [22], and Triad-7B [23], as additional baselines. Beyond these direct comparisons, we further benchmark our trained models against a set of significantly larger general-purpose vision-language models, including Qwen2.5-VL-72B, Qwen3-VL-32B, GPT-5.2-chat [33], and Gemini-3.1-Pro-Preview [11].

4.2 Main results

Quantitative results. As detailed in Table 1, our method enables compact architectures to achieve performance that rivals or exceeds the most capable proprietary and large-scale open-source systems. Our 8B model comprehensively outperforms the significantly larger Qwen3-VL-32B and GPT-5.2-chat across all scoring metrics, leading by 2.18% and 4.85% in accuracy and by 3.02% and 3.79% in F1, respectively. At the 7B scale, our model surpasses Qwen2.5-VL-72B

Table 1: Main results on the MVTec-LOCO dataset. We compare our method against SOTA large-scale models and highlight the performance gain over respective baselines. **Bold** and underline indicate the best and second-best performance among all models. Models marked with † focus on visual perception rather than reasoning, and therefore do not provide scoring metrics.

Method	Binary Metrics (↑)				Scoring Metrics (↑)			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Large-Scale & Proprietary Models								
Qwen2.5-VL-72B	60.99	87.30	35.62	47.44	52.69	78.38	20.10	29.87
Qwen3-VL-32B	70.90	88.27	52.53	65.62	62.40	82.63	36.97	50.48
GPT-5.2-chat	74.67	90.95	61.05	71.72	59.73	76.90	37.34	49.71
Gemini-3.1-Pro	74.98	74.91	84.26	**78.66**	63.64	67.62	**71.15**	**53.78**
Domain-specific Models for IAD								
AnomalyGPT-7B†	52.43	53.63	88.35	65.95	-	-	-	-
Myriad-7B†	54.27	54.37	**97.90**	69.86	-	-	-	-
Triad-ov-7B	58.06	90.03	27.36	38.45	45.03	54.64	7.93	13.45
Comparison with Baselines								
Qwen2.5-VL-3B	50.10	65.70	15.00	21.30	44.80	46.60	5.00	8.80
Ours (3B)	62.70	75.60	48.70	58.00	49.10	56.30	23.20	32.20
Gain vs. Base	**+12.60**	**+9.90**	**+33.70**	**+36.70**	**+4.30**	**+9.70**	**+18.20**	**+23.40**
Qwen2.5-VL-7B	58.43	83.95	32.53	44.11	48.81	67.98	14.16	22.39
Ours (7B)	69.20	81.70	55.80	66.20	57.10	69.50	33.60	45.00
Gain vs. Base	**+10.77**	-2.25	**+23.27**	**+22.09**	**+8.29**	**+1.52**	**+19.44**	**+22.61**
Qwen3-VL-8B	62.98	88.41	35.08	48.99	56.73	74.10	24.57	35.55
Ours (8B)	**75.16**	**93.47**	59.11	70.10	**64.58**	**90.05**	39.83	53.50
Gain vs. Base	**+12.18**	**+5.06**	**+24.03**	**+21.11**	**+7.85**	**+15.95**	**+15.26**	**+17.95**

across the board (57.10% *vs*. 52.69% scoring accuracy, 45.00% *vs*. 29.87% scoring F1). Among domain-specific models, Triad-ov-7B achieves the highest binary accuracy of 58.06% while the others suffer from severe overkill and poor precision. Constrained by logical data scarcity, this top performance among existing LMM-based IAD methods merely matches the general-purpose Qwen2.5-VL-7B baseline at 58.43%, exposing the fragility of existing datasets for logical anomaly detection. We acknowledge that inherent constraints in parameter count and training data scale leave our models with a gap in raw visual perception relative to Gemini-3.1-Pro, reflected in lower scoring recall (39.83% *vs*. 71.15%). Nevertheless, because our framework trains the model to systematically verify each inspection rule rather than make guesses, false positives are substantially reduced (90.05% *vs*. 67.62% scoring precision), and the resulting scoring accuracy slightly exceeds that of Gemini-3.1-Pro (64.58% *vs*. 63.64%).

Turning to direct baselines, SCAN delivers substantial and consistent improvements across all three model scales, with scoring F1 improving by +23.40%, +22.61%, and +17.95% at 3B, 7B, and 8B, respectively. This consistency across architectures of varying capacity confirms that the framework provides a general recipe rather than scale-dependent tuning. Recall contributes the largest share of these gains (+18.20%, +19.44%, and +15.26%). Alongside improved detection, SCAN raises the fidelity of model reasoning at every scale. Scoring accuracy, which credits a prediction only when both the detection outcome and its supporting explanation are judged correct, improves by +4.30%, +8.29%,

and +7.85% at 3B, 7B, and 8B. The strongest improvement comes from our 8B model, whose scoring precision reaches 90.05% (+15.95% over the base), indicating that nearly every flagged anomaly is supported by a correct, rule-grounded explanation. The rule-level reward in the RL stage drives this improvement by ensuring that individual rule judgments, not just the final label, receive targeted optimization.

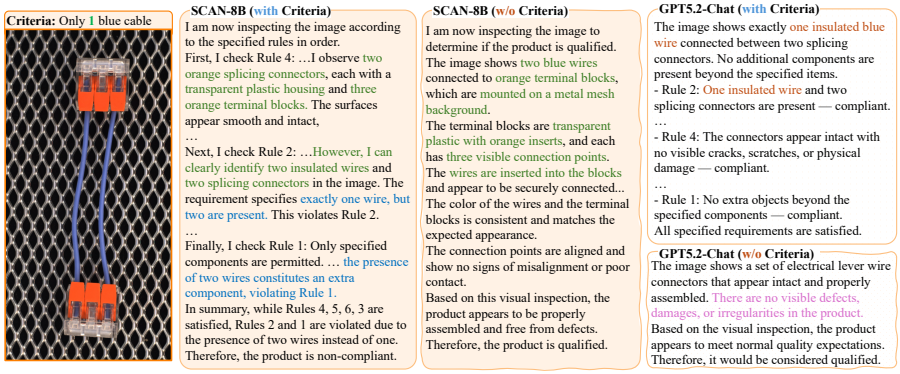

Fig. 2: Qualitative comparison between SCAN and GPT5.2-Chat. (1) **With Criteria**, SCAN demonstrates superior fine-grained visual recognition by accurately identifying both cables, whereas GPT5.2-Chat misses the extra cable. (2) **Without Criteria**, although both models reach incorrect final conclusions, SCAN exhibits a significantly more systematic logical analysis by examining the attributes, connection states, colors, and quantities of individual components.

Qualitative results. We present a qualitative comparison in Fig. 2. When provided with the criteria, SCAN-8B correctly counts two wires and flags the violation, while GPT5.2-Chat perceives only one wire and incorrectly passes the sample. This confirms that SCAN sharpens fine-grained visual discrimination in cluttered industrial scenes beyond what the strongest proprietary model achieves.

Without criteria, both models default to a normal verdict, underscoring that explicit criteria are indispensable for reliable rule-based quality inspection. However, the two models differ markedly in their reasoning traces. SCAN-8B still produces a structured, component-level analysis that accurately identifies each connector type, counts connection points, verifies insertion status, and checks color consistency. GPT5.2-Chat, by contrast, not only misperceives the visual content but also overlooks most components in the scene, reducing its inspection to a cursory check for structural damage. This indicates that SCAN has internalized a systematic verification process for both logical and structural anomalies, even in the absence of explicit rules.

Table 2: Ablation study on the effectiveness of criteria across VLM architectures. Results are evaluated on MVTec-LOCO (scoring metrics). **Bold** and underline denote the best and second-best results for configurations with criteria (✓) and without criteria (✗), respectively. Green values show the absolute improvement (Δ).

Model	Crit.	Acc. (↑)	Prec. (↑)	Rec. (↑)	F1 (↑)
Qwen2.5-VL-7B	✗	36.79	29.48	12.62	16.14
	✓	48.81	67.98	14.16	22.39
		(+12.0)	(+38.5)	(+1.5)	(+6.3)
Qwen2.5-VL-72B	✗	39.41	33.28	10.40	14.97
	✓	52.69	78.38	20.10	29.87
		(+13.3)	(+45.1)	(+9.7)	(+14.9)
Qwen3-VL-8B	✗	45.77	47.50	21.07	27.56
	✓	56.73	74.10	24.57	35.55
		(+11.0)	(+26.6)	(+3.5)	(+8.0)
Qwen3-VL-32B	✗	41.28	42.97	25.77	30.47
	✓	62.40	82.63	36.97	50.48
		(+21.1)	(+39.7)	(+11.2)	(+20.0)
GPT5.2-chat	✗	40.97	41.28	7.92	12.38
	✓	59.73	76.90	37.34	49.71
		(+18.8)	(+35.6)	(+29.4)	(+37.3)
Ours (Qwen3-VL-8B)	✗	**51.48**	**61.14**	**35.30**	**42.90**
	✓	**64.58**	**90.05**	**39.83**	**53.50**
		(+13.1)	(+28.9)	(+4.5)	(+10.6)

4.3 Ablation study

Effect of Criteria. We present an ablation study in Table 2 to evaluate the effectiveness of incorporating explicit inspection criteria across various vision-language model architectures. The results reveal a clear scaling behavior where larger models within the same family exhibit greater adaptability to product inspection rules. When equipped with explicit criteria, larger variants consistently achieve more substantial improvements across all evaluation metrics. Specifically, supplying criteria raises precision by at least 26.6% on every evaluated model, peaking at an absolute increase of 45.1% for Qwen2.5-VL-72B and 39.7% for Qwen3-VL-32B. This disproportionate enhancement in precision indicates that increased model capacity translates to a more profound comprehension of textual rules, allowing the criteria-driven approach to establish a robust operational boundary that strictly limits false positives and reduces overkill compared to smaller baselines.

Similarly, the improvement in recall exhibits a strong correlation with the inherent perceptual capabilities associated with larger model scales. As the base architecture scales from Qwen3-VL-8B to Qwen3-VL-32B, the integration of criteria yields progressively larger recall gains, shifting from a modest 3.5% increase to a substantial 11.2% improvement. This trend extends to GPT5.2-chat, which

Table 3: Effect of Dataset Components. Ablation study of our proposed data pipelines on the MVTec-LOCO dataset during SFT. The baseline model is Qwen2.5-VL-7B. Small colored values indicate the improvement (Δ) over the baseline (green for gain, red for loss). **Bold** and underline denote the largest and second-largest gains in Δ columns, respectively.

Method	Syn.	Col.	Scoring Metrics (↑)			
			Acc.	Prec.	Rec.	F1
Qwen2.5-VL-7B	✗	✗	48.81	67.98	14.16	22.39
+ Synthetic Pipeline	✓	✗	49.35 +0.54	57.26 -10.72	22.97 +8.81	32.60 +10.21
+ Collection Pipeline	✗	✓	46.65 -2.16	55.99 -11.99	23.09 +8.93	31.83 +9.44
FLAW (Ours)	✓	✓	50.70 **+1.89**	61.83 **-6.15**	23.61 **+9.45**	33.98 **+11.59**

records a striking 29.4% boost in recall when provided with explicit criteria. Such scaling confirms that while explicit rules effectively guide models to perform exhaustive logical checks and prevent the omission of subtle anomalies, this advantage remains fundamentally constrained by visual grounding capacity. Notably, when operating without criteria, the recall of GPT5.2-chat severely degrades to 7.92%, dropping well below the 12.62% achieved by the smaller Qwen2.5-VL-7B. This stark contrast emphasizes that without explicit textual guidance, even highly capable reasoning engines struggle to visually isolate logical defects.

Building upon these scaling insights, our proposed framework establishes state-of-the-art zero-shot logical anomaly detection performance without requiring massive parameters. Compared to the standard Qwen3-VL-8B baseline, our model demonstrates significantly stronger rule comprehension and perceptual acuity at the exact same scale. The extensive and fine-grained visual grounding cultivated by the FLAW dataset endows our model with exceptional perception, driving our absolute recall to 39.83%, which greatly eclipses the 24.57% of the criteria-equipped base model. Simultaneously, our two-stage training paradigm successfully bridges the gap between raw visual perception and structured reasoning. This framework-driven enhancement in rule comprehension propels our precision to 90.05%, reflecting a robust 28.9% improvement over our baseline without criteria, ultimately ensuring highly accurate attribution of logical anomalies in complex inspection scenarios.

Reliability of Scoring Metrics. We conduct a user study involving ten human annotators to evaluate the reasoning process of SCAN and validate the reliability of large language model evaluation (scoring metrics). We then compared these human annotations with the assessments generated by Qwen3-30B-A3B. The calculated Cohen's κ between the human annotators and the model reaches 0.841, which closely approaches the inter-annotator agreement of 0.877. This close alignment demonstrates that utilizing Qwen3-30B-A3B as an evaluator is a reliable approach for this specific task. Table 4 presents further robustness analysis of the scoring metrics in three distinct dimensions. These dimensions encompass the choice of the evaluator LLMs, the prompt design, and the ground truth anomaly expressions. Regarding the model dimension specifically, we incorporated Deepseek-v4-pro [8] and Qwen3.5-27B [34] into our experiments along-

Table 4: LLM-as-judge Reliability across 3 dimensions.

Setting	**Acc.**	**F1**	**Prec.**	**Rec.**
Judges	65.31 ± 0.65	54.52 ± 0.91	90.56 ± 0.45	41.10 ± 1.14
Prompts	63.97 ± 0.63	52.16 ± 1.33	90.02 ± 0.18	38.72 ± 1.12
Wording of Explanation	64.20 ± 0.35	52.54 ± 0.85	90.08 ± 0.15	39.07 ± 0.68

Table 5: Ablation study on our training framework. We evaluate the incremental impact of each stage on the MVTec-LOCO dataset. The baseline is Qwen3-VL-8B. **Bold** and underline denote the best and second-best results, respectively. SR: Sequence-level Reward; RR: Rule-level Reward; RCA: Rule-level Credit Assignment.

Method	Stages		Binary Metrics (%)				Scoring Metrics (%)			
	SFT	RL	Acc ↑	Prec ↑	Rec ↑	F1 ↑	Acc ↑	Prec ↑	Rec ↑	F1 ↑
(1) Baseline			62.98	88.41	35.08	48.99	56.73	74.10	24.57	35.55
(2) + SFT	✓		71.90	88.02	56.65	67.16	59.54	80.14	34.09	46.68
(3) + SR	✓	✓	**75.61**	85.24	**67.19**	**73.81**	60.78	76.73	**40.10**	51.92
(4) + RR	✓	✓	74.73	<u>91.31</u>	<u>61.07</u>	<u>70.82</u>	<u>62.68</u>	<u>86.91</u>	39.14	<u>52.20</u>
(5) + RR&RCA	✓	✓	<u>75.16</u>	**93.47**	59.11	70.10	**64.58**	**90.05**	<u>39.83</u>	**53.50**

side the original Qwen3-30B-A3B. The empirical results indicate that the scoring metrics exhibit low variance across all three dimensions.

Effect of Dataset Components. Table 3 evaluates the contribution of each data pipeline component using the Qwen2.5-VL-7B baseline in the SFT stage. All data variants consistently improve recall by approximately 9%, confirming that FLAW broadly strengthens logical defect sensitivity regardless of the data source. However, precision drops across all settings from the baseline's 67.98%, reflecting the absence of an explicit attribution mechanism during SFT that leads to over-prediction of anomalies. The two data sources prove complementary when combined. The full FLAW dataset achieves the highest F1 of 33.98% while limiting the precision loss to only 6.15%, compared to over 10% when either pipeline is used alone.

Effect of Supervised Fine-Tuning in SCAN. The FLAW dataset effectively enhances the model's ability to recognize logical anomalies through supervised fine-tuning. As shown in Table 5, the SFT stage (row 2) lifts scoring recall from 24.57% to 34.09%, a gain consistent with the improvement observed on Qwen2.5-VL-7B in Table 1. Reinforcement learning without explicit supervision over the reasoning process actually hurts judgment quality. When only a sequence-level reward (SR, row 3) is applied, recall continues to rise to 40.10%, yet scoring precision drops by 3.41% from 80.14% to 76.73%. This indicates that rewarding the model solely on its final answer encourages indiscriminate positive predictions, undermining the reliability of its rule-by-rule analysis.

Effect of Reinforcement Learning in SCAN. Providing feedback at the rule level effectively resolves the precision degradation observed under SR. Compared to SR, Rule-level Reward (RR, row 4) raises scoring precision from 76.73% to 86.91% while recall only marginally decreases from 40.10% to 39.14%, demonstrating that supervising intermediate reasoning steps sharpens judgment accuracy with negligible impact on detection sensitivity. Building on RR, Rule-level Credit Assignment (RCA, row 5) further pushes precision to 90.05%, an additional 3% gain over RR and nearly 10% above the SFT baseline. Meanwhile, RCA maintains recall at 39.83%, virtually matching the best recall achieved by SR. These results confirm that fine-grained credit assignment prevents the model from being positively reinforced for flawed reasoning chains, ultimately achieving the best overall scoring F1 of 53.50%.

5 Conclusion

This work reveals that the central obstacle to reliable logical anomaly detection lies not in model scale or perceptual capacity but in the absence of explicit inspection standards that anchor visual reasoning to well-defined manufacturing rules. Guided by this insight, we developed SCAN, a criteria-driven framework whose two-stage training pipeline progressively instills structured rule verification. Supervised fine-tuning teaches the model to reason against predefined criteria on a rule-by-rule basis, while reinforcement learning with token-granularity rule-level credit assignment ensures that each individual rule judgment receives targeted optimization rather than being obscured by a single sequence-level reward. Because this paradigm demands logical anomaly data at a scale that existing benchmarks cannot provide, we constructed FLAW, a dataset that pairs controlled image synthesis with targeted retrieval to produce images with rich component-level metadata from which diverse inspection criteria and logical violations are automatically derived. On MVTec LOCO, SCAN-8B surpasses both GPT-5.2-Chat and Gemini-3.1-Pro in scoring accuracy and F1, and consistent improvements observed from 3B to 8B parameters confirm that these gains originate from the training paradigm itself rather than architecture-specific tuning. More broadly, these findings suggest that grounding visual inspection in explicit, verifiable criteria offers a general principle whose applicability likely extends beyond logical anomalies to medical and video anomaly detection at large.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2023YFA1008500 and the Heilongjiang Postdoctoral Fund under Grant No. LBH-Z25148.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S.,

Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. International Journal of Computer Vision **130**(4), 947–969 (2022)
4. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: 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 (2019)
5. Chao, Y., Liu, J., Tang, J., Wu, G.: Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection. arXiv preprint arXiv:2504.11914 (2025)
6. Chen, G.H., Chen, S., Zhang, R., Chen, J., Wu, X., Zhang, Z., Chen, Z., Li, J., Wan, X., Wang, B.: Allava: Harnessing gpt4v-synthesized data for lite vision-language models. arXiv preprint arXiv:2402.11684 (2024)
7. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
8. DeepSeek-AI: Deepseek-v4: Towards highly efficient million-token context intelligence (2026)
9. Ding, L.: Adarubric: Task-adaptive rubrics for reliable llm agent evaluation and reward learning. arXiv preprint arXiv:2603.21362 (2026)
10. Fan, L., Fan, D., Hu, Z., Ding, Y., Di, D., Yi, K., Pagnucco, M., Song, Y.: Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
11. Google DeepMind: Gemini 3.1 pro: A smarter foundation for complex reasoning and agentic workflows. Tech. rep., Google DeepMind (February 2026), technical Report
12. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 1932–1940 (2024)
13. Guan, W., Lan, J., Cao, J., Tan, H., Zhu, H., Wang, W.: Emit: Enhancing mllms for industrial anomaly detection via difficulty-aware grpo. arXiv preprint arXiv:2507.21619 (2025)
14. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. Guo, Y., Xu, L., Liu, J., Dan, Y., Qiu, S.: Segment policy optimization: Effective segment-level credit assignment in rl for large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
16. Guo, Z., Liu, M., Wang, Q., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Integrating visual interpretation and linguistic reasoning for geometric problem solving. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 3988–3998 (2025)
17. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. arXiv preprint arXiv:2510.12798 (2025)
18. Kang, H., Lee, W., Kim, J., Park, H.: Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa. In: The Fourteenth International Conference on Learning Representations (2026)
19. Kazemnejad, A., Aghajohari, M., Portelance, E., Sordoni, A., Reddy, S., Courville, A., Le Roux, N.: Vineppo: Accurate credit assignment in rl for llm mathematical reasoning. In: The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24 (2024)
20. Li, X., Zhang, T., Li, Y., Yuan, H., Chen, S., Zhou, Y., Meng, J., Sun, Y., Xu, S., Qi, L., et al.: Denseworld-1m: Towards detailed dense grounded caption in the real world. arXiv preprint arXiv:2506.24102 (2025)
21. Li, Y., Cao, Y., Liu, C., Xiong, Y., Dong, X., Huang, C.: Iad-r1: Reinforcing consistent reasoning in industrial anomaly detection. arXiv preprint arXiv:2508.09178 (2025)
22. Li, Y., Wang, H., Yuan, S., Liu, M., Zhao, D., Guo, Y., Xu, C., Shi, G., Zuo, W.: Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. arXiv preprint arXiv:2310.19070 (2023)
23. Li, Y., Yuan, S., Wang, H., Li, Q., Liu, M., Xu, C., Shi, G., Zuo, W.: 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. pp. 21917–21926 (2025)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
25. Liu, S.Y., Dong, X., Lu, X., Diao, S., Belcak, P., Liu, M., Chen, M.H., Yin, H., Wang, Y.C.F., Cheng, K.T., et al.: Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. arXiv preprint arXiv:2601.05242 (2026)
26. Liu, T., Xu, R., Yu, T., Hong, I., Yang, C., Zhao, T., Wang, H.: Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. arXiv preprint arXiv:2510.07743 (2025)
27. Liu, Y., Chen, L., Liu, J., Zhu, M., Zhong, Z., Yu, B., Jia, J.: Visurf: Visual supervised-and-reinforcement fine-tuning for large vision-and-language models. arXiv preprint arXiv:2510.10606 (2025)
28. Miao, J., Du, P., Liu, Y., Wang, Y., Wang, Y.: Agentiad: Tool-augmented single-agent for industrial anomaly detection. arXiv preprint arXiv:2512.13671 (2025)
29. Nakata, H., Zou, Y., Sakai, S., Maeda, S., Gu, C., Wei, Y., Gao, S., Zhang, C.: Vid-ad: A dataset for image-level logical anomaly detection under vision-induced distraction. arXiv preprint arXiv:2603.13964 (2026)
30. Ni, M., Fan, Y., Zhang, L., Zuo, W.: Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. In: The Thirteenth International Conference on Learning Representations (2025)
31. Ni, M., Yang, Z., Li, L., Lin, C.C., Lin, K., Zuo, W., Wang, L.: Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. arXiv preprint arXiv:2505.19702 (2025)
32. Ni, T., Chen, Z., Yang, Y.: Towards open-vocabulary industrial defect understanding with a large-scale multimodal dataset. arXiv preprint arXiv:2512.24160 (2025)

33. OpenAI: GPT-5.2 System Card: Advances in Reasoning and Multi-step Planning. arXiv preprint arXiv:2601.03267 (2026)
34. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026)
35. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
36. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)
37. Wang, C., Wu, W., Zheng, J., et al.: Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
38. Wang, X., Zhuo, J., You, Z., Tan, Z., Yu, Y., Wang, S., Le, X.: Reinad: Towards real-world industrial anomaly detection with a comprehensive contrastive dataset. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
39. Wang, Y., Xia, X., Wu, X., Zhai, X., Liu, N.: Learnable assessment skills for llm-based automated scoring: Rubric construction via iterative optimization. arXiv preprint arXiv:2605.29274 (2026)
40. Xie, W., Zhao, H., Liu, W., Zhu, Y., Chen, L., Ye, M., Chen, Z., Xu, Y., Dong, S., Wang, Z., et al.: Step-wise rubric rewards for llm reasoning. arXiv preprint arXiv:2605.17291 (2026)
41. Xu, J., Lo, S.Y., Safaei, B., Patel, V.M., Dwivedi, I.: Towards zero-shot anomaly detection and reasoning with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20370–20382 (2025)
42. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
43. Yao, H., Liu, M., Yin, Z., Yan, Z., Hong, X., Zuo, W.: Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In: European Conference on Computer Vision (ECCV). pp. 1–17. Springer (2024)
44. Zang, G., Li, X., Di, D., Nie, L., Zhan, D., Song, Y., Fan, L.: Sage: A visual language model for anomaly detection via fact enhancement and entropy-aware alignment. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5030–5039 (2025)
45. Zhao, S., Lin, Y., Han, L., Zhao, Y., Wei, Y.: Omniad: Detect and understand industrial anomaly via multimodal reasoning. arXiv preprint arXiv:2505.22039 (2025)
46. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, Bangkok, Thailand (2024)

47. Zhu, W., Wang, C., Gao, B.B., Zhang, J., Jiang, G., Hu, J., Gan, Z., Wang, L., Zhou, Z., Cheng, L., et al.: Real-iad variety: Pushing industrial anomaly detection dataset to a modern era. arXiv preprint arXiv:2511.00540 (2025)
48. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision (ECCV) (2022)