Enhancing Large Vision–Language Models for Multimodal Defect Detection via SFT–GRPO Reinforcement Learning

Hung Viet Nguyen
Department of Digital Anti-Aging Healthcare
INJE University
Kimhae, 50834, Rep. of Korea
nviethung1998@live.inje.ac.kr

Hyojin Park
Gyeongnam Intelligence Innovation Center (GIIC)
Kyungnam University
Changwon, 51767 Rep. of Korea
gaiaphj@gmail.com

Namhyun Yoo
Department of Computer Engineering
Kyungnam University
Changwon, 51767 Rep. of Korea
hyun43@kyungnam.ac.kr

Jinhong Yang
Department of Medical IT
INJE University
Kimhae, 50834, Rep. of Korea
jinhong@inje.ac.kr

Abstract— **Large Vision Language Models (LVLMs) offer strong visual reasoning capabilities but their direct application to industrial defect inspection remains limited due to domain complexity, diverse defect modes, and the need for structured reporting. This paper presents a unified fine-tuning framework that combines Supervised Fine-Tuning (SFT) with Group Relative Policy Optimization (GRPO) to adapt open-source LVLMs for multimodal defect inspection in Liquefied Natural Gas (LNG) tank manufacturing. Using a balanced dataset derived from 188,631 inspection images, the proposed method enables each LVLM to perform joint defect localization, attribute prediction, and automatic generation of structured JSON inspection reports. The GRPO stage incorporates verifiable reward signals that enforce JSON validity, schema compliance, bounding box accuracy, and metadata consistency. Experimental results on four LVLM architectures demonstrate substantial performance gains, with mean Average Precision at IoU 0.5 improving from 35–39 percent to 84–89 percent and F1 scores for defect labels increasing from approximately 14 percent to above 84 percent. The best-performing model, Qwen2.5-VL-7B, achieves 88.77 percent mAP at IoU 0.5 and over 90 percent F1 in key metadata fields. These findings indicate that SFT and GRPO provide complementary benefits, enabling LVLMs to deliver accurate, interpretable, and computationally efficient inspection for next-generation manufacturing environments.**

Keywords— Large Vision Language Models, defect inspection, reinforcement learning, supervised fine-tuning, GRPO

I. INTRODUCTION

Quality assurance is a core determinant of manufacturing competitiveness. Classical methodologies such as Total Quality Management, Six Sigma, and Lean Manufacturing established systematic approaches for reducing variability and improving process reliability [1]. With the emergence of Industry 4.0 and Cyber–Physical Systems (CPS), manufacturers increasingly rely on interconnected sensors, automation, and data-driven decision-making to enhance production quality [2]. Yet despite these advances, ensuring consistent product quality remains challenging, particularly in complex or safety-critical production environments.

Defect inspection plays a critical role in quality assurance by enabling early anomaly detection and reducing scrap, rework, and downstream failures [3], [4]. Traditional inspection processes depend heavily on human operators, whose performance is limited by fatigue, subjectivity, and inconsistency. Reported misclassification rates can reach 15–20% even in precision manufacturing, with higher error rates observed in high-throughput and safety-critical industries such as oil and gas [5], [6]. These limitations have driven the adoption of computer-vision-based inspection systems using Machine Learning (ML) and Deep Learning (DL). Convolutional Neural Networks (CNNs), including YOLO-based detectors [7], are now widely deployed in industrial inspection tasks. However, their effectiveness degrades under changing defect patterns, variable illumination, or process drift, and their reliance on large labeled datasets makes adaptation costly, particularly for small and medium-sized enterprises (SMEs) [8].

Large Vision–Language Models (LVLMs) offer a promising alternative by integrating visual perception with natural-language reasoning. Recent models such as LLaVA-1.6 [9], Gemma-3 [10], and Qwen2.5-VL [11] enable context-aware inspection through multimodal reasoning, semantic interpretation, and natural-language interaction [12]. Their strong zero-shot and few-shot generalization capabilities allow adaptation to unseen defect types without extensive retraining, which is highly attractive for dynamic industrial environments. Despite these advantages, LVLM applications in industrial defect inspection remain underexplored, with most prior studies focusing on natural-image understanding rather than manufacturing-specific workflows [13], [14].

In this work, we investigate LVLM-based multimodal defect inspection in Liquefied Natural Gas (LNG) tank manufacturing, a highly specialized process involving welding, surface treatment, cable routing, pipe installation, cutting, and foam spraying. These operations exhibit diverse and evolving defect categories that challenge conventional vision-based systems. By leveraging LVLMs, we aim to enable unified defect localization, semantic interpretation, and automatic generation of structured inspection reports to support human operators.

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) - Innovative Human Resource Development for Local Intellectualization Program grant funded by the Korea government (MSIT) (IITP-2025-RS-2024-00436773), and by the "Development and Demonstration of AI Services for Manufacturing Industry Specialization" grant funded by the Korea government (Ministry of Trade, Industry and Energy) (Project no. SG20240201).

Specifically, this study makes the following contributions:

- We propose an LVLM-based multimodal inspection framework capable of defect localization, semantic reasoning, and automated report generation with interpretable outputs.
- We optimize fine-tuning and inference using Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO), enabling efficient deployment on a single 24GB GPU for on-premises industrial use while preserving data confidentiality.

Experimental results demonstrate that properly optimized LVLMs provide a scalable and interpretable alternative to conventional CNN-based inspection systems, supporting more adaptive and autonomous quality assurance pipelines for next-generation manufacturing.

II. Related Works

A. Classical Defect Detection Approaches

Traditional automated defect detection methods can be broadly categorized into three groups. Embedding-based methods [15], [16], [17] extract representations of defect-free samples using pretrained encoders and detect anomalies via distance-based similarity measures. Reconstruction-based methods [18], [19], [20] learn generative models on normal data and identify defects through reconstruction errors. CLIP-based methods [21], [22] exploit multimodal alignment between visual features and textual prompts to enable zero-shot or weakly supervised anomaly detection.

While these methods achieve strong pixel-level or image-level anomaly prediction, they lack the capability to produce semantic explanations, structured metadata, or comprehensive inspection reports—abilities increasingly required in smart manufacturing environments.

B. LVLMs for Industrial Defect Detection

Motivated by the strong perceptual and reasoning abilities of LVLMs, recent works have begun exploring their applicability to quality inspection. Several studies apply LVLMs directly, without fine-tuning, to anomaly detection or visual question answering [23], [24], [25]. For example, Chen et al. [23] introduces specialized input modules tailored to question types. However, LVLMs trained primarily for general-purpose tasks often struggle with industrial defects unless properly adapted [26].

To improve domain specificity, a number of methods perform SFT on industrial anomaly datasets [13], [26], [27]. AnomalyGPT [13] uses dual branches to generate anomaly masks and textual descriptions from synthetic data. Anomaly-OV [26] introduces a Look Twice Feature Matching (LTFM) mechanism to emphasize abnormal visual tokens. Although SFT significantly improves detection quality, SFT-based models still depend heavily on annotated data, struggle to generalize to real-world defect diversity, and optimize fixed training objectives that may not fully align with downstream reasoning tasks.

C. Reinforcement Learning and GRPO-Based Alignment

To address the limitations of SFT, recent works incorporate GRPO to align LVLMs using task-specific rewards and preference-based feedback [28], [29], [30](Li et al. 2025; Zhao et al. 2025; Zeng et al. 2025; Chao et al. 2025). LR-IAD [28] introduces focal rewards to mitigate class imbalance, while AnomalyR1 [29] proposes the Reasoned Outcome Alignment Metric (ROAM) to jointly optimize reasoning consistency and prediction accuracy.

However, standard GRPO suffers from degraded reward signals on hard samples where all candidate responses are incorrect, leading to unstable optimization and limited convergence. This motivates the development of more robust GRPO strategies capable of handling difficult cases and strengthening reasoning–detection alignment.

D. Research Gap and Contributions

Most existing defect detection studies focus on homogeneous materials or single-process settings and do not address the complexity of LNG tank manufacturing, which involves heterogeneous fabrication processes and highly diverse defect modes. Furthermore, prior LVLM-based approaches have not jointly addressed defect localization, structured metadata extraction, automated inspection-report generation, nor efficient SFT–GRPO training on a single 24GB GPU.

These gaps motivate our proposed hybrid SFT–GRPO LVLM framework, which unifies multimodal defect detection and semantic reporting while enabling practical, on-premises industrial deployment.

III. Materials And Methods

This study proposes a unified SFT–GRPO fine-tuning framework for adapting open-source LVLMs to multimodal defect inspection in LNG tank manufacturing. We evaluate four representative LVLMs—LLaVA-1.6-Mistral-7B, Gemma-3-4B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B—selected to span multiple architectural families and parameter scales under realistic on-premises constraints (single 24GB GPU). The fine-tuned models perform defect localization and structured inspection-report generation, producing JSON outputs that include defect labels, bounding boxes, and contextual metadata (e.g., tank type, location, part, and quality).

As illustrated in Fig. 1, the proposed pipeline consists of four stages: data preprocessing, prompt design, SFT–GRPO fine-tuning, and evaluation.

A. Dataset and Data Preprocessing

We use the open-access AIHub LNG Tank Quality Inspection Dataset [31], which contains 188,631 high-resolution images covering major fabrication processes such as welding, coating, insulation installation, cable routing, pipe installation, and cutting. Each image includes defect labels, bounding boxes, and contextual metadata.

To enable memory-efficient fine-tuning on 24GB GPUs, we construct a balanced subset of 22,500 images across 15 defect classes (1500 samples per class). For each class, 500 images are used for SFT, 500 for GRPO reinforcement learning, and 500 for testing, ensuring equal category representation.

All original images (size 1920×1080) are resized to 512×512 pixels, and COCO-format bounding boxes [*x, y, width, height*] are converted to corner coordinates [x_{min}, y_{min}, x_{max}, y_{max}] using standard normalization, as in (1) to (4):

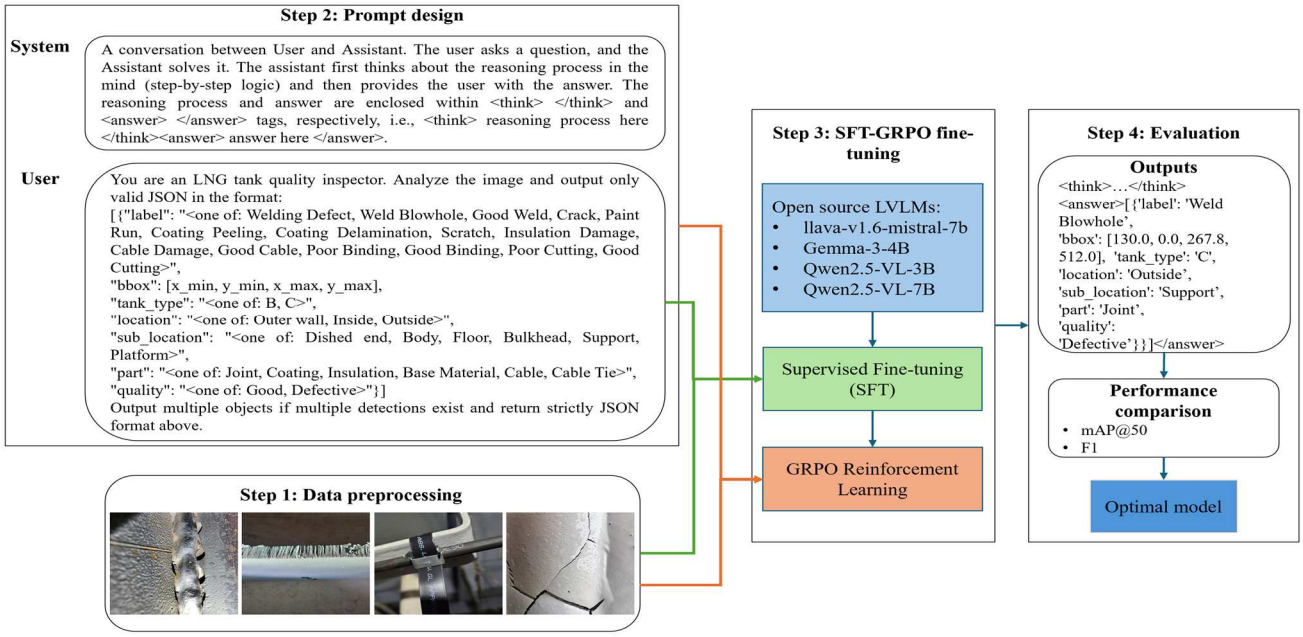

Fig. 1. The proposed unified SFT–GRPO fine-tuning framework for adapting open-source LVLMs to multimodal defect inspection in LNG tank manufacturing.

$$x_{min} = x \div 1920 \times 512 \quad (1)$$

$$y_{min} = y \div 1080 \times 512 \quad (2)$$

$$x_{max} = (x + width) \div 1920 \times 512 \quad (3)$$

$$y_{max} = (y + height) \div 1080 \times 512 \quad (4)$$

Alongside defect labels and bounding boxes, six categorical metadata fields are retained to support structured inspection-report generation: tank_type ∈ {B, C}, location ∈ {Outer wall, Inside, Outside}, sub_location ∈ {Dished end, Body, Floor, Bulkhead, Support, Platform}, part ∈ {Joint, Coating, Insulation, Base Material, Cable, Cable Tie}, and quality ∈ {Good, Defective}. These structured annotations provide the multimodal supervision required for unified defect localization, attribute prediction, and JSON-format output generation in subsequent SFT–GRPO fine-tuning.

B. *Prompt Design*

Prompt design plays a central role in enabling LVLMs to perform unified defect localization, attribute prediction, and structured inspection reporting. As shown in Fig. 1, two prompt formats are used for the SFT and GRPO stages.

For SFT, we construct the user's prompt for 7500 training samples (500 per class). Each sample pairs an image with a user instruction specifying the required JSON schema and listing all valid categories for each metadata field (label, tank_type, location, sub_location, part, quality). These schema-constrained prompts guide the LVLM to produce deterministic, machine-readable outputs.

For GRPO reinforcement learning, we utilize both the system prompt and the user prompt applied to 7500-image set for GRPO. The system prompt enforces a two-stage response: hidden reasoning enclosed in <think>...</think> followed by the JSON answer in <answer>...</answer>. The user prompt mirrors the SFT schema to maintain consistent output structure.

This design ensures consistent JSON formatting, strict schema adherence, and separation of reasoning from evaluation, enabling reliable supervision across both training stages.

C. *SFT–GRPO fine-tuning*

Fine-tuning proceeds in two stages.

1) Supervised Fine-tuning (SFT): During the first stage, each LVLM is adapted to the LNG-inspection domain using 7500 annotated training samples (500 per defect class). SFT trains the model to follow schema-constrained prompts, recognize defect patterns, and output valid JSON responses. To support training on a single 24-GB GPU, we adopt parameter-efficient fine-tuning (PEFT) using lightweight adapter modules injected into attention layers. This approach reduces trainable parameters while preserving the expressive capacity of the pretrained LVLM. All models are initialized from their Instruct-tuned 16-bit checkpoints to ensure stability and consistent optimization behavior across architectures. Training follows a unified configuration (learning rate, batch size, warm-up, scheduler), summarized in Table 1. SFT establishes the model's foundational abilities—prompt compliance, structured output generation, defect recognition, and attribute prediction—providing the base policy for reinforcement learning.

TABLE I. PEFT CONFIGURATION PARAMETERS

Parameter	Value
Optimizer	adamw_8bit
Learning rate	2e-5
Learning rate strategy	cosine

Parameter	Value
bf16	True
Warm up ratio	0.03
Epochs	3
LoRA Rank	16
Batch size	2
Gradient accumulation step	4

2) Group Relative Policy Optimization (GRPO) Reinforcement Learning: The second stage further aligns the LVLMs with downstream objectives through Group Relative Policy Optimization (GRPO). GRPO is well-suited for defect inspection because the task produces verifiable outputs—JSON format, bounding box geometry, label correctness, and metadata accuracy—allowing reward functions to provide precise optimization signals. We use another 7500-sample set (500 per class) for GRPO training. Each sample contains an image, a system prompt enforcing the <think>…</think> and <answer>…</answer> structure, and a user prompt specifying the required JSON schema. GRPO samples multiple candidate outputs per prompt and updates the model based on the advantage of each candidate relative to the group.

To improve unified detection and report generation, we design a reward suite consistent with Figure 1 and tailored to the LNG inspection schema. Rewards operate only on the <answer> slice, ensuring that chain-of-thought remains hidden while still guiding structured reasoning. The main reward components are:

- Format-Tag Reward: Ensures the output respects the strict <think>…</think> and <answer>…</answer> envelope. Reward = 1 if strictly matched; else 0.
- JSON Validity Reward: Encourages syntactically correct JSON arrays. Reward = 1 for valid JSON; else 0.
- Schema Compliance Reward: Measures whether each predicted object satisfies the required fields (label, bbox, tank_type, location, sub_location, part, material, quality) and allowed category sets.
- Count Consistency Reward: Rewards matching the number of predictions to the ground truth, normalized by count difference.
- Detection IoU-F1 Reward: Uses greedy IoU-based matching (threshold 0.5) to compute label-aware F1, rewarding correct localization while penalizing false positives and false negatives.
- Attribute Accuracy Reward: Evaluates correctness of all six metadata fields for IoU-matched object pairs.
- Anti-Gibberish Reward: Light hygiene constraint penalizing extraneous or noisy text beyond the JSON output.

These rewards collectively shape both format correctness (JSON validity, tag structure, schema compliance) and semantic accuracy (bbox localization, label prediction, metadata attributes). GRPO fine-tuning substantially improves JSON correctness, IoU-F1 detection performance, and metadata prediction accuracy over the SFT-only baseline.

D. Performance Evaluation

Model performance is assessed using the test split of 7500 images (500 samples per defect class). Following the unified output structure shown in Figure 1, evaluation focuses on two complementary aspects: defect localization and structured attribute prediction.

For localization, we compute mean Average Precision at IoU 0.5 (mAP@50), where a prediction is considered correct if the Intersection-over-Union between the predicted and ground-truth bounding box is $\geqslant 0.5$. mAP@50 summarizes per-class Average Precision (AP) and provides a reliable measure of how well the LVLM identifies and localizes defect regions.

For structured report generation, we measure F1 score, the harmonic mean of precision and recall, to evaluate classification accuracy for all categorical fields within the JSON output. Specifically, F1 is computed for the predicted label, tank_type, location, sub_location, part, and quality attributes associated with each localized defect. These metrics jointly capture the core requirements of industrial multimodal inspection: accurate spatial localization and reliable semantic interpretation.

IV. Experimental Results And Discussion

A. Experimental Setup

All experiments were conducted on a workstation with an AMD Ryzen 9 7950X CPU, NVIDIA RTX 4090 GPU (24GB VRAM), and 96GB DDR5 RAM. All LVLMs were fine-tuned using the Unsloth framework [32].

B. SFT–GRPO Fine-tuning Process Evaluation

Table II reports the performance of four LVLMs before and after applying the proposed SFT–GRPO fine-tuning framework. Prior to fine-tuning, all models exhibit limited localization capability on LNG-specific defects, with mAP@50 values of 35–39%. After SFT–GRPO, localization accuracy improves dramatically by 47–51 percentage points, yielding final mAP@50 scores between 84% and 88%. Among the evaluated models, Qwen2.5-VL-7B achieves the highest localization accuracy (88.77% mAP@50), indicating stronger spatial reasoning and more reliable bounding-box prediction.

Fine-tuning also substantially improves structured attribute prediction. Defect label F1 increases from approximately 14% to 84–90%, reflecting effective schema learning and reward-based alignment. High-level categorical fields such as tank_type, location, and quality consistently achieve over 90% F1, demonstrating robust semantic understanding of LNG manufacturing contexts. More fine-grained attributes, including sub_location and part, show slightly lower absolute performance but still benefit from 30–40 percentage point gains, confirming the effectiveness of GRPO-based reward shaping for complex multimodal reasoning. Overall, Qwen2.5-VL-7B delivers the strongest performance across both localization and attribute prediction. Notably, smaller models (Gemma-3-4B and Qwen2.5-VL-3B) also achieve large performance gains, indicating that the proposed pipeline remains effective under compact model and 24GB GPU constraints.

TABLE II. PERFORMANCE EVALUATION ON THE TEST SET BEFORE AND AFTER SFT–GRPO FINE-TUNING

Model	**mAP@50 (%) (Before → After)**	**F1 (%) (Before → After)**					
		label	tank_type	location	sub_location	part	quality
LLaVA-1.6-Mistral-7B	36.52 → 86.75	14.23 → 87.10	3.59 → 97.27	15.56 → **72.08**	10.71 → 45.87	7.27 → 62.11	29.62 → **91.24**
Gemma-3-4B	39.32 → 85.61	13.88 → 84.91	4.13 → 97.20	12.20 → 70.10	9.77 → 44.73	8.23 → 61.02	33.23 → 89.62
Qwen2.5-VL-3B	35.07 → 84.68	13.11 → 84.87	4.54 → 97.21	18.79 → 73.57	8.04 → 45.21	9.02 → 65.15	30.11 → 89.93
Qwen2.5-VL-7B	39.13 → **88.77**	14.56 → **90.19**	2.47 → **97.82**	13.92 → 73.19	8.18 → **48.33**	12.9 → **65.84**	35.67 → 91.15

These results highlight the complementary roles of the two training stages: SFT establishes prompt compliance and structured output generation, while GRPO refines localization accuracy and semantic consistency through verifiable, task-aligned rewards.

C. Ablation study

To analyze the contribution of each training strategy and contextualize performance against conventional detectors, we conduct an ablation study on Qwen2.5-VL-7B and compare it with representative YOLO-based and transformer-based detectors, as summarized in Table III.

TABLE III. ABLATIONS ON DIFFERENT TRAINING STRATEGIES OF THE OPTIMIZED QWEN2.5-VL-7B AND TRADITIONAL DETECTORS .

Model	**mAP@50 (%)**	**F1 (%)**	
		label	quality
Qwen2.5-VL-7B (SFT)	85.26	84.16	89.92
Qwen2.5-VL-7B (GRPO)	85.23	84.33	90.01
Qwen2.5-VL-7B (SFT+GRPO)	88.77	90.19	91.15
YOLOv11m	86.53	85.01	87.92
YOLOv12m	86.56	85.89	88.89
RT-DETR	86.51	84.45	86.69

For evaluation, we report mAP@50 for defect localization and F1 scores for the label and quality attributes, which represent the two most critical elements of industrial inspection reports: defect category identification and quality assessment.

When trained with SFT only, Qwen2.5-VL-7B achieves 85.26% mAP@50 and 84.16% F1 (label), indicating that supervised schema-constrained fine-tuning effectively establishes prompt compliance, structured output generation, and baseline defect-recognition capability. Training with GRPO only yields comparable localization and classification performance (85.23% mAP@50, 84.33% F1), demonstrating that verifiable reward signals can guide multimodal reasoning even without explicit paired supervision. However, both single-stage strategies underperform in structured attribute accuracy and overall consistency.

The combined SFT+GRPO strategy consistently outperforms both individual stages, achieving 88.77% mAP@50, 90.19% F1 (label), and 91.15% F1 (quality). This confirms that SFT provides stable grounding and schema adherence, while GRPO further refines spatial localization and semantic alignment, making their integration essential for unified defect detection and structured inspection reporting.

For a fair comparison with traditional detectors, YOLOv11m, YOLOv12m, and RT-DETR are each trained in two separate configurations. In the first setting, models are trained for object detection, and we report mAP@50 and F1 score for defect labels. In the second setting, the same architectures are retrained as quality classifiers, and performance is reported using F1 score for the quality attribute. This two-stage protocol reflects the fact that conventional detectors cannot jointly predict detection and semantic attributes within a single unified model.

Although YOLO-based and RT-DETR models achieve competitive localization performance (86.51–86.56% mAP@50), they remain inferior to the SFT+GRPO–optimized LVLM in both detection accuracy and semantic attribute prediction. More importantly, traditional detectors are limited to task-specific outputs and lack the ability to generate structured, multimodal inspection reports, which are critical in LNG manufacturing workflows.

V. CONCLUSION

This paper presents a unified LVLM-based multimodal inspection framework for LNG tank manufacturing, enabling joint defect localization and structured inspection-report generation within a single model. By combining parameter-efficient supervised fine-tuning (SFT) with Group Relative Policy Optimization (GRPO), the proposed approach effectively aligns LVLMs with industrial inspection requirements under practical on-premises constraints.

Experiments on a open-source LNG inspection dataset from AIHub show that the optimized Qwen2.5-VL-7B (SFT+GRPO) achieves strong localization and semantic accuracy, outperforming single-stage training strategies and traditional detectors such as YOLO and RT-DETR on key inspection attributes. Beyond detection accuracy, the proposed framework provides a distinct advantage in semantic interpretability and structured reporting, which are critical for complex industrial workflows.

Overall, these results indicate that properly aligned LVLMs can serve as a practical and scalable foundation for next-generation industrial inspection, unifying detection and semantic reporting in complex production environments. Future work will extend the framework to richer attribute sets, improve robustness under distribution shifts (e.g., lighting and

process drift), and investigate real-time deployment and integration with factory inspection workflows.

REFERENCES

[1] E. Gonzalez Santacruz, D. Romero, J. Noguez, and T. Wuest, "Integrated quality 4.0 framework for quality improvement based on Six Sigma and machine learning techniques towards zero-defect manufacturing," *The TQM Journal*, vol. 37, no. 4, pp. 1115–1155, Mar. 2024, doi: 10.1108/TQM-11-2023-0361.

[2] S. J. Oks *et al.*, "Cyber-Physical Systems in the Context of Industry 4.0: A Review, Categorization and Outlook," *Inf Syst Front*, vol. 26, no. 5, pp. 1731–1772, Oct. 2024, doi: 10.1007/s10796-022-10252-x.

[3] N. Yadav, V. Gupta, and A. Garg, "Industrial Automation Through AI-Powered Intelligent Machines—Enabling Real-Time Decision-Making," in *Recent Trends in Artificial Intelligence Towards a Smart World: Applications in Industries and Sectors*, R. Arya, S. C. Sharma, A. K. Verma, and B. Iyer, Eds., Singapore: Springer Nature, 2024, pp. 145–178. doi: 10.1007/978-981-97-6790-8_5.

[4] R. Khanam, M. Hussain, R. Hill, and P. Allen, "A Comprehensive Review of Convolutional Neural Networks for Defect Detection in Industrial Applications," *IEEE Access*, vol. 12, pp. 94250–94295, 2024, doi: 10.1109/ACCESS.2024.3425166.

[5] S. Sundaram, A. Zeid, S. Sundaram, and A. Zeid, "Artificial Intelligence-Based Smart Quality Inspection for Manufacturing," *Micromachines*, vol. 14, no. 3, Feb. 2023, doi: 10.3390/mi14030570.

[6] M. Alam, "Predictive Maintenance Strategies for Reducing Downtime in Manufacturing," *American Journal Of Industrial And Production Engineering*, vol. 6, no. 4, pp. 14–25, Aug. 2025, doi: 10.71465/ajipe.3405.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[8] S. A. Singh, A. S. Kumar, and K. A. Desai, "Comparative assessment of common pre-trained CNNs for vision-based surface defect detection of machined components," *Expert Systems with Applications*, vol. 218, p. 119623, May 2023, doi: 10.1016/j.eswa.2023.119623.

[9] H. L. Lee Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, Yong Jae, "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge," LLaVA. Accessed: Nov. 15, 2025. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[10] A. Kamath *et al.*, "Gemma 3 Technical Report," *CoRR*, vol. abs/2503.19786, Mar. 2025, [Online]. Available: https://doi.org/10.48550/arXiv.2503.19786

[11] S. Bai *et al.*, "Qwen2.5-VL Technical Report," Feb. 19, 2025, *arXiv*: arXiv:2502.13923. doi: 10.48550/arXiv.2502.13923.

[12] P. Zhou *et al.*, "When Large Vision Language Models Meet Multimodal Sequential Recommendation: An Empirical Study," in *Proceedings of the ACM on Web Conference 2025*, in WWW '25. New York, NY, USA: Association for Computing Machinery, Tháng T 2025, pp. 275–292. doi: 10.1145/3696410.3714764.

[13] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, pp. 1932–1940, Mar. 2024, doi: 10.1609/aaai.v38i3.27963.

[14] Y. Chen, Z. Zhang, L. Yu, G. Huang, and J. Huang, "Defect detection of engine cylinder liner inner wall in multi-interference environments based on Large Vision-Language Models," *J. Phys.: Conf. Ser.*, vol. 2897, no. 1, p. 012050, Oct. 2024, doi: 10.1088/1742-6596/2897/1/012050.

[15] H. Hu *et al.*, "DSMBAD: Dual-Stream Memory Bank Framework for Unified Industrial Anomaly Detection," *Electronics*, vol. 14, no. 14, July 2025, doi: 10.3390/electronics14142748.

[16] J. Hyun, S. Kim, G. Jeon, S. H. Kim, K. Bae, and B. J. Kang, "ReConPatch: Contrastive Patch Representation Learning for Industrial Anomaly Detection," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2052–2061. Accessed: Dec. 01, 2025. [Online]. Available: https://openaccess.thecvf.com/content/WACV2024/html/Hyun_ReConPatch_Contrastive_Patch_Representation_Learning_for_Industrial_Anomaly_Detection_WACV_2024_paper.html

[17] M. Li, J. He, Z. Ying, G. Li, and M. Zhou, "MemADet: A Representative Memory Bank Approach for Industrial Image Anomaly Detection," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2024, pp. 2261–2266. doi: 10.1109/SMC54092.2024.10831766.

[18] J. Guo, S. Lu, W. Zhang, F. Chen, H. Li, and H. Liao, "Dinomaly: The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 20405–20415. Accessed: Dec. 01, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Guo_Dinomaly_The_Less_Is_More_Philosophy_in_Multi-Class_Unsupervised_Anomaly_CVPR_2025_paper.html

[19] D.-C. Hoang *et al.*, "Unsupervised Visual-to-Geometric Feature Reconstruction for Vision-Based Industrial Anomaly Detection," *IEEE Access*, vol. 13, pp. 3667–3682, 2025, doi: 10.1109/ACCESS.2025.3525567.

[20] X. Zhang, M. Xu, and X. Zhou, "RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16699–16708.

[21] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, "AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 55–72. doi: 10.1007/978-3-031-72761-0_4.

[22] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19606–19616.

[23] Z. Chen, H. Chen, M. Imani, and F. Imani, "Can Multimodal Large Language Models Be Guided to Improve Industrial Anomaly Detection?," in *Volume 2B: 45th Computers and Information in Engineering Conference (CIE)*, Anaheim, California, USA: American Society of Mechanical Engineers, Aug. 2025, p. V02BT02A051. doi: 10.1115/DETC2025-168875.

[24] S. Mokhtar, A. Mousakhan, S. Galesso, J. Tayyub, and T. Brox, "Detect, Classify, Act: Categorizing Industrial Anomalies with Multi-Modal Large Language Models," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 4097–4106. Accessed: Dec. 01, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025W/VAND/html/Mokhtar_Detect_Classify_Act_Categorizing_Industrial_Anomalies_with_Multi-Modal_Large_Language_CVPRW_2025_paper.html

[25] E. Jin *et al.*, "LogicAD: Explainable Anomaly Detection via VLM-based Text Feature Extraction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, pp. 4129–4137, Apr. 2025, doi: 10.1609/aaai.v39i4.32433.

[26] J. Xu, S.-Y. Lo, B. Safaei, V. M. Patel, and I. Dwivedi, "Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 20370–20382.

[27] H. Deng, H. Luo, W. Zhai, Y. Guo, Y. Cao, and Y. Kang, "VMAD: Visual-enhanced Multimodal Large Language Model for Zero-Shot Anomaly Detection," *IEEE Transactions on Automation Science and Engineering*, pp. 1–1, 2025, doi: 10.1109/TASE.2025.3591656.

[28] P. Zeng, F. Pang, Z. Wang, and A. Yang, "LR-IAD:Mask-Free Industrial Anomaly Detection with Logical Reasoning," Apr. 28, 2025, *arXiv*: arXiv:2504.19524. doi: 10.48550/arXiv.2504.19524.

[29] Y. Chao, J. Liu, J. Tang, and G. Wu, "AnomalyR1: A GRPO-based End-to-end MLLM for Industrial Anomaly Detection," Apr. 16, 2025, *arXiv*: arXiv:2504.11914. doi: 10.48550/arXiv.2504.11914.

[30] W. Li, G. Chu, J. Chen, G.-S. Xie, C. Shan, and F. Zhao, "LAD-Reasoner: Tiny Multimodal Models are Good Reasoners for Logical Anomaly Detection," Apr. 17, 2025, *arXiv*: arXiv:2504.12749. doi: 10.48550/arXiv.2504.12749.

[31] AIHub., "LNG Tank Quality Inspection Image Data." https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71690, Oct. 2024. Accessed: Sept. 18, 2025. [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71690

[32] Daniel Han, Michael Han, and Unsloth Team, *Unsloth*. (2023). [Online]. Available: http://github.com/unslothai/unsloth