Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085203

ARTICLE

A Three-Layer Multi-Agent Framework for PHM-Enabling Autonomous Condition Monitoring of Power ICT Infrastructure in Underground Facilities

Jaekyung Lee[1,2], Byungsung Ko[2], Jiwon Lee[2], Jaeheon Park[2], Taewon Kim[2], Seoktae Kim[2] and Wonhee Kim[3,*]

[1]Department of Energy Systems Engineering, Chung-Ang University, Seoul, Republic of Korea
[2]KEPCO Research Institute, Daejeon, Republic of Korea
[3]School of Energy Systems Engineering, Chung-Ang University, Seoul, Republic of Korea

*Corresponding Author: Wonhee Kim. Email: whkim79@cau.ac.kr

Received: 07 May 2026; Accepted: 22 July 2026; Published: 15 September 2026

ABSTRACT: This study proposes the Artificial Intelligence-integrated Inspection Ecosystem (AIIE) as an autonomous condition monitoring platform to enable Prognostics and Health Management (PHM) for underground infrastructure facilities at the Korea Electric Power Corporation (KEPCO) power Information and Communication Technology (ICT) center. To address the environmental dependency of conventional systems, which necessitate extensive control logic redesigns upon changes in target facilities or environments, a three-layer abstraction architecture separating directive, orchestration, and execution roles is established, integrating a quadrupedal robot with heterogeneous sensors into a unified control structure. To overcome the limitation of relying on one general-purpose model for all inspection tasks, this study introduces a collaborative multi-agent system as its core component. This system is individually optimized for analog gauges, digital gauges, switch/LED states, and qualitative appearance defects. The proposed system achieves an Analog Gauge (AG) inspection error of 1.66% of full scale (%FS) (95% confidence interval (CI): [1.31, 2.11]) and a Digital Gauge (DG) recognition rate of 83.8% (n = 130) using a small vision-language model (sVLM). A three-way validation score of κ_w = 0.724 was obtained in identifying qualitative appearance defects, such as oil/water leaks, cleanliness, and insulation degradation. Field evaluations using 11,692 cases of real operational data confirm the system's effectiveness, achieving an automatic inspection success rate of 87.6%. Downstream PHM prognostic functions, such as anomaly detection, degradation modeling, and remaining useful life (RUL) estimation, remain as future research directions.

KEYWORDS: Prognostics and health management (PHM); condition monitoring; multi-agent artificial intelligence (AI) system; edge computing; vision-language model (VLM); power infrastructure

1 Introduction

Infrastructure facilities operating within power Information and Communication Technology (ICT) centers are crucial to national digital services. Within these centers, thousands of heterogeneous devices—including temperature-humidity sensors, uninterruptible power supply (UPS) systems, generators, power distribution facilities, and fire extinguishing systems—operate in a highly integrated configuration. In large-scale data centers where these facilities are intensively deployed in confined underground environments, facility anomalies directly lead to service interruptions. Consequently, continuous condition monitoring and early health assessment are indispensable prerequisites for Prognostics and Health Management (PHM). Conventional inspection systems rely on manual patrols, where dedicated personnel repeatedly traverse

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

designated paths to visually inspect a vast number of points. This process generates inconsistent records that are inadequate to serve as a robust foundation for data-driven PHM.

This approach entails several structural limitations. First, a patrol routine that repeatedly performs the same tasks daily diverts the capabilities of skilled personnel toward simple, repetitive labor. This degrades operational efficiency. Second, personnel shortages during night shifts create inspection gaps. The resulting longer intervals increase the risk of missing early warning signs. Third, manual recording of inspection results using spreadsheets hampers historical tracking, anomaly pattern analysis, and preventive maintenance decision-making. Fourth, variability in gauge readings and qualitative assessments, which rely on operator experience, compromises the consistency of inspection results, even for identical facilities.

Robotic automation is a promising technical alternative to address these limitations. However, underground infrastructure environments differ significantly from the operating conditions assumed by conventional industrial robotic systems. These environments are characterized by narrow passages, irregular steps, high-density equipment layouts, and poor lighting. Heterogeneous tasks—such as reading analog gauges, recognizing digital gauges, checking switch/LED states, and conducting qualitative appearance diagnostics (e.g., cleanliness and corrosion)—must be processed on a single robotic platform. In addition, the spatial registration problem of matching various detected objects with checklists in the inspection database must be accurately resolved. Moreover, conventional robotic inspection systems are typically engineered for a specific facility layout, requiring extensive redesign of control logic whenever the target site or operating environment changes.

To address these complex challenges, this study proposes the Artificial Intelligence-integrated Inspection Ecosystem (AIIE) as a Physical-AI-based, PHM-enabling autonomous condition monitoring platform. Here, Physical-AI refers to an edge-based intelligent-system paradigm in which AI inference is embedded directly into physical devices, enabling autonomous on-site decision-making and execution without cloud dependency. The AIIE targets the data acquisition and condition monitoring layers of the PHM pipeline, and we validated its performance at the underground power ICT infrastructure facilities of the Korea Electric Power Corporation (KEPCO) Power ICT Daejeon Center. By mitigating the subjectivity and repetitive burden of manual inspections, the AIIE provides a continuous, high-quality stream of operational data, establishing a solid data foundation for downstream PHM tasks, such as anomaly detection and degradation trend analysis. These downstream prognostic functions remain beyond the present scope and are addressed as future work.

The novelty and major contributions of this study are as follows:

1. Empirical validation of a PHM-oriented autonomous condition monitoring system in an operational environment: While many conventional studies remain limited to proof-of-concept validation in controlled laboratory settings, this study conducts an empirical investigation in an operational underground facility of power ICT infrastructure. By integrating the autonomous navigation of a quadrupedal robot, deep-learning-based object detection, and qualitative inference from recent vision-language models (VLMs), we demonstrate the continuous operational reliability of the AI-based condition monitoring system. This system generates structured time-series data necessary for downstream prognostics and preventive maintenance.
2. Establishment of a three-layer abstraction architecture designed for site-independent deployment: To overcome the limitations of conventional systems that require a complete redesign of control logic whenever an inspection target facility or operating environment changes, we establish a three-layer abstraction architecture that isolates roles into directive (defining health monitoring policies), orchestration (intelligent resource scheduling and control), and execution (performing unit sensing and diagnostic tasks). This architecture separates PHM policies from hardware details to facilitate

site-independent deployment. The validation of portability in heterogeneous environments remains a future research topic.

3. Multi-agent collaboration for integrated condition assessment across heterogeneous facility types: To address the complexity of industrial PHM sites, we propose a multi-agent collaborative architecture in which each agent specializes in a distinct condition monitoring modality to overcome the limitations of a single general-purpose model. Heterogeneous facilities with markedly different health indicators and reading mechanisms—including analog gauges (computer vision and geometric calibration), digital gauges (Optical Character Recognition (OCR) optimization), switch/LED states (discrete classification), and qualitative appearance defects such as cleanliness and corrosion (VLM-based inference)—are evaluated independently within a unified framework. This design improves computational efficiency and yields a health assessment engine that allows the addition or modification of sensing logic for any facility type without disrupting the overall PHM pipeline.

2 Related Work

2.1 Robot-Based Automated Inspection of Industrial Facilities

As the demand for automated inspection of social infrastructure continues to increase, studies integrating various robotic platforms and computer vision technologies have attracted growing interest. A systematic literature review by Halder and Afsari [1] analyzed 269 papers to provide a comprehensive overview of research in building and infrastructure inspection robotics. This review demonstrated that Unmanned Aerial Vehicles and Unmanned Ground Vehicles lead the inspection field, and categorized the applications of automated inspection into five major domains: maintenance inspection, construction quality inspection, construction progress monitoring, as-built modeling, and safety inspection. The review also highlighted the spatial constraints of fixed single-sensor systems and the need for mobile robotic platforms that adapt to different infrastructure environments. Lee et al. [2] surveyed robotic technologies for civil infrastructure inspection, and further introduced a hierarchical graph-based SLAM (HG-SLAM) technique that fuses IMU, camera, and 3D LiDAR data to enable UAV-based bridge inspection under GNSS denial, together with a multi-layer coverage path planner (ML-CPP) that improves flight efficiency and inspection completeness for UAV-based structural coverage.

Focusing specifically on power facilities, Wei and Rey [3] reviewed the technical status of substation inspection robots. They emphasized that multi-sensor fusion and three-dimensional (3D) Light Detection and Ranging (LiDAR)-based SLAM technologies are essential for avoiding complex obstacles in substations and planning inspection paths, identifying the development of real-time, high-precision navigation and simplified environmental perception algorithms as key challenges for future work. Mineo et al. [4] proposed a geometric and volumetric autonomous inspection framework for free-form components using a six-degree-of-freedom robotic arm and an ultrasonic testing sensor. By departing from conventional offline programming methods and solving the shortest path problem online with the A* algorithm, they demonstrated that fully autonomous single-pass inspection combining real-time robot control and sensor feedback is feasible. Konstantinidis et al. [5] proposed AROWA, an autonomous robot framework for Warehouse 4.0 health and safety inspection operations. This system uses deep learning models such as You Only Look Once (YOLO) and Mask Region-based Convolutional Neural Network (Mask R-CNN) to detect risk factors including fire, flooding, blocked passages, and non-use of personal protective equipment (PPE) without human intervention, and is integrated with a warehouse management system. However, it does not address integrated AI reading algorithms or precision measurement functions for heterogeneous equipment.

As illustrated in Fig. 1, the practical deployment of industrial quadruped robots has attracted significant attention. Boston Dynamics' Spot has been deployed for unmanned patrol inspections combining thermal

imaging cameras, gas sensors, and Red-Green-Blue (RGB) cameras in various hazardous environments such as refineries, power plants, and construction sites, with real-world use cases reported by global energy corporations. ANYmal by ANYbotics [6,7] is an explosion-proof certified, IP67-rated industrial quadruped robot specialized for offshore High-Voltage Direct Current platforms in the North Sea and European energy/chemical plants. This system is equipped with optical/thermal cameras and LiDAR on a pan-tilt head, and performs acoustic diagnostics of machinery such as pumps using a built-in microphone, and leverages autonomous patrol data for preventive maintenance. Deep Robotics in China [8] introduced quadruped robots to inspect substations and power facilities, establishing facility condition monitoring systems through autonomous navigation-based patrols and image data collection, and Energy Robotics [9] provides a software ecosystem that manages these robots through an integrated platform and links AI analysis results with maintenance decision-making.

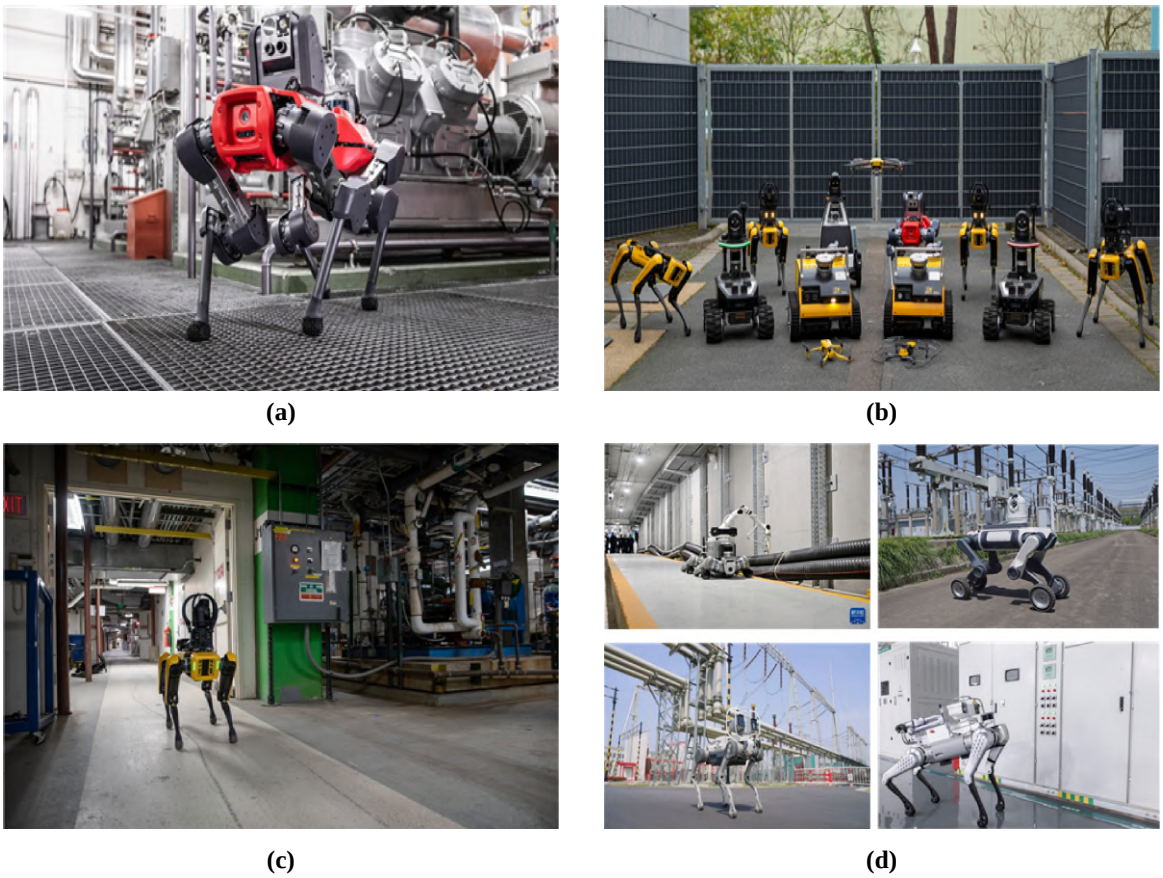


Figure 1: Major industrial quadruped robots. (**a**) ANYbotics ANYmal (**b**) Energy Robotics (**c**) Boston Dynamics Spot (**d**) Deep Robotics quadruped robot.

Underground infrastructure environments, such as the KEPCO Power ICT Center, possess unique characteristics differentiating them from conditions assumed in previous studies. Conventional wheeled mobile robots have practical limitations in path planning due to narrow passages, high-density equipment layouts, and coexisting irregular steps and obstacles. Conversely, quadruped robots not only ensure stable locomotion in unstructured terrains through their multi-jointed walking structure but also enhance visual recognition accuracy by dynamically securing an appropriate viewpoint of sensors through active posture control. Table 1 presents a comparative analysis of the characteristics of conventional wheeled mobile robots, general-purpose quadruped robots, and the approach proposed in this study.

Table 1: Comparative analysis of robotic systems.

Category	Wheeled Mobile Robot	Quadruped Robot	Proposed Approach
Mobility	Flat-terrain optimized	Complex-terrain capable	Environment-customized
Obstacle negotiation	Limited	Excellent	Structure-based negotiation
Inspection coverage	Path-limited	Expandable	Entire area coverage
Locomotion stability	Stable on flat ground	Stable on unstructured ground	Environment-tuned
System architecture	Commercial/Simple	Partially closed	Open (KEP-ROS)
AI application	Data acquisition-centric	Platform-dependent	In-house-developed AI
Cost	Low	High	Optimized
Maintenance	Easy	Complex	Managed
Scalability	Limited	Partially possible	Expandable
Target environment	Structured environments	Unstructured environments	Specific facilities

The studies above demonstrate mobile robotic automation for individual inspection domains, but most remain restricted to navigation or detection specialized for a single equipment type or facility, without addressing the integrated inspection of heterogeneous equipment modalities required in enclosed underground environments. As a result, direct application to the integrated inspection of heterogeneous equipment within enclosed and complex environments, such as underground power ICT infrastructure, remains challenging. This study addresses this research gap by integrating a three-layer (directive-orchestration-execution) multi-agent architecture with a quadrupedal robot to propose a system capable of processing analog gauges, digital gauges, switches, LEDs, and qualitative anomaly diagnosis within a single integrated framework.

2.2 Vision-Based Automatic Gauge Reading

In industrial robotic inspection, automatic gauge reading is a critical technology directly linked to system safety. Accordingly, both analog and digital gauges have been studied extensively.

2.2.1 Analog Gauge Reading

A primary challenge in analog gauge reading is the significant decline in pointer orientation estimation accuracy, which is caused by image quality degradation factors inherent to real-world environments, such as illumination variations, oblique viewpoints, reflections, and occlusions. To address this issue, Dong et al. [10] proposed a vector detection network (VDN) that models the pointer as a two-dimensional vector. The VDN simultaneously estimates a heatmap for the pointer tip and a scalar map representing the orientation using a ResNet-based backbone, thereby jointly detecting multiple needles. An end-to-end pipeline was constructed to first detect the gauge by deploying a region-based convolutional neural network (R-CNN) and subsequently derive the reading through template matching, demonstrating real-time reading capabilities on their proprietary Pointer-10K dataset.

Approaches integrating gauge detection, viewpoint rectification, and value conversion into a single pipeline have also been widely studied. GAUREAD, proposed by Milana et al. [11], detects circular analog gauges by deploying YOLOv4-tiny and applies a geometric perspective rectification algorithm that overcomes the limitations of the Hough transform to correct gauge faces distorted into ellipses due to oblique viewpoints. Consequently, a reading error within 3% for viewpoints below 20° and within 9% for viewpoints up to 50° was attained. Reitsma et al. [12] from the Autonomous Systems Lab at ETH Zurich proposed an interpretable, step-by-step segmentation pipeline designed for robotic deployment in real-world environments. They employed YOLOv8-based instance segmentation and orthogonal distance regression to fit the needle line. Considering that the scale range of the gauge is not known a priori, they dynamically

extracted numbers and units by deploying an OCR model combining DBNet and ABINet, reporting a relative error of less than 2%. To reduce the cost of label acquisition, Leon-Alcazar et al. [13] addressed the data scarcity issue by generating a synthetic dataset rendered using Blender. They constructed a pipeline based on MobileNetV2 and EfficientNet that regresses the angles of the needle, start point, and end point as a 360-degree classification problem, reporting a mean error of 4.55° and demonstrating a 52% error reduction rate compared to previous studies.

Automated industrial inspections suffer from a chronic data scarcity problem, where data are biased toward normal states, making it difficult to acquire images of critical anomalies or diverse lighting conditions. Zhang et al. [14] proposed ADD-GAN to generate synthetic defect images of substation equipment for unmanned patrol inspection, raising the mAP of a YOLOv7 detector from 71.9% to 81.5% under defect-image scarcity. In a similar manner, Xu et al. [15] addressed data scarcity by converting safety regulations into structured prompts and using text-to-image generative models, such as DALL · E 3, to generate synthetic datasets for high-altitude work sites.

However, these prior studies share a common limitation: most are specialized for a single type of analog gauge or assume controlled capturing conditions where preprocessing is omitted. To overcome these limitations, our analog gauge inspection agent adopts the P2-YOLO-Pose structure [16], which simultaneously detects the center of the rotation axis and the pointer tip through YOLO-Pose-based keypoint extraction. It is designed as a hybrid pipeline that preemptively resolves lens distortion and oblique viewpoint issues through homography-based orthographic projection using elliptical region of interest (ROI) optimization. This design is intended to achieve more robust pointer orientation inference under lens distortion and oblique viewpoints than prior approaches based on the OpenCV Hough transform [17] that utilize the geometric characteristics of circular analog gauges or ellipse fitting [12].

2.2.2 Digital Gauge Reading

For digital gauge reading, arbitrary-shape scene text detection technology is directly relevant. TextSnake, proposed by Long et al. [18], employs a geometric model that represents text as a sequence of overlapping disks along a symmetric axis. This method effectively detects scale variations and curved text, reaching state-of-the-art performance in curved and multi-oriented text detection at that time, and laying the foundation for subsequent scene text detection research. Conventional OCR-based approaches have advanced rapidly. Early methods relied on pattern-matching techniques, such as Tesseract [19]. Subsequent approaches adopted deep-learning-based models, including EasyOCR [20] and PP-OCR [21]. Most recently, Transformer-based technologies such as TrOCR [22] and GLM-OCR [23] have demonstrated high performance in scene text recognition. However, these OCR methods suffer from issues unique to seven-segment displays in power facility environments, such as font misrecognition, false detection of surrounding nameplate or model-name text caused by unspecified ROIs, and missing decimal points. Our quantitative evaluation demonstrates that directly applying OCR to digital gauge reading yields only 46.9% accuracy. Accordingly, this study aims to overcome the limitations of conventional OCR by leveraging the ability of small vision-language models (sVLMs) to read values in context. Recently, Lin et al. [24] proposed MeasureBench, which systematically benchmarks the gauge reading capability of VLMs and shows that even state-of-the-art VLMs struggle with fine-grained measurement reading, with imprecise spatial grounding of pointers and scale markings as a dominant failure mode that motivates task-specific, context-aware adaptation rather than naive zero-shot inference.

2.3 Industrial Application of Vision-Language Models

Large-scale VLMs, trained on extensive web data, have demonstrated zero-shot inference capabilities in various visual tasks, leading to their rapidly expanding application in industrial safety and inspection. A comprehensive survey by Zhang et al. [25] categorized the development of VLMs into three methodological pillars: pre-training objectives (contrastive, generative, and alignment objectives), downstream transfer learning (prompt tuning and feature adapters), and knowledge distillation. The study quantitatively discusses the zero-shot applicability of a single VLM to cover a wide range of visual recognition tasks, extending from simple image classification to fine-grained object detection and segmentation. Awais et al. [26] investigated the lineage of vision foundation models and emphasized that textual and visual prompting achieve task adaptation without requiring additional parameter retraining. However, they identified critical unresolved challenges for practical industrial deployment, including vulnerability to hallucinations (where large models fabricate responses by relying solely on the prior knowledge of text prompts while ignoring visual inputs), a lack of understanding of complex physical laws in the real world, and bias stemming from training datasets.

To address these limitations, a growing body of research has demonstrated the practical applicability of VLMs in industrial safety fields with constrained computational resources. Xu et al. [15] combined generative AI-based data augmentation and object detection results with structured prompt inputs for VLMs to analyze PPE compliance in high-altitude work, recording a compliance reasoning accuracy of 87.6% with the Qwen2.5-VL:7B baseline. While large VLMs possess strong contextual reasoning capabilities, they are unsuitable for real-time edge monitoring due to excessive computational demands. Conversely, sVLMs with parameters under 4B are computationally efficient but suffer from decreased accuracy in complex construction environments and are prone to hallucinations that cause them to miss critical risk factors. To resolve this trade-off, Adil et al. [27] proposed a detection-guided framework that integrates the spatial precision of an object detection model (YOLOv11n) with the multimodal reasoning of an sVLM. By converting the spatial coordinate and class information of detected workers and construction machinery into structured text prompts to guide the reasoning of the sVLM, the F1-score for construction site risk detection improved from 34.5% to 50.6% when using the Gemma-3:4B model. This approach was demonstrated to reduce speculative false positives and substantially improve the semantic quality of the generated rationales (BERTScore) from 0.61 to 0.82, while introducing an inference overhead of only 2.5 ms per image. The few-shot adaptation capabilities of VLMs to address data scarcity in specialized domains are also being actively researched. In the industrial inspection domain, Ueno et al. [28] fine-tuned a VLM with web-collected product images and inspection-criteria texts, and applied in-context learning so that a single exemplar image with explanatory text could adapt the model to new products without retraining. Their method achieved an F1-score of 0.950 (Matthews correlation coefficient (MCC) of 0.804) on MVTec AD in a one-shot setting, demonstrating that minimal in-context learning enables rapid adaptation to specialized inspection targets. Moreover, training large-scale VLMs from scratch in resource-constrained research environments or specialized industrial domains requires prohibitive computational resources. To overcome this challenge, Cao et al. [29] proposed a generalized domain prompt learning framework that combines a small number of prompt samples (16 per class) with small-scale domain-specific foundation models. They demonstrated that natural-image-based VLMs could be successfully transferred to five specialized domains, including remote sensing, medicine, and geology, without significant computational overhead by leveraging quaternion networks and low-rank adaptation techniques. This finding aligns with our objective of maximizing the domain adaptation capability of VLMs under constrained computing resources, such as edge robotic environments. Regarding inference efficiency, Giedra et al. [30] benchmarked twelve compact open-weight large language models (LLMs) (sub-1B to 8B parameters) deployed using Ollama on an NVIDIA

Jetson Orin Nano, linking response quality with measured throughput, power, and energy efficiency. GPU-accelerated inference more than doubled the median throughput (from 7.12 to 18.13 tokens/s) while improving energy efficiency from 0.453 to 0.823 tokens/J, quantitatively confirming that carefully selected small models can satisfy the latency and power constraints of edge deployment. Guided by this efficiency analysis and the visual perception limitations of existing VLMs, we selected the Qwen-VL family [31], which features an architecture specialized for fine-grained visual understanding, text reading, and visual grounding, as the baseline model in this study. Qwen-VL combines a visual receptor with a language model and is trained through a three-stage training pipeline leveraging a multilingual multimodal corpus to enhance detailed geometric cognitive capabilities. Recently, in addition to Qwen3-VL:8B [32], various multimodal VLMs with optimized performance-to-size ratios, such as Qwen3.5-Omni [33] and Gemma 4 [34], have been developed. Considering both edge-device inference efficiency and fine-grained visual understanding capabilities, Qwen3-VL:8B is deployed as the primary on-device model. Furthermore, an optimized pipeline is proposed to minimize hallucinations and support reliable status reading and anomaly diagnosis in complex industrial equipment inspection environments.

While much of the existing industrial VLM application literature has focused on cloud-based large-scale models or binary classification, this study proposes a pipeline specialized for industrial qualitative diagnosis. This pipeline integrates visual question answering (VQA) based on a five-point rubric, 4× bicubic-upscaling-based hallucination suppression, and rubric-based three-way expert validation. This approach is distinguished from prior studies in that it outputs the degradation severity as a quantitative scale, extending beyond simple binary anomaly detection.

2.4 AI-Fused Spatial Alignment and Inference Optimization

In robotic inspection systems, spatial matching that precisely correlates equipment objects detected by a camera with inspection points in a database requires the seamless integration of 3D localization and object matching algorithms. Fan et al. [35] proposed a collaborative mapping framework based on ORB-SLAM3 for robotic navigation and localization in unknown environments, demonstrating improved map construction efficiency and real-time performance. In monocular depth estimation, deep-learning-based relative depth map estimation technologies have advanced, led by MiDaS [36]. Recently, highly generalized models that generate dense depth maps, such as Depth Anything 3 [37], have been widely used for 3D spatial reconstruction combined with camera extrinsic parameters. However, deep-learning-based depth models impose practical constraints in edge robotic environments where hundreds of objects must be processed in real time, owing to high inference latency, high computational resource consumption, and the difficulty of absolute scale recovery. For object detection-based spatial partitioning, edge computing environments with demanding real-time processing requirements necessitate a strict balance between model inference speed and accuracy. Ataei et al. [38] compared YOLOv7, YOLOv7-tiny, and the two-stage Faster R-CNN detector for wind turbine structural damage detection, reporting that YOLOv7 achieved the highest accuracy (82.4% mAP@50) among the three. Furthermore, according to a study by Khanam and Hussain [39] analyzing the latest YOLOv11 architecture, YOLOv11 introduces the more computationally efficient C3k2 block, a refined Cross Stage Partial (CSP) bottleneck, alongside a dedicated C2PSA block that enhances spatial attention in the feature maps, together restricting the parameter count while significantly enhancing feature extraction capabilities for small or occluded objects. Inheriting the development of the YOLO series initiated by Redmon et al. [40], this study constructed a YOLOv8n-based 68-class detection model, yielding a mean average precision (mAP@0.5) of 97.31% under training conditions with n = 11,000 images (Exp1). The proposed area-based Z-depth approximation ($Z_k \propto 1/\sqrt{A_k}$) exploits the geometric principles of the pinhole camera model—whereby identical objects appear larger when positioned closer to the camera—to

separate foreground and background objects solely based on bounding box area, bypassing the need for deep-learning-based depth models. Compared to the 44.1% error rate of the Y-axis grid baseline in an ablation study (n = 34), the proposed method recorded a 0% matching error across all scenes (Exp3), indicating that the foreground/background separation problem can be effectively addressed using only a computationally negligible geometric approximation.

Regarding label matching, approaches such as ontology matching with large language model embeddings [41] and subword-level embeddings for lexical matching [42] have been explored to resolve lexical discrepancies between heterogeneous systems. However, these methods are computationally expensive and may fail to ensure the deterministic behavior required for real-time inspection systems in edge environments. This study develops a four-tier cascade label matching engine that sequentially applies exact matching (Tier 1), suffix tolerance (Tier 2), explicit LABEL_MAP (Tier 3), and stem containment (Tier 4) following normalization preprocessing. This engine achieved a 0% matching error rate for all detected samples (Exp2), indicating that a deterministic matching system can serve as a practical alternative for industrial deployment without requiring complex embedding-based models.

2.5 Prognostics and Health Management of Critical Infrastructure

PHM is a systematic engineering field that integrates data acquisition, health indicator construction, fault diagnosis, remaining useful life (RUL) prediction, and maintenance decision support to guarantee the reliability and availability of complex systems [43]. Lei et al. [43] provided a comprehensive review of machinery prognostics covering the entire pipeline from data acquisition to RUL prediction, emphasizing that the quality and continuity of condition data acquired from monitored assets represent a fundamental prerequisite upon which all downstream prognostic processes depend. Si et al. [44] systematically reviewed statistical data-driven approaches for RUL estimation, clarifying that observable health indicators obtained through condition monitoring serve as the primary inputs for RUL models, and that the diversity and consistency of these indicators directly govern prediction reliability.

With the rapid advancement of artificial intelligence, data-driven methodologies have become the dominant paradigm for the PHM of industrial systems. Zhao et al. [45] surveyed deep learning applications for machinery condition monitoring, including autoencoders, deep belief networks, convolutional neural networks, and recurrent neural networks, demonstrating that deep learning provides powerful tools for extracting health-related features from high-dimensional sensor data. Zhang et al. [46] surveyed data-driven methodologies for predictive maintenance (PdM) across industrial equipment, classifying applications according to machine learning and deep learning algorithm types. They highlighted that most existing PdM research relies on public benchmark datasets of well-instrumented rotating machinery, while relatively few studies have addressed heterogeneous and multimodal equipment environments characteristic of actual infrastructure sites. Serradilla et al. [47] surveyed deep learning models for predictive maintenance across industrial domains, noting that data acquisition quality, continuity, and multimodality represent major bottlenecks that restrict the practical deployment of PdM in complex industrial environments.

A key challenge commonly highlighted by these survey studies is the data acquisition gap. In environments where thousands of heterogeneous inspection points—encompassing analog gauges, digital gauges, switches and LED indicators, and qualitative visual condition inspection items—cannot be cost-effectively instrumented with dedicated fixed sensors, the continuous and structured health data stream required for AI-based PHM pipelines is absent. Traditional data acquisition methods—such as fixed sensor arrays, periodic manual readings, or supervisory control and data acquisition (SCADA) system logs—are suitable for well-instrumented rotating machinery. However, they exhibit severe limitations in physically complex

and heterogeneous equipment environments where inspection methods and spatial arrangements vary significantly across zones.

This data acquisition gap serves as the primary motivation for the autonomous mobile inspection platform proposed in this study. By positioning the proposed AIIE as the data acquisition and condition monitoring backbone of the PHM pipeline, this study enables structured, continuous, and auditable health records across heterogeneous equipment types. These records provide the multimodal condition data streams necessary to support downstream fault diagnosis, anomaly trend analysis, and predictive maintenance decision-making in underground power ICT infrastructure.

3 Proposed System Architecture

This section describes the configuration and development details of the patrolling, inspection, and diagnosis technologies developed for a mobile robot, applicable to the underground infrastructure environment of the KEPCO Power ICT Center. Based on the inspection targets and phased automation plan established through the inspection environment analysis (Section 3.1), an inspection system integrating a robotic platform, sensor configurations, autonomous driving technology, and image-based AI analysis technology was developed. The system was implemented to enable the robot to autonomously navigate the underground infrastructure, inspect equipment conditions, and determine the presence of anomalies based on the acquired data. Rather than relying on commercial platforms, we prioritized developing inspection technologies optimized for power facility environments through a proprietary robot operating system and AI analysis framework.

3.1 Inspection Environment Analysis

We derived the patrolling, inspection, and diagnosis items for the underground infrastructure of the KEPCO Power ICT Center based on the patrol inspection manuals currently in use.

3.1.1 Derivation of Inspection Items and Data Analysis

As illustrated in Fig. 2, the underground infrastructure comprises 11 inspection zones: the energy storage system (ESS) room, UPS room, machine room, battery room, fuel tank room, generator room, fire extinguishing agent room, Halon fire extinguisher room, control room, electrical room, and other common areas. Among these, seven zones—the machine room, fuel tank room, generator room, battery room, electrical room, UPS room, and ESS room—were selected as robot-assisted inspection areas. For each selected zone, we reviewed the types, installation formats, inspection methods, and accessibility of the major equipment. Through this survey, a total of 1240 inspection points were identified in the underground infrastructure of the KEPCO Power ICT Center, and all collected data were used in the subsequent analysis. The main inspection areas and major equipment types are summarized in Table 2.

We defined inspection items in detail according to the actual unit of inspection execution, rather than calculating them solely based on the quantity of equipment. Even for the same equipment, if the inspection objectives differed—such as visual inspection, gauge reading, status indicator checking, leakage inspection, and environmental condition monitoring—they were classified as independent inspection points. This approach fully incorporated the actual work units performed in the conventional human-centric inspection system. We classified the derived inspection items into observation, reading, manipulation, and environmental measurement types according to their inspection methods and data characteristics, as summarized in Table 3. Observation-type items involve checking for visual anomalies, signs of leakage, corrosion and contamination, and the status of indicators. Reading-type items involve reading and determining gauge values, including circular analog gauges, digital gauges, and instrument displays. Manipulation-type items

require physical intervention, such as button operations, panel opening and closing, and equipment operation checks, whereas environmental measurement-type items are classified as those measuring surrounding environmental conditions, such as temperature, humidity, volatile organic compounds (VOCs), nitrogen oxides (NO_x), and noise.

This classification clarifies the sensing methods and work characteristics required for each inspection item, establishing a baseline for analyzing robotic applicability. A significant portion of the derived inspection points corresponds to repetitive and standardized tasks, such as visual inspections of equipment exteriors, gauge readings, and status indicator checks, examples of which are shown in Fig. 3.



Figure 2: Inspection map of the underground infrastructure of the KEPCO Power ICT Center.

Table 2: Inspection zones and major equipment in the KEPCO Power ICT Center.

Zone	Major Equipment	Inspection Item Type
Machine room	Pumps, pipes, valves	Visual, leakage, noise
Generator room	Generators, control panels	Gauge, status, presence of anomaly
Battery room	Batteries, racks	Visual, leakage, status
Electrical room	Molded case circuit breakers, panels	Gauge, status, degradation severity, indicator lights
UPS room	UPS panels	Status, alarms, indicator lights
ESS room	ESS panels	Temperature, status, presence of anomaly
Fuel tank room, etc.	Fuel tanks, pipes	Leakage, corrosion, status
Common	Thermo-hygrostats, fire extinguishers	Visual, gauge, temperature

Table 3: Classification criteria by type of inspection items.

Category	Definition	Key Contents
Observation	Visual inspection	Leakage, corrosion, contamination, status
Reading	Gauge/indicator value verification	Analog gauges, digital values
Manipulation	Physical intervention required	Buttons, opening/closing, operation, etc.
Environmental measurement	Sensor-based measurement	Temperature/humidity, gas, noise, etc.



Figure 3: Various forms of equipment including panels, gauges, and piping. (**a**) Pressure vessels and gauges (**b**) Halon fire extinguishers (**c**) Control panels (**d**) Switch and button (**e**) Multiple piping structures.

3.1.2 Classification of Inspection Points

Based on the analysis of robotic applicability for the derived inspection items, we classified each inspection point into immediately inspectable, development-required, or uninspectable categories according to the current state of technology. In this classification, we weighed practical field applicability alongside technical feasibility—robot accessibility to the target location, equipment visibility from the robot's viewpoint, lighting and spatial conditions for data acquisition, and suitability for repetitive automated inspection.

Immediately inspectable points refer to items that the robot can inspect using current technology without additional development. This primarily includes inspection tasks that can be performed in a non-contact manner, such as visual verification, gauge reading, status indicator checking, and environmental data acquisition. Representative examples include checking for anomalies on equipment exteriors, inspecting for leakage and contamination, checking alarm and indicator states, reading analog and digital gauges, and measuring environmental data such as temperature and humidity. These items feature repetitive and standardized characteristics, representing areas where the automation effect is expected to be greatest, as the robot can determine states while navigating along a predetermined path.

Development-required inspection points refer to items that have robotic feasibility but require additional technical development or improvements in equipment and environments. Representative examples include inspections requiring the opening and closing of equipment panels and racks, inspections of equipment located in narrow spaces or shaded areas, zones where image-based recognition is challenging due to unfavorable lighting conditions, and interface manipulation items designed for human operators. This is because existing equipment was designed assuming human-centric inspection. Uninspectable points refer to areas where robotic application is challenging given the current state of technology and field conditions, making human-centric inspection unavoidable at this stage. This applies to cases where direct contact, precise manual operations, or internal equipment verification are essential due to the nature of the tasks. Examples include direct measurements using testers, internal inspections after opening equipment, continuous button operations, and complex field responses by operators (Fig. 4). This classification serves as a criterion for clearly establishing the current limitations of automation.

3.1.3 Prioritization of Robot Application and Definition of Inspection Targets

Based on the inspection point classification, we established the priority of robot deployment and defined the scope of automation targets through a phased strategy that considers not only technological feasibility but also on-site effectiveness and operational efficiency. Robot deployment was first prioritized for immediately inspectable items, which account for approximately 57.8% of the total inspection points. Among these, priority went to items with high inspection frequency, items difficult to perform during nighttime or under limited manpower conditions, and items involving hazardous or low-accessibility environments, where automation yields substantial safety benefits.

We defined development-required points as mid-to-long-term targets and divided them into two phases: a technology-supplemented implementation phase, covering items resolvable through relatively simple improvements such as adding sensors or improving imaging conditions, and an environment-improvement-based expansion phase, covering items that require facility-side changes such as automatic opening and closing devices, interface improvements, and structural accessibility for robots.

Uninspectable items were excluded from the current automation scope and remain human-centric inspection areas; their automation feasibility will be continuously reviewed alongside future technological advancements and structural improvements of the facilities.

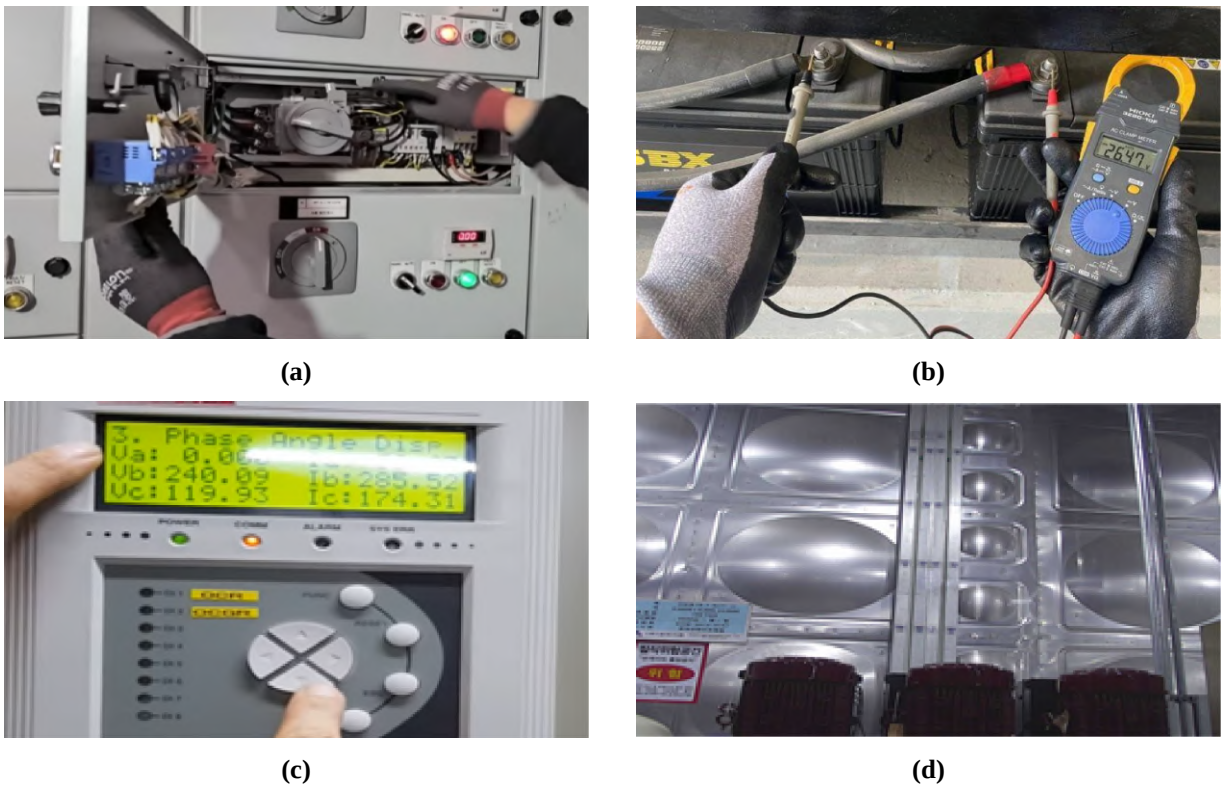


Figure 4: Key uninspectable inspection items. (**a**) Panel internal inspection (**b**) Direct measurement (**c**) Button operation (**d**) Long-range inspection.

3.1.4 Inspection Scope and Coverage Analysis

The seven zones selected for robotic inspection (Fig. 2) each follow different inspection items and management standards depending on the function and operational purpose of their equipment.

The underground infrastructure exhibits high equipment density and distinct environmental characteristics for each zone. Some zones may experience high-temperature, high-humidity conditions or generate gas and noise, and tight clearances between equipment or cramped passageways are common. These structural characteristics directly influence mobility and accessibility during inspection. Because inspection locations vary—including panel interiors, floors, and the upper parts of equipment—inspection requires repetitive movement and diverse viewpoints.

We classify the inspection targets of the KEPCO Power ICT Center underground infrastructure by equipment type, as presented in Table 4. Among the total 1240 inspection points, 717 (57.8%) were set as autonomous robot inspection targets, while the remaining 523 (42.2%) are inspected in parallel by human operators due to physical access constraints (e.g., inside sealed cabinets and high-altitude equipment).

The KEPCO Power ICT Center is equipped with a monitoring system in the disaster prevention room that allows remote verification of operational information and status data for some equipment. However, in actual operations, patrol inspections to directly verify on-site conditions based on these data are performed concurrently, and inspection results are managed manually or using spreadsheet software. In addition, on-site inspections include items that are difficult to verify solely through the central monitoring system, such as equipment appearance anomalies, signs of leakage, corrosion and contamination, and the status of

instruments and indicators, going beyond simple numerical checks. Consequently, numerous inspection items are still performed through direct visual inspection and judgment by operators.

Table 4: Inspection point coverage by equipment type.

Equipment Type	Total	Robot Inspection	Reading Method
UPS (uninterruptible power supply)	320	210	Analog gauge (AG) + digital gauge (DG)
Battery/Charger	280	180	DG + SW/LED status determination
Cooling Facilities	240	140	AG + VLM qualitative diagnosis
Switchgear	210	120	SW/LED + DG
Others (Environmental measurement)	190	67	Environment sensor hub
Total	1240	717 (57.8%)	—

As described above, the underground infrastructure of the KEPCO Power ICT Center has a structure where diverse equipment and complex environmental conditions coexist, characterized by clearly distinguished scopes and attributes of the inspection targets. Furthermore, a complementary equipment status inspection system is maintained through an operating framework where the central monitoring system and on-site patrol inspections are executed in parallel.

3.2 KEP-ROS: Open Multi-Robot Operating Framework

Although commercial quadruped robot platforms offer high mobility performance, proprietary payload policies and overseas-dependent maintenance create practical barriers to customized functional expansion and long-term cost efficiency on-site. To overcome the limitations of such closed operational structures, we designed and implemented KEP-ROS (KEPCO Robot Operating System), an open multi-robot operating framework.

KEP-ROS is built around three core design principles and remains independent of specific robot platforms. The first, protocol abstraction, integrates the heterogeneous communication protocols of gRPC-based quadruped robots and ROS2-based general-purpose robots into a single layer, enabling upper applications to operate regardless of the robot type. The second, mission lifecycle automation, automates the entire process from inspection schedule registration to robot navigation, data collection, and result storage via a distributed asynchronous task pipeline, thereby minimizing human intervention. The third, field-customized scalability, designs all AI inspectors and sensor modules as independent components that can be replaced or added, allowing flexible adaptation to changes in the equipment environment.

To position these design principles against a representative commercial platform, Table 5 compares the key functional differences between the Boston Dynamics Spot platform and KEP-ROS.

Table 5: Comparison between Boston Dynamics Spot platform and KEP-ROS.

Category	Boston Dynamics	KEP-ROS
AI/Sensor scalability	Only proprietary payloads allowed	Flexible expansion of customized modules
Precision control	Closed API, customized control unavailable	Fully open, direct gRPC control
Payload	Only Boston Dynamics-compatible products supported	Flexible application of general payloads
Maintenance/repair	Dependent on overseas repairs, incurring high costs	In-house maintenance feasible
Video streaming	WebSocket	WebRTC
AI specialization	General-purpose AI	Facility-specialized AI models
Interoperability	Dependent on a single platform	Heterogeneous integration with ROS2 bridge

The primary differences between the two platforms lie in scalability and cost structure. While the Boston Dynamics Spot officially supports only manufacturer-specific payloads, which leads to a rapid increase in cost when adding customized sensors and AI modules, KEP-ROS achieves equivalent functionality by using customized payloads. By providing a fully open control interface, KEP-ROS enables precision control essential for on-site inspections—such as waypoint stops, pan-tilt-zoom (PTZ) control, and data acquisition triggers—and its low-latency streaming supports immediate situational awareness in remote monitoring environments. By replacing the structural vulnerability of relying on overseas repairs in the event of failures with an in-house maintenance system, KEP-ROS secures operational continuity for the long-term operation of the power infrastructure.

The implementation of the KEP-ROS system comprises five core components. The technology stack includes a FastAPI + Celery/Redis backend, Spot SDK (gRPC) and ROS Bridge dual-protocol robot control, a React 19 + Three.js 3D dashboard, a relational database, and Docker Compose container orchestration. The operational workflow follows an eight-stage circular structure: mission planning → scheduling → execution → data acquisition → data transmission → AI analysis → result monitoring → safety management. The database is structured with eight core tables: robot, mission, schedule, mission_result, inspection_point, inspection_data, inspection_result, and event_log. The multimodal sensor configuration integrates RGB cameras, thermal imaging cameras, and Internet of Things (IoT) environmental sensors for temperature, humidity, VOCs, NO_x, and noise, where environmental data serve as a prerequisite for determining the safety of the inspection zone. Operational strategies include zone-based autonomous patrol, map-based pose estimation and path planning, real-time obstacle avoidance, and schedule-based automatic circular mission management. Details on the system design and implementation are provided in Supplementary Material S1.

4 Computational Modeling and Methodology

4.1 AI Diagnosis Pipeline Architecture

We propose a three-layer AI diagnosis architecture that links inspection standard definition, data distribution, detailed diagnosis, and result storage in a single flow. The overall structure is divided into three layers: the Directive Layer, Orchestration Layer, and Execution Layer, which are functionally separated such that the inspection standards of the upper layer are sequentially transmitted to the actual analysis of the lower layer. In particular, this architecture is designed to maximize the reliability and maintainability of the system by clearly separating deterministic logic and probabilistic AI inference. The overall AI diagnosis pipeline consists of eight steps, as illustrated in Fig. 5.

The Directive Layer is the highest-level command layer that defines customized inspection standards and acts as a trust anchor for the entire inspection process. Standard information, such as missions for each inspection zone, diagnostic items for each location, inspection types, and normal ranges, is stored in this layer, while inspection points, missions, and schedule information for each site are maintained and referenced by the system. The inspection standards defined in this layer are directly transmitted to the task routing of the Orchestration Layer and the threshold values of the inspectors in the Execution Layer, thereby establishing a deterministic control foundation for the entire pipeline.

While safety inspections in industrial sites require strict rule-based operations that exclude non-deterministic elements, AI models derive results based on statistical probabilities and therefore carry inherent uncertainty. By clearly separating these two elements, the Directive Layer provides the following system engineering advantages:

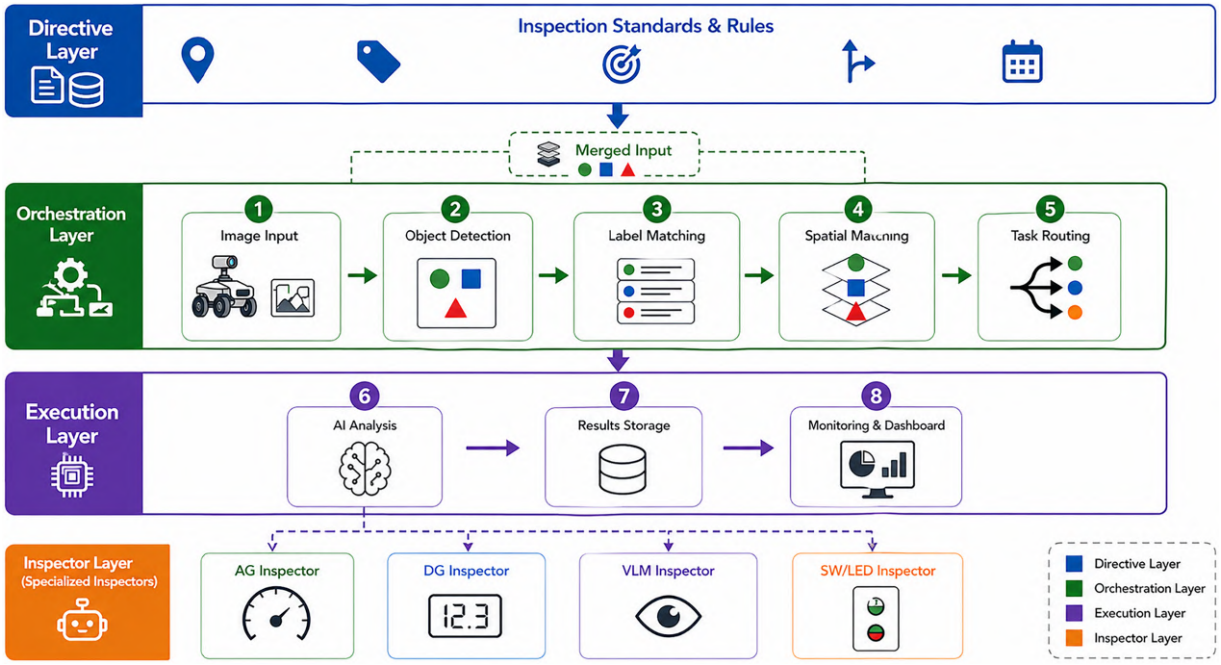


Figure 5: Overall flow of the AI inspection and diagnosis pipeline.

First, regarding the centralized management of operational standards, without this layer the threshold values would have to be hardcoded inside each AI model. For instance, if the normal range of a pressure gauge needs to shift from 1.2–1.8 to 1.0–1.6 MPa under equipment degradation, conventional architectures would require modifying the hardcoded logic and redeploying the entire system. In contrast, in the proposed architecture, modifications to the normal range in the Directive Layer are immediately reflected, requiring no changes to the AI models or robot control codes. In industrial environments where baseline adjustments occur frequently on-site, this architecture directly reduces operational costs and downtime.

Second, the architecture also preserves platform independence: if inspection missions are structurally dependent on a specific robot platform, replacing hardware would require redefining the inspection point locations, baseline values, inspection types, and the entire routing logic according to the new platform. In this architecture, because each inspection point P_i is defined with respect to the global world frame of the site rather than the coordinate system of a specific robot, the spatial and viewpoint information (L_i), heterogeneous data protocol (T_i), reference data governance (G_i), and analysis/verification logic (V_i) are reused without modification even if the robot platform is replaced. Even if a platform is replaced with another having a different body type or camera mounting height, only the viewpoint parameters require redefinition, while the Execution Layer's diagnostic logic is unaffected.

Third, system scalability benefits as well. Conventional approaches require direct insertion of custom detection logic into the code when new equipment is added to the site. In this structure, simply adding the metadata of a new inspection point (e.g., location, inspection type, and normal range) as a row in the database allows the Orchestration Layer to automatically recognize and process the point, without further code-level intervention.

Fourth, overcoming AI black-box limitations is another advantage: AI models can produce probabilistically different outputs even for the same input, making it difficult to trace the grounds of decision-making retrospectively. By externalizing the reference data governance (G_i) and analysis/verification logic (V_i)

delivered to each inspector as explicit metadata, the Directive Layer enables operators to verify the criteria upon which the AI based its decision directly at the log level. If the decision result for a specific instrument is questionable, the operator can query the G_i value of the corresponding point to immediately verify the normal range used by the AI. This secures the auditability of the AI decision and builds operator trust in the system. Each inspection point P_i is modeled as the following metadata set:

$$P_i = \{L_i, T_i, G_i, V_i\} \tag{1}$$

where L_i, T_i, G_i, and V_i denote spatial and viewpoint information, the heterogeneous data protocol, reference data governance, and analysis/verification logic, respectively.

The spatial and viewpoint information ($L_i \in \mathbb{R}^3$) includes three-dimensional (3D) coordinates and appropriate viewing-angle parameters to maximize data quality, functioning as a geometric constraint to avoid on-site interference and thereby ensuring a consistent imaging environment. The heterogeneous data protocol (T_i) defines control schemes for heterogeneous hardware, such as RGB, thermal imaging, and IoT sensors, and serves as a reference to chronologically synchronize the acquired multimodal data. The reference data governance (G_i) stores ground truth values calculated based on facility specifications and operational history, which serve as the criteria for converting AI model outputs into concrete pass/fail judgments. The analysis and verification logic (V_i) defines a routing key for allocating data to the appropriate agent and post-processing logic for verifying result integrity, filtering statistical outliers to enhance diagnostic precision.

4.1.1 Orchestration Layer

The orchestration layer serves as an intermediate layer that converts the inspection specifications of the directive layer into executable analysis tasks. It ensures that the images collected by the robot are transmitted to the appropriate AI inspectors based on the correct sequence and reference values. In the absence of this layer, operations such as image acquisition, object detection, inspection point matching, and agent allocation would be distributed across independent scripts, which would fail to ensure data consistency between consecutive stages. This layer integrates asynchronous task queues and spatial intelligence logic to process Stages 1–5 of the pipeline.

In Stage 1 (image input), when the robot arrives at the inspection point, measurement data, including visible light and thermal images, are transmitted to the analysis server. Immediately upon detecting the arrival of a new image, the corresponding data are registered as the QUEUED state and inserted into the task queue. Since the status from the queue registration to the completion of processing is recorded, pending tasks can be resumed upon reconnection despite communication failures.

In Stage 2 (object detection), the diagnostic system sequentially polls tasks in the QUEUED state to detect objects. Bounding boxes and class names of facility components, such as gauges, switches, LEDs, and valves, are extracted from the input image. These results serve as the input for the subsequent Stages 3–5. In environments with densely packed small-scale equipment, such as underground switchgear, the detection rate for small objects using the standard YOLO model decreases. To address this issue, we additionally trained a P2 head leveraging high-resolution feature maps to enhance small-object detection performance.

In Stage 3 (label matching), vocabulary discrepancies between the detection model's class names and the inspection framework structurally occur in typical industrial environments. To resolve this, we designed a four-tier cascade label matching engine; its rationale, algorithm, and ablation results are described in Section 4.2.1.

In Stage 4 (spatial matching), in scenes where front and rear components coexist, sorting based solely on 2D coordinates mixes foreground and background objects, causing the checklist's 1:1 mapping to fail. We resolve this with an area-based Z-depth approximation followed by grid sorting, as detailed in Section 4.2.2.

In Stage 5 (task routing), once label matching is completed, the cropped ROI images are automatically allocated to one of four analysis agents according to the routing key defined in V_i. Forcing a single general-purpose model to handle all inspection categories degrades the accuracy of individual tasks, because operations with distinct precision requirements (e.g., gauge reading and qualitative anomaly detection) must be processed by the same model. To prevent this performance degradation, the inspectors are partitioned into four categories: AG, DG, qualitative appearance (VLM), and discrete state (SW/LED) inspectors. The ROI image, reference data governance (G_i), and heterogeneous data protocol (T_i) are then encapsulated into a single task package and delivered.

4.1.2 Execution Layer

The execution layer performs the actual diagnosis through four specialized inspectors categorized by the characteristics of the inspection items, consistent with the task-specific partitioning rationale described in Stage 5.

In Stage 6 (AI analysis), each inspector operates as follows. The AG Inspector reads analog gauges through five-keypoint pose estimation, followed by homography projection correction and angle-to-value conversion (Section 4.3). The DG Inspector reads digital gauges, including seven-segment displays, via an sVLM (structured-prompt VLM; Qwen3-VL:8B) (Section 4.5). The Qualitative Inspector rates appearance states, such as corrosion, oil leakage, contamination, and insulation anomalies, on a five-point rubric scale (Section 4.6). The SW/LED Inspector determines the discrete On/Off, Auto/Manual, and illumination states of switches and LEDs through class-based object recognition (Section 4.4).

In Stage 7 (result storage), the readings of each inspector are stored as numerical and textual state values. The original and result image paths are recorded together to facilitate post-analysis and historical traceability.

In Stage 8 (monitoring and dashboard), the stored readings are visualized in real-time on a result verification dashboard, and items exceeding thresholds automatically trigger alarms.

4.2 Label Matching and Spatial Alignment

4.2.1 Four-Tier Hierarchical Label Matching

In industrial field deployments, the object detection model and the inspection framework that consumes its outputs are frequently developed by different teams at different times, or externally developed models are integrated retroactively. Because each model represents the same facility with a different class name, a vocabulary discrepancy in item names between the object detection model and the inspection framework is a structural problem that inevitably arises in the system design phase. This discrepancy does not stem from a single cause but originates from three structurally distinct issues.

First, for components such as LEDs and switches where the current state is embedded in the class name, the equipment item itself must be identifiable regardless of the detected state. For instance, knob switch objects may include state information such as Auto (SW_Knob_Auto) or Manual (SW_Knob_Manual); however, from an operational standpoint, the switch state changes dynamically over time. Thus, the item should be matched if it conforms to a generic SW_Knob_* template. If the inspection checklist defines the target simply as SW_Knob, matching fails for both states unless the state suffixes are removed, leading to a missing diagnosis.

Second, even if gauges perform the same function (having the same measured physical quantity and scale start/end points), class names representing the same equipment vary across models if class name notations differ by manufacturer or if different object detection models are trained with disparate naming conventions. In this scenario, recognizing identical equipment through simple string comparison is infeasible, rendering the distribution of diagnostic agents impossible.

Third, accumulated typographical errors or notation inconsistencies occurring during database registration of inspection points manifest as latent defects where matching for specific items persistently fails during field operations. Regardless of the underlying cause, a matching failure implies the omission of diagnostic results for the corresponding inspection point, and repeated omissions degrade the operational reliability of the entire inspection system.

However, because these three problems require different solutions, they cannot be addressed simultaneously using a single matching rule.

Accordingly, we designed a four-tier cascade method that sequentially applies rules starting from the highest precision. Because applying low-precision rules first can lead to mismatching among similar items, each tier is strictly configured to proceed to the next stage only when unresolved in the preceding step.

The first tier, exact match, verifies full matching after lowercase conversion and special character removal, prioritizing identical cases post-conversion to prevent redundant operations in subsequent stages. The second tier, suffix tolerance, verifies matching after removing state suffixes (such as _on and _off), identifying the inspection item itself regardless of the status at detection and thereby preventing matching failures for LED and switch components. The third tier, LABEL_MAP lookup, performs explicit transformation using a predefined mapping table, resolving structural discrepancies between gauges performing the same function under different manufacturer notations or between labels trained across different YOLO models. The fourth tier, stem containment, absorbs human errors or notation inconsistencies originating from the directive layer and is applied only to items that remain unresolved in the preceding three stages.

Once a match is confirmed in any stage, subsequent stages are bypassed immediately, preventing the application of lower-precision rules to already resolved items. This maintains matching precision while systematically processing the three causes of discrepancy, each with different characteristics, within a single pipeline.

The contribution and quantitative results of each stage are described in detail in Section 5.2.

4.2.2 Spatial Alignment: Z-Depth Approximation and Grid Sorting

In control panels within underground power ICT infrastructure, switches and indicator lights of the same type are occasionally arranged in overlapping front and rear rows. In this scenario, because the foreground and background objects are projected onto a single 2D image plane, determining the relative depth of each component using only YOLO-generated bounding boxes is challenging.

This issue stems primarily from the perspective effects and camera viewing angles when 3D physical space is projected onto a 2D plane. Owing to the viewpoint characteristics where the robot captures the facility either horizontally or from a lower angle looking upward, background equipment located far in the 3D space is projected near the vanishing point at the top of the image (corresponding to smaller Y values). As a result, the Y coordinates of front-row and back-row equipment can overlap or become reversed. In scenes containing both front and rear rows, the rightmost object in the front row and the leftmost object in the rear row may present highly similar Y coordinates on the 2D plane, leading to erroneous grouping into the same row. Under such conditions, simply assigning inspection points based on the X-axis sequence

causes foreground and background objects to be mapped to the directive layer's checklist in a mixed manner, leading to mapping discrepancies.

This study addresses the issue through a three-step process. First, we partition detected objects into front and rear pools using Z-depth approximation, with the bounding box area serving as a surrogate variable for the inverse perspective relationship. Subsequently, X-axis sorting based on grid order is performed within each row to complete the 1:1 mapping with the checklist items in the directive layer.

In a pinhole camera model, objects of identical physical size are projected with larger areas on the image plane as they are positioned closer to the camera. Since control panel environments with repeating layouts of identical switches and indicator lights satisfy this condition, we estimate the relative depth of objects using only the bounding box area A_k, without requiring additional depth sensors.

$$Z_k \approx \frac{f\sqrt{C}}{A_k^{1/2}} \tag{2}$$

Here, f represents the focal length, and C denotes the physical size calibration coefficient. Based on the relationship $Z_k \propto 1/\sqrt{A_k}$, a larger area implies that the object is located closer to the camera. Thus, sorting the bounding boxes in descending order of their areas is equivalent to front-priority sorting.

The separation of front and rear pools is performed as follows. Because a $(rear)$ tag is pre-inserted in the database entries of equipment located in the rear row within the directive layer, this tag is parsed to determine the expected number of front objects (N_{front}) and rear objects (N_{rear}) in the corresponding scene. Subsequently, the entire set of detected objects $\mathcal{D}$ is sorted in descending order of area based on Eq. (2). The top N_{front} objects are allocated to the front pool, and the remaining N_{rear} objects are assigned to the rear pool.

$$\text{Pool}(obj_k) = \begin{cases} \text{Front} & A_k \geq A_{(N_{front})} \\ \text{Rear} & A_k < A_{(N_{front})} \end{cases} \tag{3}$$

Within each pool where the front/rear separation is completed, objects are grouped into rows according to the grid layout of the panel, and sorted along the X-axis to establish a 1:1 correspondence with the checklist items in the directive layer. We determine rows using a tolerance-based approach, in which two bounding boxes are classified into the same row if the difference between their Y-axis center coordinates c_y is less than half the height h of the reference box.

$$\left| c_y(obj_i) - c_y(obj_j) \right| < \varepsilon_y, \qquad \varepsilon_y = 0.5\ h \tag{4}$$

This tolerance is designed to accommodate practical scenarios where objects in the same row do not align exactly along the Y-axis due to minor camera angle variations or uneven equipment installation heights. Objects classified into the same row are finally sorted in ascending order of their X coordinates.

The quantitative comparison results are described in detail in Section 5.3.

Consequently, the entire spatial alignment pipeline achieves a 1:1 mapping of checklist items in multi-layered layouts using solely 2D bounding box information, without relying on depth sensors or 3D reconstruction.

4.3 Analog Gauge Inspector (AG Inspector)

Although analog gauges serve as crucial indicators for intuitively conveying the operational state of power systems, implementing automated reading using autonomous robots faces two fundamental bottlenecks. First, conventional polar unwrapping techniques are effective only under frontal viewing conditions; they introduce non-linear scale errors on elliptically distorted dials during oblique imaging, which frequently occurs along robot paths. Second, the axis-aligned bounding box (AABB) employed in standard object detection exhibits structural limitations, as the proportion of background noise increases significantly during the rotation of thin, elongated pointers [16].

To overcome these limitations simultaneously, this study redefines gauge reading as a Structural Pose Estimation problem rather than simple object detection, and deploys an end-to-end pipeline consisting of three stages: keypoint detection, geometric calibration, and physical value conversion.

The geometric structure of the gauge is modeled as a topological skeleton $K = \{k_{start}, k_{mid}, k_{center}, k_{end}, k_{tip}\}$ consisting of five semantic keypoints. k_{start}, k_{mid}, and k_{end} represent the start, middle, and end positions of the scale, respectively, and establish the angular reference for the reading range. k_{center} denotes the mechanical rotation axis of the pointer, serving as the geometric origin for all trigonometric functions and the invariant reference point for virtual point generation. k_{tip} indicates the tip of the pointer, which is the keypoint requiring the highest sub-pixel precision to determine the measured value. Unlike the AABB approach, this keypoint-based method maintains the structural context regardless of the rotation angle of the pointer and effectively suppresses background elements (e.g., pipes and cables). To overcome the limitations of circular gauges that lack corner registration points, virtual points (VPs) are generated at locations point-symmetric to the scale start and end points with respect to the rotation center (k_{center}). This secures at least four correspondence points required for correction [16].

Based on YOLOv11-Pose, we optimized the network architecture by incorporating a high-resolution P2 feature layer with stride 4 and eliminating the P5 layer (stride 32) designed for large object detection. Because the minimum stride of standard YOLOv11 is 8, the resolution of the P3 feature map for a 1280 × 1280 input is only 160 × 160. By integrating the P2 layer, this resolution is doubled to 320 × 320, which enables the pointer tip, only a few pixels wide, to be regressed on a grid with four times more cells (i.e., twice the linear resolution). The removal of the P5 layer offsets the computational overhead caused by the P2 layer, which restricts the FPS degradation to less than 1%. Regarding the loss function configuration, the weights for the Object Keypoint Similarity loss and the bounding box loss were set to $\lambda_{pose} = 31.0$ and $\lambda_{box} = 7.5$, respectively. This configuration reflects the inherent nature of visual metrology, where minor coordinate errors of the pointer tip directly translate into errors in the final angular conversion [16].

4.4 SW/LED Inspector

SW/LED Inspector: YOLO Class-Based Discrete State Determination

Selector switches and LED indicator lights are discrete inspection points distributed across power facilities and switchboard panels, indicating the operating states of the equipment. Instead of introducing separate color analysis or classification models, this study employs a scheme that encodes state information directly into the class names returned by the YOLO detection model. Since the detection classes are structured in the format of {type}_{state} (e.g., LED_Red_on and SW_Knob_on), both object detection and state classification are performed simultaneously.

We determine state by comparing the class name specified during label matching with the inspection criteria type defined in the directive layer. To account for notational variations across state suffixes (on/off/opened/closed/left/right/center), the suffix-tolerance rule of the label matching engine (Section 4.2.1)

is applied. The final PASS/FAIL decision is determined by comparing the target state value specified for the object in the directive layer with the label returned by the detected object.

4.5 Digital Gauge Inspector (DG Inspector)

DG Inspector: sVLM-Based Numerical Reading of Digital Gauges

Numerical reading of digital gauges and variable-frequency drive panels faces structural limitations that are difficult to resolve using general-purpose OCR engines alone. A preliminary performance evaluation conducted during the design phase using PaddleOCR revealed two fundamental issues. First, because PaddleOCR indiscriminately detects all text within the input image, numbers on nameplates (e.g., 3/380/60) or model labels (e.g., H—0060) are misidentified as target inspection values. Second, because there is no mechanism to specify the reading target (e.g., cooling water temperature or booster pressure), domain-irrelevant text is incorporated into the reading results. With 1240 inspection points at this site, manually specifying target value locations for each point is highly impractical. This indicated that practical field deployment requires an additional ROI detector and a post-processing pipeline. We evaluated PaddleOCR in the design phase as a representative baseline for general-purpose OCR, and confirmed the validity of adopting the final sVLM through multi-backend comparison experiments involving five OCR engines and three VLMs (n = 130), as detailed in Section 5.5.

To address these issues, Qwen3-VL:8B, the sVLM, is adopted as the core reading engine. The sVLM is explicitly instructed on the values to read and the output format through structured prompts predefined according to the type of equipment under inspection. For instance, for a booster pump, the system provides a type-specific prompt that describes the target values and output format, such as: "Read the actual Set Pressure and Current Pressure. . .Output format: (Set—Current)". This approach disregards domain-irrelevant text while supplying the visual context required for recognizing seven-segment fonts and decimal points.

During preprocessing, low-resolution cropped images are enlarged by a factor of four using bicubic interpolation before model input to enhance pixel details. The backend supports multiple local and remote inference options for deployment flexibility. The response post-processing stage performs output formatting, such as separator standardization (pipe → comma) and semicolon removal.

4.6 Qualitative Inspector

Qualitative Inspector: Qualitative Appearance State Diagnosis

Among the equipment inspection items, appearance-based conditions—such as cleanliness, corrosion, oil leakage, and pipe insulation—are difficult to quantify. To process these qualitative items, the Qualitative Inspector uses a vision-language model based on Qwen3-VL:8B.

The system configuration (VLM_PROMPTS) maps each inspection type to a structured VQA prompt. The prompts are designed to strictly constrain the output format; for example, the specific VQA output requirements applied during image analysis are as follows:

(1) Cleanliness (1: Poor/contaminated ~5: Clean)
(2) Water leakage (1: Severe leakage ~5: No leakage)
(3) Oil leakage (1: Severe leakage ~5: No leakage)
(4) Corrosion/rust (1: Severe corrosion ~5: No corrosion)
(5) Insulation condition (1: Damaged/exposed sheath ~5: Undamaged)

The model infers the aging and damage state of the equipment based on the rubric-defined five-point scale. By processing the input image and prompt concurrently, the model goes beyond simple object

detection, returning semantically interpreted state results in a structured format. The inference backend offers similar deployment flexibility as the DG Inspector. In this study, we predominantly used Ollama.

Because the Qualitative Inspector performs direct inference on individually captured images without passing through the YOLO detection pipeline, its diagnostic reliability is not constrained by detection coverage; it is instead assessed through a three-way validation scheme. The rubric design, the four-expert blind evaluation protocol, and the quantitative validation results are detailed in Section 5.6.

5 Experiments and Results

Since the AI diagnostic pipeline sequentially couples object detection (Stage 2), label matching (Stage 3), spatial alignment (Stage 4), and inspector reading (Stage 6), errors from preceding stages propagate and accumulate through all subsequent stages, ultimately degrading the reliability of the final decision. Accordingly, we divide the evaluation into stage-wise validation (Experiments 1–3) to independently verify each pipeline stage, and inspector-specific validation (Experiments 4–6) to assess the reading accuracy of each inspector. Table 6 outlines the mapping between each experiment and its target pipeline stage, along with the respective validation objectives.

Table 6: Mapping between experiments and pipeline stages.

NO.	Experiment	Pipeline Stage	Validation Objective
Exp 1	Pipeline entry point detection model	Stage 2 (Detection)	Securing input quality for all subsequent stages
Exp 2	Four-tier cascade label matching	Stage 3 (Label Matching)	Matching of directive/orchestration items
Exp 3	Area-based Z-axis depth spatial matching	Stage 4 (Spatial Alignment)	Association between detected objects and locations
Exp 4	Analog gauge reading reliability	Stage 6 (AG Inspector)	Evaluation of analog gauge reading reliability
Exp 5	Digital gauge reading reliability	Stage 6 (DG Inspector)	Evaluation of digital gauge reading methodology
Exp 6	Qualitative appearance diagnosis	Stage 6 (Qualitative Inspector)	Evaluation of qualitative diagnosis reliability

5.1 Experimental Environment

We conducted all experiments in the underground environment of the KEPCO Power ICT Center. We employed Boston Dynamics' Spot as the robot (maintaining a battery state of charge (SOC) $\geq$ 30%), and used an NVIDIA Jetson Thor as the inference server. For VLM inference, we used Ollama (Qwen3-VL:8B, Q4 GGUF quantization).

Camera and lighting conditions: We employed a SpotCam PTZ (payload) camera with a resolution of 1920 × 1080 under an on-site illuminance of 15–366 lx (fluorescent-lit underground environment). Images for AG/DG reading were captured from a vertical frontal view at a distance of 0.5–1.5 m from the inspection points, while images for the Qualitative Inspector were captured with a wide-angle view to encompass the entire equipment.

Annotation procedure: AG GT values were obtained by manually reading the analog gauge values from Spot-captured on-site images. DG GT values were obtained by manually reading the digital display values from Spot-captured on-site images and recording them in Excel. GT labels for qualitative inspection items were generated using the five-point rubric as a structured scoring guideline. SW/LED GT labels were finalized after independent review by two annotators.

Train/validation split (YOLO): The total dataset was randomly split into training and validation sets at an 80:20 ratio. No separate held-out test split was reserved, and the reported mAP is based on the validation split.

Inference hardware: NVIDIA Jetson Thor (JetPack 7.1-b112, CUDA 12.6). We calculated the latency as the average of 20 runs under a warm model state.

In Stage 2 of the pipeline (object detection), objects are detected at the orchestration layer, and different detection paths operate independently for each inspector type according to the information from the directive layer. Table 7 summarizes the detection paths for each inspector.

Table 7: Detection model structure for each inspector.

Inspector	Detection/Analysis Model	Remarks
Classifier	(Detection) YOLOv8n: 68 classes	Analog gauge, digital gauge, switch/LED, fire extinguisher, etc.
AG Inspector	(Analysis) P2-YOLO-Pose	Keypoint-only model; see [16] for details
DG Inspector	(Analysis) Qwen3-VL:8B	Digital value reading using sVLM
Qualitative Inspector	(Analysis) Qwen3-VL:8B	Direct VQA inference based on separately captured images

The AG Inspector uses the common model for object detection and subsequently extracts keypoints using P2-YOLO-Pose [16]. As the Qualitative Inspector operates independently of the YOLO detection pipeline (Section 4.6), this evaluation targets only the common entry model, YOLOv8n (68 classes). The detection accuracy of this model dictates the input quality for subsequent label matching (Stage 3), spatial alignment (Stage 4), and inspector reading (Stage 6). Because undetected objects cannot be recovered in any subsequent stage, the entry model's performance directly establishes the upper bound of the overall coverage for each inspector path.

The training dataset comprises approximately 11,000 images (68 classes) of switches, LEDs, and gauge panels from the underground facility of the KEPCO Power ICT Center. In the training curves of Fig. 6, the mAP@0.5 converged rapidly after 20 epochs and showed a stable learning curve through the best epoch. The key performance metrics at the best epoch (60) are as follows:

- mAP@0.5: 97.31%
- mAP@0.5:0.95: 86.26%
- Precision: 90.41%, Recall: 96.85%

5.2 Label Matching

Because unresolved lexical mismatches between the orchestration and directive layers block inspector assignment and propagate as omissions through all subsequent stages, we evaluated the step-by-step contribution of the proposed four-tier cascade matching strategy (Table 8).

The four-tier cascade achieved a 0% match error rate: Tier 1 (exact match) alone resolved 74.0% of cases, and the remaining 12.7% and 13.3% were sequentially absorbed by Tier 2 (suffix tolerance) and Tier 3 (LABEL_MAP lookup), respectively. Tier 4 (stem containment) was not applied in any actual cases as all discrepancies were successfully resolved in the preceding stages, thus functioning purely as a safety net. These results demonstrate that all three causes of discrepancy identified in Section 4.2.1 are systematically absorbed within a single cascaded pipeline.

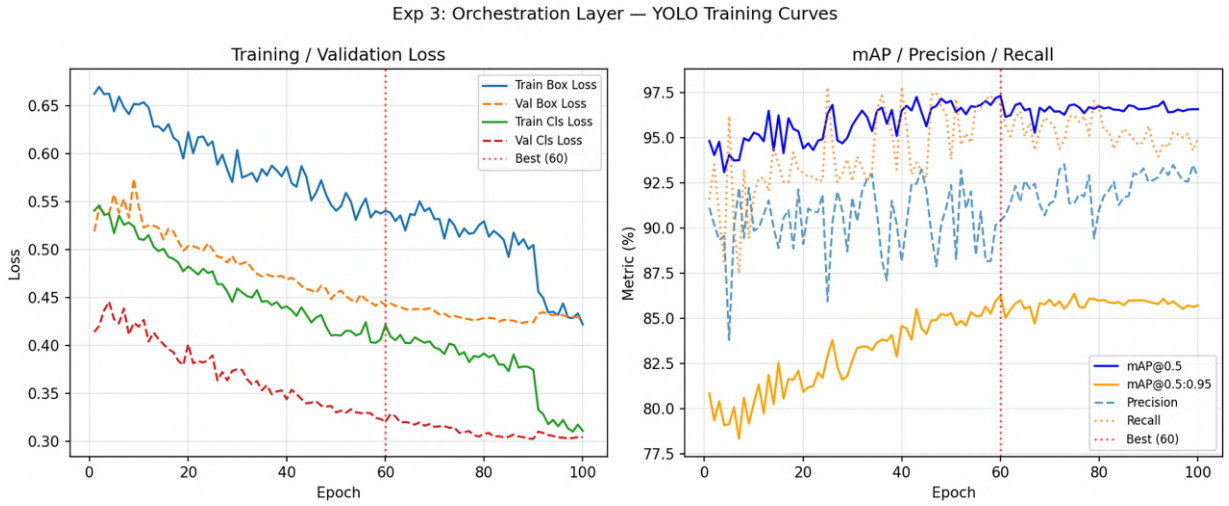


Figure 6: Classifier (YOLO) training curves.

Table 8: Four-tier cascade label matching: cumulative success rate.

Stage	Matching Strategy	Incremental Success	Cumulative Success	Conditions
Preproc	Canonical normalization	—	—	Lowercase conversion + separator removal
1	Exact match	74.0%	74.0%	Exact match
2	Suffix tolerance	12.7%	86.7%	Cases containing a state suffix
3	LABEL_MAP lookup	13.3%	100.0%	Cases registered in LABEL_MAP
4	Stem containment	0.0%	100.0%	Cases containing a core stem segment

5.3 Spatial Matching

We validated the area-based Z-depth approximation at four inspection points containing mixed front/rear layouts within the operational database. We applied three methods to perform spatial matching on the objects detected by YOLO at each inspection point (total $n = 34$): (i) Y-axis grid alignment alone (Baseline), (ii) the monocular depth estimation model MiDaS-small [36] (Baseline-2), and (iii) the area-based Z-depth approximation (Proposed) described in Section 4.2.2, after which we measured the match error rates.

In panel environments with co-located front and rear equipment, mismapped objects invalidate the inspection results even when detection and label matching succeed; the accuracy of spatial alignment (Stage 4) must therefore be evaluated separately from detection accuracy. This ablation study quantifies the reliability of each method under these conditions, as summarized in Table 9.

As an additional baseline, we employed the monocular camera-based depth estimation model MiDaS-small [36]. In MiDaS-small, a relative depth map is estimated from a single RGB image, and the front/rear order is determined based on the average depth value within each bounding box. Across the four scenes ($n = 34$), MiDaS-small exhibited a mismatch rate of 64.7%, indicating lower reliability than the Y-axis grid-only Baseline (44.1%). This result is attributed to the narrow depth range of the control panel environment and the presence of reflective metallic surfaces, which degrade the accuracy of the relative depth map generated by the monocular depth estimation model. In contrast, the proposed method achieved a 0%

mismatch rate across all scenes by leveraging only the reciprocal of the bounding box area ($z_{\text{proxy}} = 1/\tilde{A}_k$, where $\tilde{A}_k = A_k/(W \times H)$) without deploying an additional depth estimation model.

Table 9: Spatial matching ablation: Y-axis grid vs. MiDaS-small vs. area-based Z-depth ($n = 34$).

Method	Match Error Rate	Remarks
Y-axis grid only (Baseline)	44.1% (average)	Mismatches occurred in mixed front/rear scenes
MiDaS-small [36] (Baseline-2)	64.7% (average)	Monocular RGB → relative depth map-based front/rear classification
Area-based Z-depth (Proposed)	0%	No mismatches across all scenes

When analyzing the results for each scene in Fig. 7, we observe two distinct patterns.

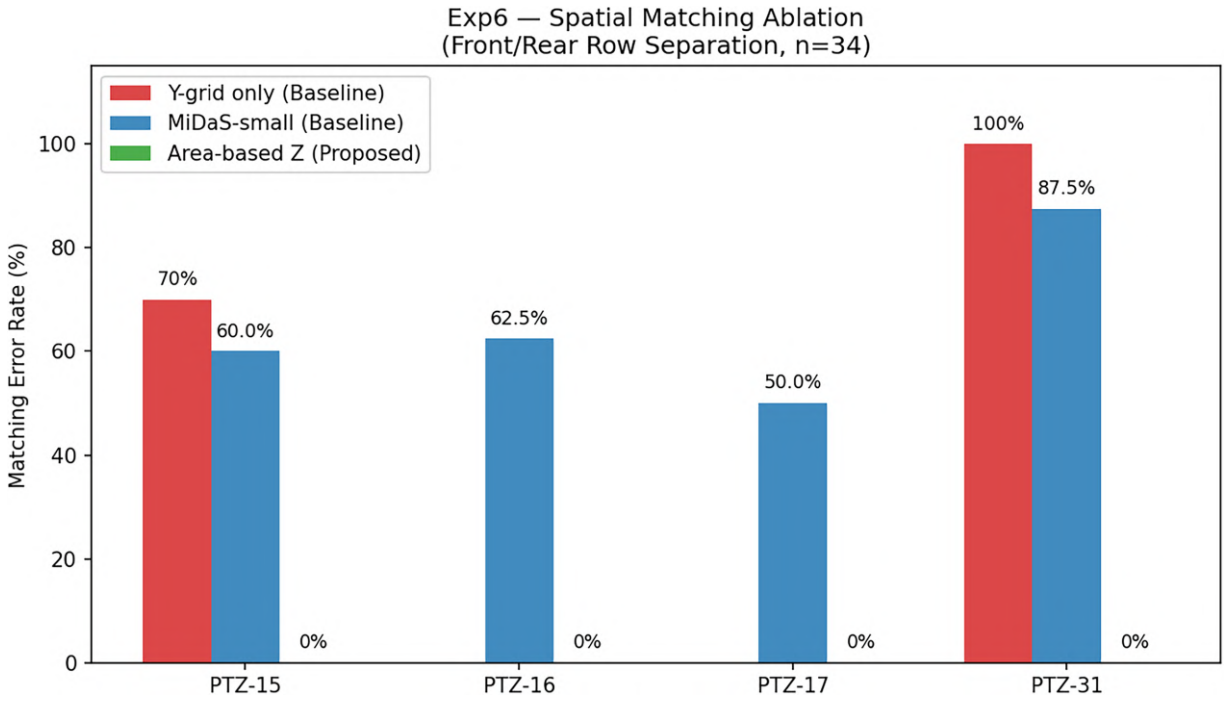


Figure 7: Spatial match error rates by scene ($n = 34$): Y-axis grid, MiDaS-small, and area-based Z-depth.

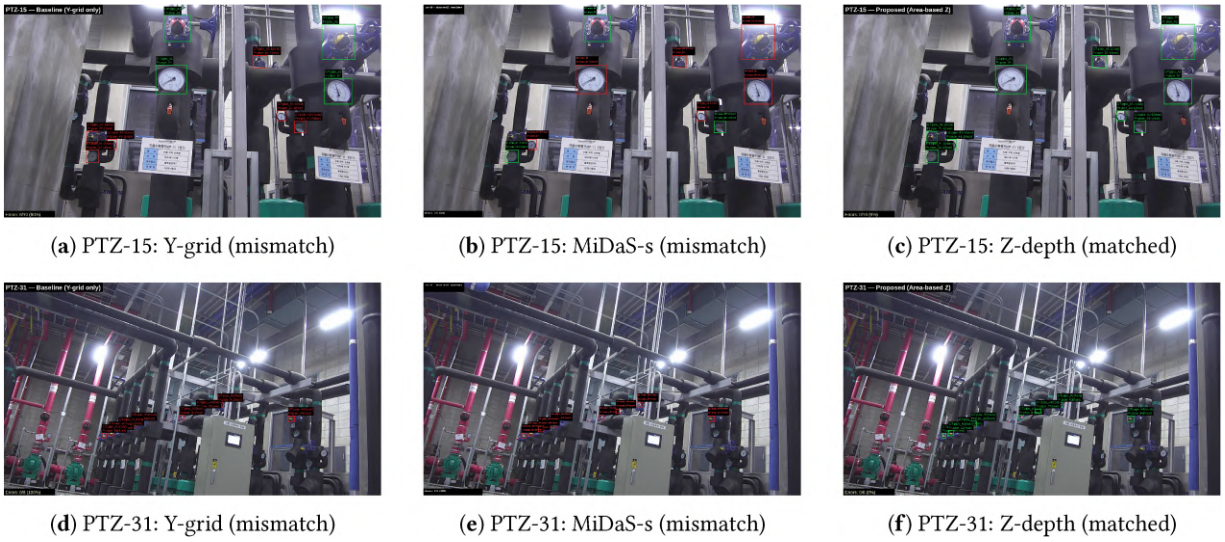
At inspection points PTZ-16 and PTZ-17, the front and rear objects were sufficiently separated along the Y-axis, resulting in a 0% match error rate for both the Y-axis grid Baseline and the Proposed method. This result implies that the Y-axis alignment method is not inherently unreliable; rather, it fails only under specific conditions where the Y-coordinates of the front and rear objects overlap. However, MiDaS-small recorded mismatch rates of 62.5% and 50.0% in the same PTZ-16 and PTZ-17 scenes, respectively. This failure occurs because MiDaS-small relies solely on its estimated depth map for alignment, without leveraging the domain-specific geometric regularity of Y-axis separation. The reliability of the depth map degraded on metallic pipes and black insulation surfaces, leading to mismatches even in scenes with sufficient Y-axis separation.

In contrast, at inspection point PTZ-15, the Baseline recorded a match error rate of 70%, and at PTZ-31, it exhibited a 100% error rate, indicating that all items in the scene were incorrectly matched. Because both scenes are characterized by overlapping Y-coordinate ranges between the front and rear equipment, the alignment order in the depth direction could not be distinguished using simple Y-axis alignment alone. The average error rate of the Y-axis grid Baseline across all four scenes combined was 44.1%, resulting in nearly half of the inspection points being assigned to incorrect equipment.

The proposed method maintained this 0% error rate even in the more challenging PTZ-15 and PTZ-31 scenes. Because the area-based front/rear pool partitioning (Section 4.2.2) precedes the X-axis alignment, this reliance on purely 2D bounding box information allows integration into the existing YOLO-based pipeline without requiring additional hardware and with a negligible increase in inference cost.

However, because this validation is limited to four scenes ($n = 34$) from a single facility, further validation is required to assess its generalizability to heterogeneous environments with different equipment densities and depth distributions. This limitation is detailed in the Future Work subsection of Section 6.

Fig. 8 presents a qualitative comparison of the three methods on the PTZ-15 and PTZ-31 scenes, where the quantitative differences of Table 9 are most pronounced.



(**a**) PTZ-15: Y-grid (mismatch) (**b**) PTZ-15: MiDaS-s (mismatch) (**c**) PTZ-15: Z-depth (matched)

(**d**) PTZ-31: Y-grid (mismatch) (**e**) PTZ-31: MiDaS-s (mismatch) (**f**) PTZ-31: Z-depth (matched)

Figure 8: Qualitative comparison of front/rear matching (Y-grid/MiDaS-small/Z-depth).

5.4 AG Inspector

The AG Inspector detects the pointer tip (k_{tip}) and the scale start/end points (k_{start}, k_{end}) using P2-YOLO-Pose-based keypoint extraction, and computes the reading based on the angular ratio among these three keypoints. The training and structural performance of the analog gauge inspection model are described in detail in a separate paper [16]. This experiment therefore focuses on validating the field reading accuracy of the system equipped with the model.

The experimental dataset consists of 200 analog gauge images collected from an actual underground power ICT facility. Because the mean absolute percentage error exhibits numerical instability, diverging when the GT value approaches zero, we selected the percentage of full scale (%FS), which normalizes the error against the full scale of the gauge range, as the evaluation metric.

We used bootstrap resampling (B = 5000) to obtain an overall %FS error of 1.66% with a 95% confidence interval (CI) of [1.31, 2.11] (Table 10).

Table 10: AG inspector: bootstrap %FS error by category and type (n = 200). MAE: mean absolute error.

Category/Type	*n*	Min	Max	MAE	%FS	Boot 95% CI
Pressure	102			0.1128	2.57%	[2.02, 3.34]
AG_Pressure01_P-0-1	20	0	1.0	0.0183	1.83%	[1.32, 2.35]
AG_Pressure02_P-0-1.5	20	0	1.5	0.0562	3.75%	[1.56, 7.20]
AG_Pressure03_P-0-1.5	20	0	1.5	0.0188	1.26%	[0.98, 1.56]
AG_Pressure05_P-0-16	10	0	16.0	0.2340	1.46%	[1.03, 1.90]
AG_Pressure06_P-0-1.5	12	0	1.5	0.0552	3.68%	[2.80, 4.55]
AG_Pressure_Fire-ext	20	0	10.0	0.3319	3.32%	[2.61, 4.10]
Environmental	38			0.9130	0.98%	[0.71, 1.25]
AG_Temperature_T-0-100	20	0	100.0	1.2483	1.25%	[0.89, 1.60]
AG_Thermo-hygro	18	−30	50.0	0.5405	0.67%	[0.34, 1.08]
Ammeter	60			0.6094	0.56%	[0.07, 1.20]
AG_Ammeter01_A-0-300	20	0	300.0	0.2333	0.08%	[0.00, 0.20]
AG_Ammeter03_A-0-100	20	0	100.0	0.0000	0.00%	[0.00, 0.00]
AG_Ammeter04_A-0-100	20	0	100.0	1.5948	1.59%	[0.16, 3.17]
Overall	200			0.4138	1.66%	[1.31, 2.11]

The top row of Fig. 9 presents representative reading examples for a thermometer (Fig. 9a) and a pressure gauge (Fig. 9b). Among the three categories, the pressure category (n = 102, 2.57%) exhibited the highest error. Even among pressure gauges sharing an identical scale range (P-0-1.5), an error discrepancy of approximately 2.5%p exists between AG_Pressure02 (3.75%) and AG_Pressure03 (1.26%). This finding suggests that the precision of keypoint estimation varies with the installation position and camera angle, even for the same gauge type. For the fire extinguisher pressure gauge (AG_Pressure_Fire-ext, n = 20, 3.32%), the keypoint estimation error increased because the fire extinguisher was tilted laterally during imaging; this case corresponds to Fig. 9c.

The ammeter category (n = 60, 0.56%) exhibited the lowest error among the three categories. This reflects the operational profile of power ICT patrol inspections, in which the target equipment is mostly inactive at the time of inspection, resulting in a majority of samples having a GT value of 0 A. When the needle is positioned at zero, k_{tip} is fixed near the scale start point (k_{start}), thereby stabilizing keypoint estimation. The virtually error-free performance of AG_Ammeter03_A-0-100 (0.00%) and AG_Ammeter01_A-0-300 (0.08%) demonstrates this effect. Despite having the same scale range (A-0-100), AG_Ammeter04_A-0-100 exhibited a relatively high error of 1.59%. This higher error is attributed to the inclusion of cases where GT ≠ 0, which increases the reading difficulty. The bootstrap CI of the ammeter, proportionally wide relative to its point estimate ([0.07, 1.20] around a mean of 0.56%), is due to the co-existence of samples concentrated at GT = 0 and samples with GT ≠ 0, which leads to a large %FS variance (0.00%–1.59%) among the types. The environmental category (n = 38, 0.98%) exhibited lower error than the pressure category, with the thermo-hygrometer (0.67%) showing lower error than the thermometer (1.25%).

The bottom row of Fig. 9 shows the detection results of the baseline methods for the same pressure gauge. Although the Hough-Circle method detected the gauge boundary as a circle and estimated the needle

direction (Fig. 9d), it failed to detect both the needle and the scale simultaneously in most cases. The ETH analog_gauge_reader detected the scale using circle fitting and marked the scale keypoints with red dots (Fig. 9e). Fig. 9f shows the needle segmentation result, which segments the needle region and estimates its central axis; however, both the needle and the scale were detected successfully in only 6.7% of cases.

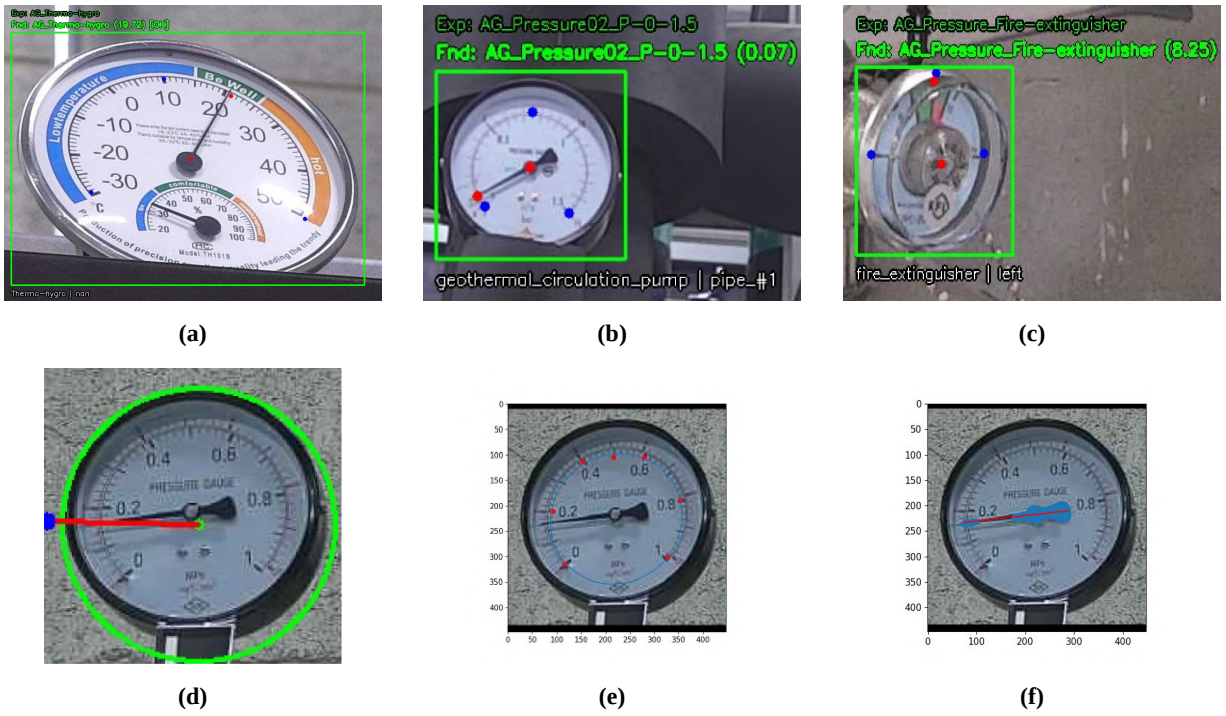


Figure 9: Examples of analog gauge readings and comparison with baselines. (**a**) Thermometer reading (**b**) Pressure gauge reading (**c**) Fire extinguisher pressure gauge (**d**) Hough-Circle (**e**) ETH scale keypoints (**f**) ETH needle segmentation.

For baseline comparison, we applied Hough-Circle [17] and ETH analog_gauge_reader [12] to the same dataset (n = 200), using the success rate of needle and scale keypoint detection as a common evaluation metric. Table 11 presents the results broken down by gauge category and sub-type.

In the "All" row of Table 11, Hough-Circle exhibits a success rate discrepancy of 39.5%p between scale detection (59.0%) and needle detection (19.5%).

Although identifying the gauge boundary using Hough circle detection is partially effective, needle angle estimation relies on simple line detection, making it difficult to distinguish the needle from background line segments. The ETH analog_gauge_reader exhibited low success rates for both needle detection (18.5%) and scale detection (17.5%). AG_Pressure06 exhibited 0% Hough scale detection due to an unusual gauge face geometry, yet ETH attained 83.3% needle detection on the same type, illustrating complementary failure modes. The AG Inspector recorded 99.5% for both needle and scale keypoint detection across all gauge types, and a fallback mechanism produced readings even when keypoint detection failed, yielding outputs for all 200 samples.

Table 11: AG inspector: baseline comparison by gauge category and sub-type (n = 200).

Category/Type	n	Hough		ETH Pipeline		AG Inspector	
		Needle	Scale	Needle	Scale	Needle	Scale
Pressure	102	6.9%	75.5%	27.5%	20.6%	99.0%	99.0%
AG_Pressure_Fire-ext	20	0.0%	90.0%	0.0%	5.0%	100.0%	100.0%
AG_Pressure01_P-0-1	20	5.0%	85.0%	30.0%	30.0%	100.0%	100.0%
AG_Pressure02_P-0-1.5	20	10.0%	60.0%	45.0%	10.0%	95.0%	95.0%
AG_Pressure03_P-0-1.5	20	0.0%	100.0%	0.0%	0.0%	100.0%	100.0%
AG_Pressure05_P-0-16	10	40.0%	100.0%	30.0%	50.0%	100.0%	100.0%
AG_Pressure06_P-0-1.5	12	0.0%	0.0%	83.3%	58.3%	100.0%	100.0%
Environmental	38	18.4%	34.2%	13.2%	7.9%	100.0%	100.0%
AG_Temperature_T-0-100	20	35.0%	65.0%	0.0%	15.0%	100.0%	100.0%
AG_Thermo-hygro	18	0.0%	0.0%	27.8%	0.0%	100.0%	100.0%
Ammeter	60	41.7%	46.7%	6.7%	18.3%	100.0%	100.0%
AG_Ammeter01_A-0-300	20	35.0%	40.0%	0.0%	5.0%	100.0%	100.0%
AG_Ammeter03_A-0-100	20	50.0%	45.0%	20.0%	15.0%	100.0%	100.0%
AG_Ammeter04_A-0-100	20	40.0%	55.0%	0.0%	35.0%	100.0%	100.0%
All	200	19.5%	59.0%	18.5%	17.5%	99.5%	99.5%

5.5 DG Inspector

The DG Inspector reads numeric values from the digital gauge ROIs detected by YOLOv8n. For a fair comparison, we applied identical bounding box (bbox) ROI cropping to all methods. A reading was scored correct if a number within ±10% of the manually parsed GT value appeared anywhere in the response (any-in-response, n = 130). Table 12 presents the per-label error breakdown for five OCR engines (Tesseract [19], TrOCR [22], EasyOCR [20], PaddleOCR [21], GLM-OCR [23]) and three VLM models (sVLM Qwen3-VL:8B [32], Qwen3.5:9b [33], gemma4:e4b [34]).

Consistent with the structural limitation identified in Section 4.5—indiscriminate recognition of all text within the ROI—OCR-based methods failed to distinguish target measurements from SCADA timestamps and panel labels; even with ROI cropping, PaddleOCR reached only 46.9%. By semantically specifying the target measurement through structured prompts, the current production sVLM (Qwen3-VL:8B) achieved 83.8%. Under the same evaluation criteria, gemma4:e4b (86.9%) and Qwen3.5:9b (84.6%) showed numerically higher accuracy, though the difference was not statistically significant, as detailed below; further improvements via backbone replacement are left for future work. The representative digital gauge reading results for each type are illustrated in Fig. 10.

Table 12: DG inspector: per-label error breakdown across OCR engines and VLMs (n = 130).

Method	Type	Label (n)								Total (130)
		AC (30)	Pump (20)	TR (25)	D-Meter (30)	DIM (8)	B-Pump (8)	Gen-S (5)	GIMAC (4)	
[OCR]										
EasyOCR	ok	1	1	8	1	1	0	3	0	15 (12%)
	rw	29	18	17	27	6	8	2	4	111
	nr	0	1	0	2	1	0	0	0	4
TrOCR	ok	0	0	0	0	0	0	0	0	0 (0%)
	rw	30	20	25	30	8	8	5	4	130
	nr	0	0	0	0	0	0	0	0	0
PaddleOCR	ok	18	7	15	14	2	0	1	4	61 (47%)
	rw	11	3	10	14	5	8	4	0	55
	nr	1	10	0	2	1	0	0	0	14
Tesseract	ok	3	0	1	2	0	0	0	0	6 (5%)
	rw	24	14	23	22	5	7	5	4	104
	nr	3	6	1	6	3	1	0	0	20
GLM-OCR	ok	30	16	22	20	0	0	0	4	92 (71%)
	rw	0	4	3	10	8	8	5	0	38
	nr	0	0	0	0	0	0	0	0	0
[VLM]										
gemma4:e4b	ok	30	17	25	19	5	8	5	4	113 (87%)
	rw	0	3	0	11	3	0	0	0	17
	nr	0	0	0	0	0	0	0	0	0
Qwen3.5:9b	ok	30	12	25	21	5	8	5	4	110 (85%)
	rw	0	7	0	7	3	0	0	0	17
	nr	0	1	0	2	0	0	0	0	3
Qwen3-VL:8B	ok	30	14	25	17	6	8	5	4	109 (84%)
	rw	0	6	0	6	1	0	0	0	13
	nr	0	0	0	7	1	0	0	0	8

Note: "ok" = GT in response; "rw" = all incorrect reads combined; "nr" = no read.

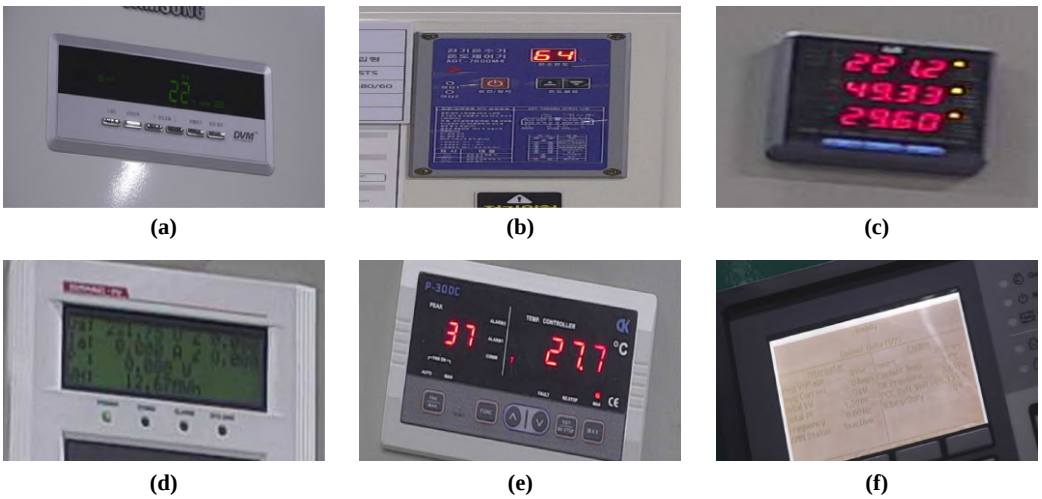


Figure 10: Digital gauge reading results of representative samples by type. (**a**) Air conditioner GT: 22/sVLM: 22 (**b**) Water heater GT: 64/sVLM: 64 (**c**) Pump GT: 221.2/sVLM: 221.2 (**d**) Integrated meter GT: 221.75/sVLM: 221.25 (**e**) Transformer temperature GT: 37/sVLM: 37 (**f**) On-site detection error GT: 38/sVLM: Unreadable.

We used Cochran's Q test and pairwise McNemar tests (Edwards' continuity correction) to statistically validate performance across all $n = 130$ samples, treating parse failures as incorrect reads (Table 13). Cochran's Q test confirmed significant overall performance differences among the eight compared methods ($Q = 541.5$, $df = 7$, $p < 0.001$). Pairwise McNemar tests demonstrated that the proposed sVLM (Qwen3-VL:8B; 83.8%, Wilson 95% CI: [76.6%, 89.2%]) significantly outperformed all conventional OCR-based methods—EasyOCR ($b = 94$, $c = 0$, $\chi^2 = 92.01$, $p < 0.001$), Tesseract ($b = 105$, $c = 2$, $\chi^2 = 97.23$, $p < 0.001$), and PaddleOCR ($b = 52$, $c = 4$, $\chi^2 = 39.45$, $p < 0.001$)—as well as the OCR-specialized VLM GLM-OCR ($b = 23$, $c = 6$, $\chi^2 = 8.83$, $p = 0.003$). No statistically significant difference was found against gemma4:e4b ($b = 3$, $c = 7$, $\chi^2 = 0.90$, $p = 0.343$) or Qwen3.5:9b ($b = 5$, $c = 6$, $\chi^2 = 0.00$, $p = 1.000$). TrOCR failed to extract any numeric value across all 130 samples (acc = 0.0%; Wilson 95% CI: [0.0%, 2.9%]); we excluded McNemar analysis due to a degenerate discordance matrix ($c = 0$).

Table 13: DG inspector: accuracy and statistical validation of OCR and VLM baselines ($n = 130$).

Method	Acc (%)	Wilson 95% CI	n	McNemar (b, c)	χ^2	p-Value
Qwen3-VL:8B (Reference)	83.8	[76.6, 89.2]	130	— (Reference)	—	—
gemma4:e4b	86.9	[80.1, 91.7]	130	$b = 3$, $c = 7$	0.90	0.343
Qwen3.5:9b	84.6	[77.4, 89.8]	130	$b = 5$, $c = 6$	0.00	1.000
GLM-OCR	70.8	[62.4, 77.9]	130	$b = 23$, $c = 6$	8.83	0.003
PaddleOCR	46.9	[38.6, 55.5]	130	$b = 52$, $c = 4$	39.45	<0.001
EasyOCR	11.5	[7.1, 18.2]	130	$b = 94$, $c = 0$	92.01	<0.001
Tesseract	4.6	[2.1, 9.7]	130	$b = 105$, $c = 2$	97.23	<0.001
TrOCR	0.0	[0.0, 2.9]	130	Insuf. power	—	—

Note: McNemar: Edwards' continuity correction; b = reference-only correct, c = baseline-only correct.

5.6 Qualitative Inspector

To ensure the evaluation reliability of the Qualitative Inspector, we introduced a three-way validation framework, establishing a five-point rubric for six categories—including cleanliness, water leak, oil leak, combined water/oil leak, corrosion, and insulation damage—and illustrating the evaluation criteria with 30 anchor images (Fig. 11). The combined leak (WO) category aggregates cases involving concurrent water and oil leakage for evaluation purposes; it is not an independent VQA output dimension. The full text of the rubric is provided in Appendix A. Subsequently, the $n = 30$ images were evaluated in a blind manner by four independent experts.

The analysis results of the three validation legs are presented in Table 14. In Leg A (inter-expert agreement), an average pairwise weighted kappa of 0.793 was obtained among the four experts, indicating the high reliability of the GT. Leg B (expert group vs. GT), which measures the agreement between the rounded average of the four experts' scores and the GT, achieved an overall κ_w of 0.937. The Leg B κ_w of 1.000 in the cleanliness, water leak, oil leak, and combined leak categories indicates that the rounded average of the four experts' scores perfectly matched the GT labels across all five levels, demonstrating the effectiveness of the visually unambiguous rubric anchors. For corrosion (R, $\kappa_w = 0.878$) and insulation damage (IS, $\kappa_w = 0.737$), the expert consensus did not reach perfect agreement with the GT owing to the inherent subjectivity in assessing damage severity. In Leg C (VLM vs. GT), which compares the final decision of the Qualitative Inspector with the GT, a Pearson r of 0.904 and a weighted kappa κ_w of 0.724 were achieved for all 30 samples, indicating the high qualitative diagnostic performance of the inspector.

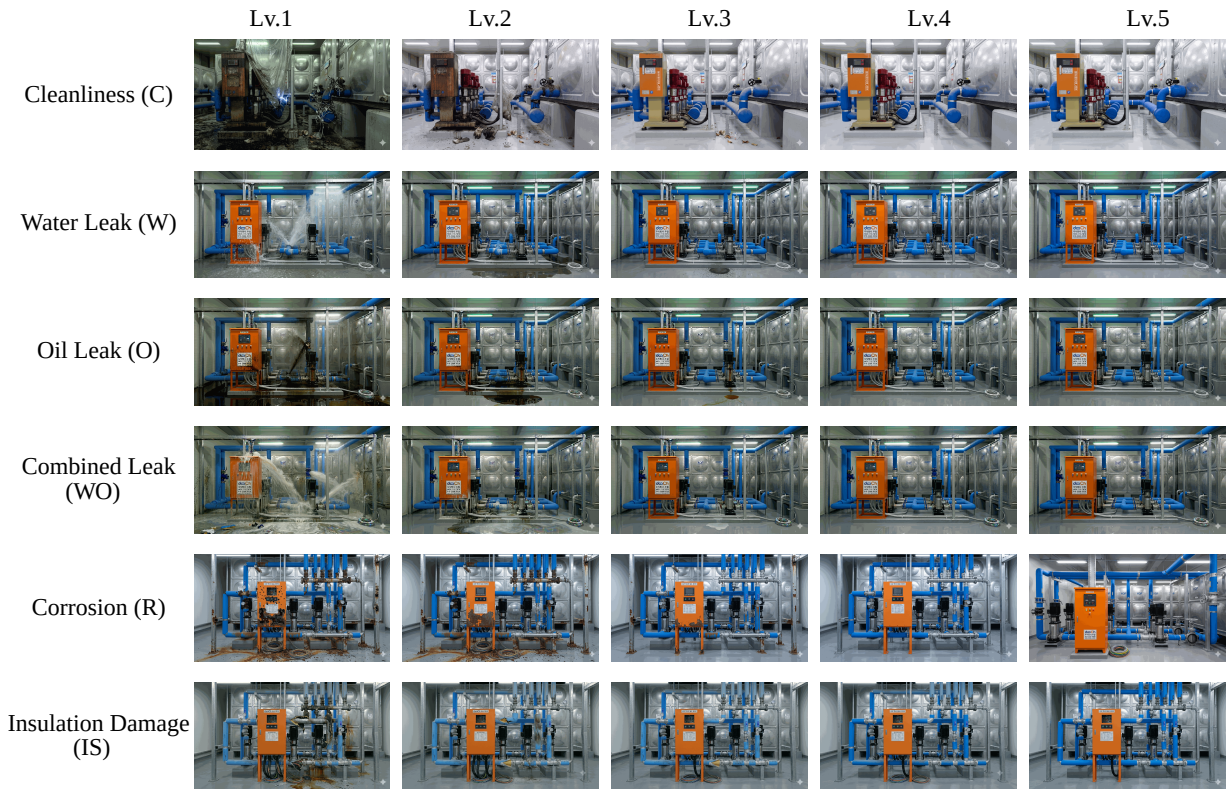


Figure 11: Rubric anchor images used for three-way validation (n = 30).

Table 14: Three-way validation results by category (κ_w, Pearson r).

Category	Leg A (κ_w)	Leg B (κ_w)	Leg C (κ_w)	Leg C (Pearson r)
Cleanliness (C)	0.867	1.000	0.634	0.866
Water Leak (W)	0.855	1.000	0.773	0.930
Oil Leak (O)	0.871	1.000	0.773	0.930
Combined Leak (WO)	0.816	1.000	0.773	0.930
Corrosion (R)	0.800	0.878	0.634	0.866
Insulation Damage (IS)	0.524	0.737	0.750	0.938
Overall (n = 30)	0.793	0.937	0.724	0.904
Category Average	0.789	0.936	0.723	0.910

In a preliminary comparison, prompting without the five-point rubric yielded r = 0.446, whereas incorporating the rubric into the prompt improved performance to r = 0.904. For corrosion (R), an inter-expert (Leg A) agreement of κ_w = 0.800 was achieved. Although the inter-expert agreement for insulation damage (IS) was improved to κ_w = 0.524 (moderate) after applying the rubric, the remaining subjectivity warrants cautious interpretation. The GT quality establishes an upper bound for AI performance evaluation; therefore, for items where rating criteria diverge even among experts, such as insulation damage, refining the GT labeling protocol is a prerequisite.

5.7 Overall System Performance Summary and Coverage

Taken together, these experimental results indicate that system reliability is determined by the cumulative accuracy secured independently at each pipeline stage. In the detection stage, the pipeline's input quality is ensured with an mAP@0.5 of 97.31%. By achieving a matching error rate of 0% in both label matching and spatial alignment, detection results are transferred to the inspectors without any item mismatch.

In the inspector stage, we verified the reliability of the AG Inspector (%FS 1.66%, n = 200), the DG Inspector (83.8%, n = 130), the Qualitative Inspector (r = 0.904, Leg C three-way validation), and the SW/LED Inspector (mAP@0.5 97.31%).

Table 15 summarizes the key performance indicators for the four types of inspectors, and Table 16 presents the overall inspection status of a total of 11,692 inspection trials.

Table 15: Performance summary of inspector agents.

Inspector	Key Indicator	Remarks
AG Inspector	%FS 1.66% CI [1.31, 2.11]	P2-YOLO-Pose keypoint; n = 200, Bootstrap B = 5000
DG Inspector	sVLM 83.8%	n = 130, identical bbox ROI crop, 8-way comparison
Qualitative Inspector	Leg C: r = 0.904, κ_w = 0.724	Three-way validation, n = 30
SW/LED Inspector	mAP@0.5 = 97.31%	YOLO-based classifier

Table 16: Summary of inspection status by inspection point.

Inspector	T1 (Success)	T2 (Detection Failure)	T3 (Scope-EXcess)	Total	Success Rate*
Analog gauge	738	164	0	902	81.82%
Digital gauge	792	193	40	1025	80.41%
Qualitative diagnosis	167	0	0	167	100.00%
SW/LED/Fire Extinguisher	8111	1032	455	9598	88.71%
Total	9808	1389	495	11,692	87.59%

Note: *Success rate: T3 (Scope-excess) is excluded from the population.

The 11,692 patrol inspections, accumulated across seven inspection zones over approximately one month of operational deployment, are classified into 9808 cases of T1 (automated inspection success), 1389 cases of T2 (detection failure), and 495 cases of T3 (scope-excess detection), yielding a success rate of 87.6% based on the registered inspection points (T1 + T2).

T1 denotes cases where the registered inspection points in the database were correctly detected and the readings were successfully obtained. T2 represents instances where the detector failed to locate the registered points, resulting in a 'Not Found' record. T3 corresponds to cases where the detection model successfully detected the object, but the object was excluded from evaluation as unregistered equipment in the database. Because T3 is attributed to the absence of management database entries rather than AI detection failure, it is excluded from the denominator of the success rate calculation.

The observed causes of T2 are as follows. First, owing to deviations in measured illuminance across zones (ranging from a minimum of 15 lx to a maximum of 366 lx), high-illuminance areas such as battery rooms (366 lx) and electrical rooms (211 lx) suffer from direct reflection of fluorescent lights on the front protective cover of digital gauges. This lighting variation and reflection lead to failures in boundary detection.

Second, positioning errors of the robot (approximately 10–20 cm) during waypoint stops cause targets to either fall outside the field of view or be occluded by surrounding structures. For certain facilities, consecutive

'Not Found' records accumulated over time because targets remained occluded by materials stacked over an extended period.

Third, in confined zones such as machine rooms and generator rooms (passage width of approximately 1.5 m), oblique lateral image capture causes aspect ratio distortion, which leads to failures in YOLO anchor matching and subsequent drops in confidence scores below the threshold.

Fourth, resolution limitations for small objects: those measuring a few centimeters or less, such as LED_Green-dot_on (83 cases) and SW_Knob_Left (75 cases), fail to be detected due to degraded image contrast and pixel sparsity in low-illuminance environments, including fuel tank rooms (15 lx) and UPS rooms (60 lx).

By zone, structural occlusion is the primary cause of failures in generator rooms (270 cases, 19.44%), whereas panel reflection interference is the dominant cause in electrical rooms (250 cases, 18.00%). As for reading-stage reliability, no VLM hallucinations were observed (0 cases) during the operational period; however, because they cannot be entirely eliminated, an operator verification process is recommended.

To verify the real-time applicability on edge devices, we empirically measured the inference latency and resource utilization of each module on the NVIDIA Jetson Thor over 20 runs (Table 17). The DG Inspector and Qualitative Inspector latencies reported in the table were measured under warm conditions, with models preloaded in memory. Under cold-start conditions (reloading the model after Ollama TTL expiration), the process could exceed 55 s; this latency is mitigated in operational environments by deploying keep-alive configurations or executing periodic warm-up calls.

Table 17: Per-module inference latency and resource usage (Jetson Thor, runs = 20).

Module	Latency (Mean ± Std)	Remarks
YOLO detection	(97.3 ± 5.4) ms	Classifier
AG Inspector	(36.5 ± 1.3) ms	—
SW/LED Inspector	(96.4 ± 2.8) ms	—
DG Inspector	6.8 s	Warm run
Qualitative Inspector	7.3 s	Warm run
DB insert	~0 ms	Based on SQLite
Power (Average/Peak)	42 W/68 W	tegrastats
Memory	11.4 GB	Including YOLO + VLM
Total patrol time	170 min	7 zones

6 Discussion

6.1 Effect of Multi-Expert Agent Collaborative Design

This study proposes distributing roles among expert agents equipped with algorithms optimized for specific tasks, rather than processing heterogeneous inspection tasks with a single general-purpose model. The performances of the AG Inspector (%FS 1.66%), DG Inspector (83.8%), SW/LED Inspector (mAP@0.5 97.31%), and Qualitative Inspector (Leg C $r = 0.904$, $\kappa_w = 0.724$) empirically validate the viability of this strategy. First: Heterogeneous multimodal integration and site-independent design. Unlike prior works that address a single sensing modality in isolation, the proposed system integrates four heterogeneous modalities—the AG, DG, SW/LED, and Qualitative Inspectors—into a unified pipeline, with each inspector attaining the performance levels reported above. Furthermore, the Directive–Orchestration–Execution architecture ensures site portability: deploying the system to a new facility requires only updating

the Directive layer (database registry of inspection points), while all AI models and orchestration logic remain unchanged.

Second: Validation in a real operational environment. We validated the proposed system using 11,692 autonomous patrol cases accumulated over approximately one month of operational deployment across seven inspection zones in an underground power ICT infrastructure. Under real-world conditions including illumination variation, physical occlusion, and deployment-induced noise, an autonomous inspection success rate of 87.6% was achieved based on the registered inspection points (n_{T1+T2} = 11,197; T1 = 9808). These results demonstrate the long-term operational reliability of the proposed system under real-world conditions that are typically absent from controlled laboratory evaluations.

Third: Structured database output and edge-autonomous execution. The inspection output is not a simple classification label but a relational database record structured at the inspection-point unit level, providing a continuous, auditable data stream for downstream PHM time-series trend analysis. The entire pipeline runs autonomously on a single edge board (NVIDIA Jetson Thor) without cloud dependency—a deployment constraint rarely addressed in cloud-dependent VLM-based inspection studies, but required for real-time operation in network-constrained underground environments.

The accuracy discrepancy between the general-purpose OCR (PaddleOCR) and the sVLM (Qwen3-VL:8B) in the DG Inspector (46.9% vs. 83.8%, under the identical ROI condition with n = 130; Section 5.5) does not merely reflect a difference in model performance; it originates from their contrasting design philosophies. The sVLM's semantic reasoning capability, reinforced by structured prompting, is key to overcoming the disambiguation limitation—distinguishing target measurements from SCADA timestamps and panel labels—that constrains OCR-based approaches.

6.2 Practical Implications of Area-Based Z-Depth Approximation

The reliable scene separation achieved by the proposed area-based Z-depth approximation across all scenes in the spatial alignment experiment (Exp3) carries two practical implications.

First, computationally, separating the foreground and background solely by calculating the bounding box area, without deploying a separate depth estimation model (e.g., MiDaS-small or Depth Anything 3), offers a computational advantage in real-time edge environments. Although deep learning-based monocular depth estimation entails an inference latency of tens to hundreds of milliseconds per frame, the area-based alignment requires only a sorting operation of $O(n \log n)$. Second, regarding generalizability, this approximation naturally holds because it reflects the domain characteristics of industrial control panel rooms, where identical types of equipment are arranged in a regular grid structure. However, this approximation assumes that the physical sizes of the front- and back-row equipment are similar. Separation errors may therefore occur in environments where heterogeneous equipment of different sizes coexists. As shown for the PTZ-16/17 scenes (Section 5.3), MiDaS-small's failure to leverage this geometric regularity resulted in mismatch rates of 62.5% and 50.0% even where the Y-axis grid alone sufficed.

Notably, the area-based Z-depth approximation is a site-specific methodology whose validity rests on the two conditions above—regular front/back-row grid arrangement and comparable equipment sizes—together with a camera viewpoint that captures both rows simultaneously. Porting the system to sites with modified camera positions, atypical arrangements, or heterogeneous equipment sizes therefore requires separately verifying these prerequisites, making immediate generalization challenging.

6.3 Qualitative Diagnostic Reliability of Small VLMs and GT Design Issues

The three-way validation of the Qualitative Inspector (Leg C, $r = 0.904$, $\kappa_w = 0.724$) suggests that small VLMs can perform fine-grained status assessments at a rubric-based five-point scale, extending beyond simple binary classification. Conversely, for the water leakage item, one expert evaluated even minor traces as defects, whereas another expert judged them as normal provided there were no operational issues, thereby confirming the discrepancy in the GT.

These issues offer important implications for the future construction of qualitative diagnostic systems. Before evaluating AI model reliability, the reliability of expert annotations must itself be validated: the cleanliness item's high inter-expert agreement (Leg A $\kappa_w = 0.867$) vs. insulation's low agreement ($\kappa_w = 0.524$) shows that GT quality varies by item, so clear anchor-image definitions and a prior agreement process are prerequisites for continuous-scale labeling.

6.4 Implications for AI-Based PHM Pipelines

The experimental results of this study demonstrate that the proposed AIIE constitutes a viable and scalable data acquisition and condition monitoring layer for AI-based PHM of underground power ICT infrastructure.

Within the PHM pipeline, the proposed system is responsible for the first two stages, namely data acquisition and condition monitoring, while simultaneously generating structured and timestamped health records required for subsequent prognostics and maintenance decision support. The quantitative readings generated by the AG Inspector and DG Inspector represent time-series operational parameters that can be directly input into anomaly detection or RUL prediction models. The discrete state output of the SW/LED Inspector provides binary health indicators suitable for fault diagnosis logic, and the rubric-based five-point scale qualitative scores generated by the Qualitative Inspector encode degradation severity in a structured format that facilitates trend analysis.

Specific PHM scenarios implementable using the proposed system are as follows. Through multiple autonomous patrol cycles, the AIIE accumulates a longitudinal database of heterogeneous equipment health indicators. Statistical process control or machine learning-based anomaly detection models applied to this database can identify gradual degradation trends (e.g., a continuous rise in coolant pressure gauge readings or a progressive deterioration of the insulation appearance score) and trigger preventive maintenance actions before functional failure occurs. The three-layer Directive–Orchestration–Execution architecture ensures that these monitoring scenarios can be reconfigured or extended to new equipment without modifying the underlying sensing agents. This supports PHM strategies that evolve as the infrastructure ages.

With these capabilities, the AIIE functions as a mobile sensing backbone for AI-based PHM in environments where dense stationary sensor deployment is impractical, extending beyond its role as an inspection tool.

6.5 Limitations and Future Work

We conducted this study at a single facility, the KEPCO Power ICT Center. The generalizability of the proposed system to various facility types, lighting conditions, and equipment configurations therefore requires further validation. Furthermore, deploying the system to new sites requires registering inspection points in the database and reconstructing patrol routes.

In addition, the following limitations exist: (i) performance validation under systematic low-light conditions was not conducted; (ii) manual inspection was maintained for 42.2% (523 points) of the total

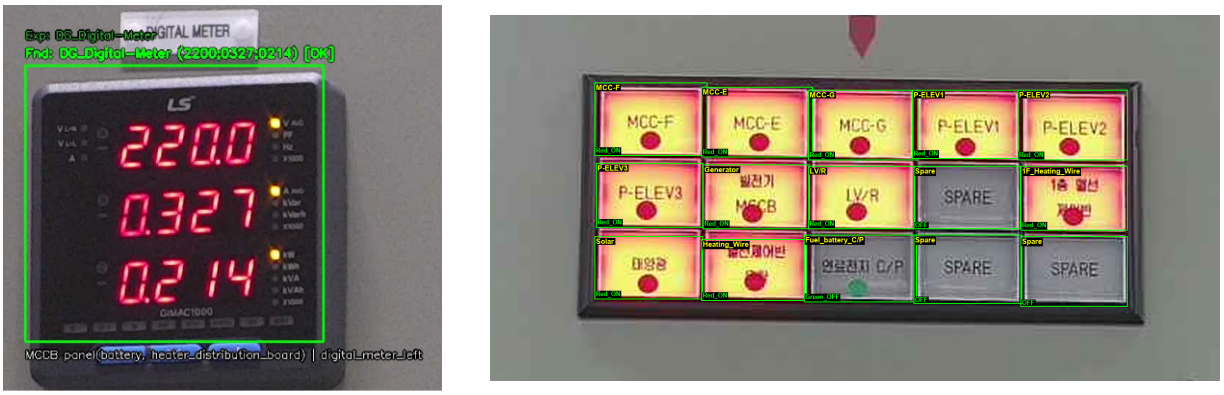
inspection points due to physical inaccessibility, structural constraints of the equipment, or the characteristics of the operational and environmental measurement items; (iii) a risk of VLM hallucination remains, which is mitigated through structured prompts but cannot be entirely eliminated, necessitating an operator verification process.

In particular, the reliability of the area-based Z-depth approximation requires separate validation in environments with different equipment densities and depth distributions, which we identify as a direction for future work.

It remains a future challenge to incorporate manipulative inspection capabilities—such as contact-based measurements or panel-opening and- closing operations—by modifying the robot's operating environment or deploying additional payloads.

Fig. 12 shows the challenges encountered during field deployment and directions for future improvement.

Analysis of the data collected in the operating environment revealed two major misrecognition patterns when reading 7-segment digital gauges, as illustrated in Fig. 12a. The first is the failure to recognize decimal points. Due to changes in illumination or subtle camera shake, the light-emitting decimal points of LEDs are either treated as noise or not clearly distinguished from the background, resulting in scale errors such as misreading '220.0' as '2200'. The second pattern is the over-segmentation error for characters with a narrow physical width, such as the digit '1'. In 7-segment fonts, the digit '1' occupies a very narrow pixel region compared to other digits and is displayed close to the right margin. The model therefore misidentifies it as a space rather than a continuous character, causing adjacent digits to be segmented into separate data fragments. These limitations remain to be resolved in future work through prompt engineering and preprocessing techniques specialized for gauge displays.



(a) **(b)**

Figure 12: Challenges and directions for future improvement in field deployment (**a**) Digital gauge readout error (GT: 220.0, 0.327, 0.214/sVLM: 2200, 0327, 0214) (**b**) Semantic linking of detected LED panel locations.

Another major limitation of the current implementation, shown in Fig. 12b, is the substantial human effort required to associate the physical locations of detected LED panels with the semantic information of the equipment.

The current implementation of the directive layer relies on a labor-intensive process in which administrators manually input metadata to map the operational states of each target panel to its corresponding text label (e.g., MCC-F and P-ELEV3). This manual mapping approach severely degrades maintenance efficiency when the inspection facilities are expanded or when panel layouts are modified.

Consequently, an approach employing VLMs is required to simultaneously interpret spatial layouts and visual text information contained in images. By utilizing VLMs to automatically recognize OCR information and equipment states on panels and link them to the semantic structure of a digital twin or an inspection database, the system can achieve intelligent semantic linking while minimizing manual definitions in the predefined directive layer.

This automated pipeline would therefore provide a foundation for generating and delivering high-level contextual information, such as "Anomaly detected in the solar indicator on panel P-ELEV3," eliminating the need for manual intervention when reporting inspection results to operators.

7 Conclusion

This paper proposed AIIE, a Physical-AI-based, PHM-enabling autonomous health monitoring platform for underground power ICT infrastructure, and validated its effectiveness in real operating environments including machine and electrical rooms. By integrating four specialized inspectors—AG, DG, SW/LED, and qualitative diagnosis—into a three-layer Directive–Orchestration–Execution microservice architecture, the system functions as a PHM sensing backbone that can be deployed in heterogeneous industrial environments without architectural redesign.

To address spatial matching challenges, this study proposed an area-based Z-depth approximation that separates front and rear equipment rows using only the reciprocal of the bounding box area ($z_{\text{proxy}} = 1/\tilde{A}_k$), bypassing monocular depth estimation. In ablation experiments ($n = 34$, 4 scenes), the proposed method attained a 0% matching error rate across all scenes, outperforming the Y-axis grid baseline (44.1%) and the MiDaS-small baseline (64.7%), including in scenes with sufficient Y-axis separation where MiDaS-small alone still failed (Section 5.3). The proposed method achieved this with no additional computational burden.

The individual inspectors achieved the following quantitative performance: the AG Inspector recorded a %FS error of 1.66% with a 95% bootstrap CI of [1.31, 2.11] ($n = 200$). The DG Inspector recorded a reading accuracy of 83.8% in comparison experiments involving eight methods ($n = 130$), a 36.9%p improvement over the conventional OCR baseline (PaddleOCR, 46.9%). The SW/LED Inspector achieved a status-determination accuracy of mAP@0.5 = 97.31% using a YOLO-based classifier. The Qualitative Inspector recorded inter-expert agreement of $\bar{\kappa}_w = 0.793$ (Leg A), expert-vs-GT agreement of $\kappa_w = 0.937$ (Leg B), and VLM-vs-GT agreement of Pearson $r = 0.904$ and $\kappa_w = 0.724$ (Leg C) in a rubric-based three-way validation, quantitatively demonstrating the feasibility of deploying small edge VLMs for qualitative industrial health assessments.

At the system level, a successful automated inspection rate of 87.6% was attained based on the 11,197 inspection cases involving registered inspection points (T1 + T2), collected among 11,692 real-world autonomous patrol cases. The remaining T2 cases are non-detections caused by constraints in patrol viewpoints and path coverage. This reflects a system coverage metric rather than classification errors of the detection models.

However, 523 of the 1240 inspection points (42.2%) still require manual inspection, and this validation remains restricted to a single site. Future research directions include: (i) expanding physical coverage; (ii) applying the area-based Z-depth approximation to heterogeneous sites and validating adaptive calibration; (iii) replacing the DG Inspector backbone with alternative VLMs such as gemma4:e4b (86.9%); and (iv) integrating anomaly trend analysis and preventive maintenance subtasks using long-term patrol data.

Acknowledgement: None.

Funding Statement: This work was supported by the Korea Electric Power Corporation (KEPCO) (Project No. R25IA04).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization: Jaekyung Lee and Wonhee Kim; Methodology: Jaekyung Lee and Byungsung Ko; Software: Jaekyung Lee and Byungsung Ko; Validation: Jaekyung Lee, Taewon Kim, Seoktae Kim, Jaeheon Park and Jiwon Lee; Formal Analysis: Jaekyung Lee and Byungsung Ko; Investigation: Jaekyung Lee, Byungsung Ko and Taewon Kim; Data Curation: Jaekyung Lee, Jaeheon Park and Jiwon Lee; Original Draft: Jaekyung Lee; Review and Editing: Jaekyung Lee, Jaeheon Park, Taewon Kim and Wonhee Kim; Supervision: Wonhee Kim. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used and/or analyzed during the current study are not publicly available due to confidentiality agreements with the Korea Electric Power Corporation (KEPCO), but can be obtained from the corresponding author upon reasonable request and with the approval of KEPCO.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmc.2026.085203/s1.

Appendix A Experimental Reproducibility Information

Appendix A.1 Dataset and Annotation

The training dataset (68 classes) consists of approximately 11,000 images of switches, LEDs, and gauge panels captured in the underground facilities at the KEPCO Power ICT Center, and the resulting detector achieved mAP@0.5 = 97.31%. The training and validation datasets were split in an 80:20 ratio. A single trained annotator performed image annotation using the X-AnyLabeling tool [48], excluding objects with less than 30% visibility.

Appendix A.2 Image Acquisition Equipment and Field Illumination

The acquisition equipment comprised a Boston Dynamics SpotCam Plus IR PTZ camera (2 MP, 1920 × 1080, 30× optical zoom, IP65 rated). We used only fluorescent ceiling lighting, without auxiliary illumination. The measured illuminance in each area was: machine room, 102 lx; generator room, 82 lx; fuel tank room, 15 lx; UPS room, 60 lx; battery room, 366 lx; electrical room, 211 lx; and ESS room, 172 lx.

Appendix A.3 Inference Hardware and VLM Operational Environment

The inference hardware and VLM deployment configuration follow Section 5.1. We deployed the VLM using Qwen3-VL:8B (Q4 GGUF, 8.8B parameters) through Ollama.

Appendix A.4 VLM Prompt Template

A representative example of the rubric-based VLM prompt used in the Qualitative Inspector (Class_Clean_Water_Oil_Rust):

Rate each item 1–5. RULE: 1 = worst problem, 5 = no problem.

Output format: (score|score|score|score). No other text.

(1) Cleaning State: 5 = Pristine, zero dust/debris|4 = Slight dust, no foreign objects|3 = Visible dust, minor debris away from equipment|2 = Heavy contamination, hazardous debris near equipment|1 = Conductive sludge on equipment, metallic debris on live parts
(2) Water Leakage: 5 = Completely dry, no water|4 = Surface sweating only, no drips|3 = Active drip streaks, drops forming|2 = Water pooling on floor|1 = Gushing or spraying from joint failure
(3) Oil Leakage: 5 = Completely dry, no oil|4 = Faint oil sheen on fitting only|3 = Oil drip streaks, drops forming|2 = Oil pooling on floor or tray|1 = Oil spraying or gushing from fracture
(4) Corrosion/Rust: 5 = Zero rust, paint perfect|4 = <5% surface area, pinpoint spots|3 = 5%–15% surface area, paint peeling|2 = >15% surface area, scaling|1 = Through-holes, structural failure

When evaluating a single category, only the prompt for the corresponding item is used. The rubric for Insulation State is as follows:

(5) Insulation State—look at ALL pipes/cables, score the WORST one: 5 = Perfect, smooth, intact, zero damage|4 = Slight fading, minor tape lifting|3 = Surface scratches or small seam crack, covering still attached|2 = Clear crack, inner material exposed|1 = Covering torn or missing, bare metal exposed, charring.

Appendix A.5 AG Angle-to-Value Conversion Procedure

Below is the pseudocode for the markerless homography correction and angle-to-value conversion of AG Inspector Stage 3.

Algorithm A1: AG angle-to-value conversion procedure

Input: Keypoints $\{k_{\text{center}}, k_{\text{start}}, k_{\text{end}}, k_{\text{tip}}\}$, $V_{\min}$, $V_{\max}$

Stage 3a—Markerless Homography Correction (VP-based):

$k'_{\text{start}} \leftarrow 2 \cdot k_{\text{center}} - k_{\text{start}}$ (Centrosymmetric virtual point)

$k'_{\text{end}} \leftarrow 2 \cdot k_{\text{center}} - k_{\text{end}}$ (Centrosymmetric virtual point)

$P_{\text{src}} \leftarrow \{k_{\text{start}}, k_{\text{end}}, k'_{\text{start}}, k'_{\text{end}}\}$ (≥4 corresponding points)

$H \leftarrow findHomography(P_{\text{src}}, P_{\text{dst}})$, P_{dst}: Reference points for front-view unit circle

Apply H to each keypoint → Obtain corrected coordinate k_r

Stage 3b—Angle-to-Value Conversion (Based on corrected keypoints):

$\theta_{\text{tip}}, \theta_{\text{start}}, \theta_{\text{end}} \leftarrow atan2(\cdot)$ (Counter-clockwise normalization)

$\Delta\theta(a, b) = (a - b) \bmod 360$ (Clockwise arc length)

ratio $\leftarrow \Delta\theta(\theta_{\text{start}}, \theta_{\text{tip}}) / \Delta\theta(\theta_{\text{start}}, \theta_{\text{end}})$

$V_{\text{result}} \leftarrow V_{\min} + \text{ratio} \times (V_{\max} - V_{\min})$

Output: V_{result} (Measured value in engineering units)

Appendix A.6 Z-depth Approximation Procedure

Below is the pseudocode for the area-based Z-depth approximation to resolve front/rear row ambiguity in spatial matching.

Algorithm A2: Z-depth approximation procedure

Input: Bounding box (x_1, y_1, x_2, y_2), Image dimensions $W \times H$

1: area $\leftarrow (x_2 - x_1) \times (y_2 - y_1)$

2: $\tilde{A} \leftarrow \text{area}/(W \times H)$ (Normalization, $[0,1]$)

3: $z_{\text{proxy}} \leftarrow 1/\tilde{A}$ (Large box = Close = Small z)

4: Sort z_{proxy} in ascending order → Separate front row (small z) and rear row (large z)

Output: z_{proxy} rank (resolution of front/rear matching ambiguity). This approach achieved a 0% matching error rate in 4 field scenes ($n = 34$ detections) (Table 9). Refer to Section 6.5 for limitations of the single-facility demonstration scale.

References

1. Halder S, Afsari K. Robots in inspection and monitoring of buildings and infrastructure: a systematic review. Appl Sci. 2023;13(4):2304. doi:10.3390/app13042304.
2. Lee AJ, Song W, Yu B, Choi D, Tirtawardhana C, Myung H. Survey of robotics technologies for civil infrastructure inspection. J Infrastruct Intell Resil. 2023;2(1):100018. doi:10.1016/j.iintel.2022.100018.
3. Wei X, Rey WP. Advancements in substation inspection robots: a review of research and development. E3S Web Conf. 2024;528(4):02017. doi:10.1051/e3sconf/202452802017.
4. Mineo C, Cerniglia D, Poole A. Autonomous robotic sensing for simultaneous geometric and volumetric inspection of free-form parts. J Intell Robot Syst. 2022;105(3):54. doi:10.1007/s10846-022-01673-6.
5. Konstantinidis FK, Balaska V, Symeonidis S, Mouroutsos SG, Gasteratos A. AROWA: an autonomous robot framework for Warehouse 4.0 health and safety inspection. In: Proceedings of the 30th Mediterranean Conference on Control and Automation (MED); 2022 Jun 28–Jul 1; Athens, Greece. p. 494–9.
6. Gehring C, Fankhauser P, Isler L, Diethelm R, Bachmann S, Potz M, et al. ANYmal in the field: solving industrial inspection of an offshore HVDC platform. In: Field and service robotics. Cham, Switzerland: Springer; 2021. p. 247–60.
7. Hutter M, Gehring C, Lauber A, Gunther F, Bellicoso CD, Tsounis V, et al. ANYmal—toward legged robots for harsh environments. Adv Robot. 2017;31(17):918–31. doi:10.1080/01691864.2017.1378591.
8. DeepRobotics. Industrial solutions for quadruped robots: digital power and infrastructure inspection. 2026 [cited 2026 Jan 1]. Available from: https://www.deeprobotics.cn/en/index/industry.html#part1.
9. Energy Robotics. Automating industrial inspection rounds with autonomous mobile robots. 2024 [cited 2026 Jan 1]. Available from: https://www.energy-robotics.com/post/automating-industrial-inspection-rounds-with-autonomous-mobile-robots.
10. Dong Z, Gao Y, Yan Y, Chen F. Vector detection network: an application study on robots reading analog meters in the wild. arXiv:2105.14522. 2021.
11. Milana E, Ramírez-Agudelo OH, Estevam Schmiedt J. Autonomous reading of gauges in unstructured environments. Sensors. 2022;22(17):6681. doi:10.3390/s22176681.
12. Reitsma M, Keller J, Blomqvist K, Siegwart R. Under pressure: learning-based analog gauge reading in the wild. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 14–20. doi:10.1109/icra57147.2024.10610793.
13. Leon-Alcazar J, Alnumay Y, Zheng C, Trigui H, Patel S, Ghanem B. Learning to read analog gauges from synthetic data. arXiv:2308.14583. 2023.
14. Zhang N, Yang G, Hu F, Yu H, Fan J, Xu S. A novel adversarial deep learning method for substation defect image generation. Sensors. 2024;24(14):4512. doi:10.3390/s24144512.
15. Xu W, Yi W, Tan Y. Generative AI-driven data augmentation and object-guided vision-language reasoning for PPE compliance analysis in work-at-height. Adv Eng Inform. 2026;71:104364. doi:10.1016/j.aei.2026.104364.
16. Lee J, Kim Y, Ko B, Kim T, Park J, Lee J, et al. Robust analog gauge reading via virtual point-based geometric rectification and P2-YOLO-pose. Comput Model Eng Sci. 2026;147(1):35. doi:10.32604/cmes.2026.080624.

17. Bradski G. The OpenCV library. Dr Dobb's J Softw Tools. 2000;25(11):120–3.
18. Long S, Ruan J, Zhang W, He X, Wu W, Yao C. TextSnake: a flexible representation for detecting text of arbitrary shapes. In: Computer Vision—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 19–35. doi:10.1007/978-3-030-01216-8_2.
19. Smith R. An overview of the tesseract OCR engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2; 2007 Sep 23–26; Curitiba, Brazil. p. 629–33. doi:10.1109/icdar.2007.4376991.
20. JaidedAI. EasyOCR: ready-to-use OCR with 80+ languages supported. GitHub repository. 2020 [cited 2026 Jan 1]. Available from: https://github.com/JaidedAI/EasyOCR.
21. Du Y, Li C, Guo R, Yin X, Liu W, Zhou J, et al. PP-OCR: a practical ultra lightweight OCR system. arXiv:2009.09941. 2020. doi:10.48550/arXiv.2009.09941.
22. Li M, Lv T, Chen J, Cui L, Lu Y, Florencio D, et al. TrOCR: transformer-based optical character recognition with pre-trained models. arXiv:2109.10282. 2021.
23. Duan S, Xue Y, Wang W, Su Z, Liu H, Yang S, et al. GLM-OCR technical report. arXiv:2603.10910. 2026. doi:10.48550/arXiv.2603.10910.
24. Lin F, Liu Y, Xu H, Chen Y, He Z, Zhao M, et al. Do vision-language models measure up? Benchmarking visual measurement reading with MeasureBench. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2026 Jun 3–7; Denver, CO, USA.
25. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. arXiv:2304.00685. 2024.
26. Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, et al. Foundational models defining a new era in vision: a survey and outlook. arXiv:2307.13721. 2023.
27. Adil M, Ahmed M, Aqib M, Gonzalez VA, Lee G, Mei Q. Integration of object detection and small VLMs for construction safety hazard identification. arXiv:2604.05210. 2026.
28. Ueno S, Hayashi Y, Nakatsuka S, Yamada Y, Aizawa H, Kato K. Vision-language in-context learning driven few-shot visual inspection model. In: Proceedings of the 20th International Conference on Computer Vision Theory and Applications (VISAPP); 2025 Feb 26–28; Porto, Portugal. p. 253–60.
29. Cao Q, Chen Y, Lu L, Sun H, Zeng Z, Yang X, et al. Generalized domain prompt learning for accessible scientific vision-language models. Nexus. 2025;2(2):100069. doi:10.1016/j.ynexs.2025.100069.
30. Giedra H, Matuzevičius D, Sledevič T, Shubitidze G, Serackis A. Benchmarking large language model inference on limited-resource edge systems. Electronics. 2026;15(11):2451. doi:10.3390/electronics15112451.
31. Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond. arXiv:2308.12966. 2023.
32. Qwen Team. Qwen3-VL technical report. Model: qwen3-vl:8b via Ollama. arXiv:2511.21631. 2025. doi:10.48550/arXiv.2511.21631.
33. Qwen Team. Qwen3.5-Omni technical report. arXiv:2604.15804. 2026. doi:10.48550/arxiv.2604.15804.
34. Google DeepMind. Gemma 4 model card. Model: gemma4:e4b via Ollama. 2025 [cited 2026 Jan 1]. Available from: https://ai.google.dev/gemma/docs/core/model_card_4.
35. Fan C, Li Z, Ding W, Zhou H, Qian K. Integrating artificial intelligence with SLAM technology for robotic navigation and localization in unknown environments. Appl Comput Eng. 2024;67(1):8–13. doi:10.54254/2755-2721/67/2024ma0056.
36. Ranftl R, Lasinger K, Hafner D, Schindler K, Koltun V. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Trans Pattern Anal Mach Intell. 2022;44(3):1623–37. doi:10.1109/TPAMI.2020.3019967.
37. Lin H, Chen S, Liew J, Chen DY, Li Z, Shi G, et al. Depth anything 3: recovering the visual space from any views. arXiv:2511.10647. 2025. doi:10.48550/arXiv.2511.10647.
38. Ataei ST, Zadeh PM, Ataei S. Vision-based autonomous structural damage detection using data-driven methods. arXiv:2501.16662. 2025.
39. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024.

40. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/cvpr.2016.91.
41. Sousa G, Lima R, Trojahn C. Complex ontology matching with large language model embeddings. arXiv:2502.13619. 2025.
42. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5(1):135–46. doi:10.1162/tacl_a_00051.
43. Lei Y, Li N, Guo L, Li N, Yan T, Lin J. Machinery health prognostics: a systematic review from data acquisition to RUL prediction. Mech Syst Signal Process. 2018;104:799–834. doi:10.1016/j.ymssp.2017.11.016.
44. Si XS, Wang W, Hu CH, Zhou DH. Remaining useful life estimation—a review on the statistical data driven approaches. Eur J Oper Res. 2011;213(1):1–14. doi:10.1016/j.ejor.2010.11.018.
45. Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. Mech Syst Signal Process. 2019;115(1):213–37. doi:10.1016/j.ymssp.2018.05.050.
46. Zhang W, Yang D, Wang H. Data-driven methods for predictive maintenance of industrial equipment: a survey. IEEE Syst J. 2019;13(3):2213–27. doi:10.1109/jsyst.2019.2905565.
47. Serradilla O, Zugasti E, Rodriguez J, Zurutuza U. Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. Appl Intell. 2022;52(10):10934–64. doi:10.1007/s10489-021-03004-y.
48. Wang W. Advanced auto labeling solution with added features. GitHub; 2023 [cited 2026 Jan 1]. Available from: https://github.com/CVHub520/X-AnyLabeling.