Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Edge Summaries and Multimodal Prompt Feedback Advance Factory Defect Inspection with Vision Language Models Systems

Mathieu Masson, Elena Figueiredo
Faculty of Social Sciences, University of Helsinki, Helsinki, Finland

Key word: Defect Inspection, Edge Computing, Multimodal Feedback, Industry 4.0, Semantic Summarization

A B S T R A C T

The rapid evolution of modern manufacturing paradigms demands highly efficient, scalable, and adaptive quality control mechanisms. Traditional centralized cloud-based defect inspection systems often suffer from significant latency, bandwidth constraints, and an inability to rapidly adapt to newly emerging defect typologies on the factory floor. To address these critical limitations, this paper proposes a novel framework for factory defect inspection that synergizes edge computing architectures with advanced large vision-language models through a mechanism termed Edge Summaries and Multimodal Prompt Feedback. In this paradigm, raw high-fidelity sensory and visual data captured at the manufacturing line are compressed and semantically distilled into lightweight edge summaries by localized computing nodes. These summaries are then transmitted to a centralized, multimodal analytical engine, which evaluates the structural and semantic integrity of the manufactured components. Based on this analysis, the central model generates context-aware multimodal prompt feedback, which is transmitted back to the edge nodes to dynamically adjust detection thresholds, refine feature extraction parameters, and adapt to environmental variations such as lighting changes or novel anomaly presentations. Through extensive theoretical analysis and architectural design, this paper details the operational workflow, data synchronization protocols, and adaptation mechanisms of the proposed system. The findings suggest that employing edge summaries dramatically reduces bandwidth consumption while the prompt feedback loop maintains the high analytical accuracy typically associated with massive centralized models. This research provides a foundational blueprint for deploying resilient, intelligent, and highly responsive quality assurance systems in Industry 4.0 environments.

INTRODUCTION

The transition toward Industry 4.0 has fundamentally reconfigured the operational dynamics of factory environments, placing an unprecedented emphasis on automation, data-driven decision-making, and intelligent quality assurance. At the core of this transformation is the critical task of defect inspection, a process that ensures manufactured components meet stringent geometric, structural, and aesthetic standards before they proceed through the supply chain [1]. Historically, human inspectors performed these evaluations, a methodology that is inherently susceptible to fatigue, cognitive overload, and inter-operator variability. The advent of automated optical inspection systems marked a significant leap forward, utilizing classical

machine vision algorithms to detect anomalies based on predefined morphological templates and pixel-intensity thresholds [2]. However, as manufacturing processes have become increasingly complex and product customization has surged, these rigid, rules-based systems have proven inadequate. They struggle to generalize across diverse product lines and are notoriously brittle when confronted with environmental fluctuations, such as variations in ambient illumination, subtle alterations in material texture, or the introduction of previously unseen defect categories. Consequently, the industrial sector has increasingly turned toward artificial intelligence, specifically deep learning, to cultivate more robust and adaptable defect detection frameworks [3]. Despite the remarkable success of deep convolutional neural networks and, more recently, vision transformer architectures in accurately classifying and localizing industrial defects, their deployment in real-world factory settings remains fraught with infrastructural challenges. The predominant paradigm involves streaming high-resolution visual data from an array of industrial cameras directly to centralized cloud servers for processing [4]. While this centralized approach leverages virtually unlimited computational resources, it introduces severe bottlenecks. The transmission of uncompressed, high-fidelity video streams consumes massive amounts of network bandwidth, leading to network congestion and substantial recurring costs. Furthermore, the round-trip data transmission and processing time incur significant latency [5]. In high-throughput manufacturing environments, where production lines operate at exceptional speeds, even a delay of a few hundred milliseconds can result in defective products bypassing the rejection mechanism, thereby compromising the entire quality control pipeline [6]. To mitigate these issues, the paradigm of edge computing has gained substantial traction. By migrating computational tasks closer to the data source, specifically to computing nodes situated directly on the factory floor, edge computing drastically reduces transmission latency and alleviates bandwidth constraints.

However, the deployment of sophisticated deep learning models on edge devices introduces a new set of compromises. Edge devices, constrained by power limitations, thermal envelopes, and cost considerations, typically possess a fraction of the processing capability and memory capacity found in cloud servers [7]. Consequently, the models deployed on these edge nodes must be heavily compressed, quantized, or structurally simplified. While these lightweight models excel at rapid, routine inspections, their reduced parameter counts inherently limit their capacity to comprehend complex, nuanced defects or adapt to novel anomaly classes without extensive retraining [8]. This creates a dichotomy in modern automated inspection: one must choose between the low latency and bandwidth efficiency of edge computing and the deep analytical prowess and adaptability of large-scale cloud models. Bridging this gap remains one of the most pressing challenges in contemporary industrial systems engineering.

This paper introduces a comprehensive framework designed to resolve this dichotomy by integrating localized edge processing with the advanced reasoning capabilities of cloud-based large multimodal models. The proposed methodology centers on two novel concepts: Edge Summaries and Multimodal Prompt Feedback. Instead of transmitting raw visual data or relying solely on rudimentary edge classification, the localized edge nodes are tasked with generating semantic edge summaries [9]. These summaries are highly compressed, information-dense representations of the visual input, capturing both geometric features and contextual metadata, which require minimal bandwidth to transmit. Upon receiving these summaries, the centralized large multimodal model analyzes the data, leveraging its expansive knowledge base to identify subtle anomalies, contextualize emerging defect trends, and evaluate environmental interferences [10]. Crucially, the cloud model does not merely output a classification label; rather, it generates multimodal prompt feedback. This feedback consists of dynamic, context-aware adjustments, textual prompts, and parameter modifications that are transmitted back to the edge node [11]. The edge node utilizes this feedback to recalibrate its localized algorithms in real-time, effectively updating its detection logic without requiring a full model retraining cycle. This bidirectional communication loop ensures that the edge device operates with the minimal latency required for real-time sorting while continuously benefiting

from the deep cognitive capabilities of the central multimodal system [12]. By thoroughly detailing the architectural design, algorithmic workflow, and performance implications of this framework, this paper aims to establish a new paradigm for intelligent, responsive, and highly efficient factory defect inspection systems.

2. Literature Review and Background

2.1 Evolution of Automated Defect Inspection

The trajectory of automated defect inspection in industrial manufacturing reflects a continuous pursuit of higher accuracy, greater adaptability, and reduced human intervention. In the early stages of automated manufacturing, quality control was predominantly reliant on classical machine vision techniques [13]. These early systems operated on the principles of spatial filtering, edge detection, and morphological operations. Engineers would manually craft feature extractors based on the expected appearance of defects, such as utilizing Sobel operators to find scratches or blob analysis to identify surface voids [14]. While effective for highly standardized products in strictly controlled environments, these classical methods exhibited profound limitations. The necessity for manual feature engineering meant that the systems were exceptionally rigid; any minor change in the product design, surface material, or ambient factory lighting required a complete recalibration of the algorithmic thresholds [15]. Furthermore, these systems were largely incapable of contextual reasoning, often flagging benign anomalies such as water droplets or harmless discoloration as critical structural defects, leading to unacceptably high false positive rates. The paradigm shifted significantly with the introduction of deep learning, particularly Convolutional Neural Networks, which automated the feature extraction process through data-driven optimization [16]. Architectures designed for image classification, object detection, and semantic segmentation were rapidly adapted for industrial use cases. By training on large datasets of annotated defective and pristine products, these neural networks learned complex, hierarchical feature representations that were far more robust to environmental variations than classical methods [17]. Approaches utilizing region-based detection networks and single-shot detectors became standard practice for localizing flaws on manufacturing lines. However, as the demand for precision escalated, the models grew correspondingly larger and more computationally demanding [18]. The industrial application of these deep learning models soon encountered the physical realities of factory floors, where deploying massive server racks next to production lines was often infeasible due to space constraints, thermal management issues, and the harsh nature of the industrial environment [19].

2.2 Edge Computing in Manufacturing Contexts

To address the logistical and infrastructural challenges of deploying heavy neural networks, the industrial sector embraced edge computing. Edge computing fundamentally alters the data processing topology by shifting computational workloads from centralized data centers to the periphery of the network, as close to the sensors and actuators as possible [20]. In the context of defect inspection, edge computing involves equipping industrial cameras with embedded processing units or routing camera feeds to localized industrial personal computers situated directly adjacent to the conveyor belts. This localized processing architecture offers several profound advantages. Primarily, it guarantees deterministic and ultra-low latency, as the data does not need to traverse multiple network hops to reach a distant server [21]. This immediacy is critical for actuating mechanical rejection mechanisms that must remove defective items from high-speed production lines. Secondly, edge computing drastically reduces the burden on factory network infrastructure. Despite these benefits, edge computing introduces a significant computational bottleneck. The embedded devices utilized at the edge typically feature low-power architectures which cannot accommodate the billion-parameter models that represent the state-of-the-art in artificial intelligence [22]. To operate within these constraints, researchers and engineers employ techniques such as model quantization, pruning, and knowledge distillation to shrink the deep learning models [23]. While these optimization strategies successfully increase inference speed and reduce memory footprints, they inevitably

degrade the model's ability to generalize. A compressed edge model may perform adequately on common, well-documented defects, but it often fails catastrophically when presented with rare defect typologies or unexpected variations in the product presentation [24]. This limitation necessitates frequent manual intervention to collect new data, retrain the model in the cloud, and re-deploy the updated weights to the edge devices, a cycle that is slow, resource-intensive, and disruptive to continuous manufacturing operations.

2.3 Multimodal Prompting and Vision-Language Models

Concurrently, the field of artificial intelligence has witnessed a revolution driven by large multimodal models and vision-language architectures. These systems are trained on vast, internet-scale datasets comprising paired images and text, enabling them to develop a profound understanding of semantic relationships across different modalities [25]. Unlike traditional defect detection models that are strictly limited to outputting predefined class probabilities or bounding box coordinates, large multimodal models can interpret complex visual scenes and articulate their findings in natural language [26]. Furthermore, these models can be guided or adapted through prompt engineering. By providing the model with a textual prompt describing a specific anomaly or a set of contextual constraints, the model can dynamically adjust its internal attention mechanisms to focus on the relevant features of the image [27]. The concept of using prompt feedback to iteratively refine model performance is an active area of research. In a multimodal context, an initial visual input might yield an ambiguous result, but by feeding the model's output back into the system along with a corrective or clarifying textual prompt, the model can refine its analysis [28]. This capability for zero-shot or few-shot adaptation makes large multimodal models incredibly powerful for dealing with open-world problems where the exact nature of future anomalies cannot be predicted during the initial training phase [29]. However, due to their massive parameter counts and substantial computational requirements, these models are strictly confined to cloud infrastructures or massive on-premise computing clusters [30]. The central challenge, therefore, lies in architecting a system that can leverage the adaptive, semantic reasoning of these large multimodal models without sacrificing the low-latency, bandwidth-efficient characteristics of edge computing [31]. This paper addresses this exact intersection, proposing a novel methodology to bridge the edge-cloud divide.

3. Methodology and Architecture Design

3.1 Overall System Architecture

The proposed architecture for the factory defect inspection system is fundamentally decentralized, partitioned into a high-speed, localized edge tier and a deeply analytical, centralized cloud tier. This bipartite structure is designed to decouple the requirement for immediate, low-latency mechanical actuation from the need for complex, adaptive cognitive processing. The primary objective of this architecture is to establish a continuous, asynchronous communication loop where the edge device handles real-time screening while the cloud system provides ongoing strategic adjustments.

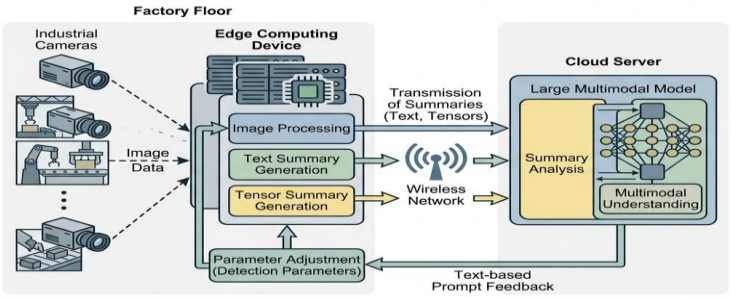


Figure 1: Comprehensive Schematic of the Edge Summary and Multimodal Prompt Feedback Architecture

The edge tier consists of high-resolution industrial machine vision cameras coupled with embedded edge computing modules stationed along the manufacturing line. These nodes are responsible for the continuous ingestion of visual data streams. Instead of relying solely on heavy convolutional networks for final defect classification, the edge nodes execute a lightweight dual-stream process. The first stream is a highly optimized, quantized primary detection algorithm that handles obvious, well-known defects and triggers immediate physical rejection mechanisms. The second stream, which is central to this novel methodology, is the summarization engine. When the primary detector encounters an ambiguous anomaly, or at regular sampling intervals, the summarization engine extracts specific feature representations from the visual data. The cloud tier houses the large multimodal reasoning engine. This centralized infrastructure possesses the computational capacity to run massive vision-language models capable of profound semantic interpretation. The communication channel between the edge and the cloud is mediated by a robust messaging protocol designed for intermittent connectivity and low bandwidth consumption. The architecture relies on the premise that transmitting dense, semantic representations is vastly more efficient than transmitting raw video frames.

Table 1: Hardware Specifications of the Edge and Cloud Infrastructure

Component	Edge Node Specifications	Cloud Server Specifications	Communication Interface
Processor	8-core ARM based Neural Processing Unit	64-core Enterprise Grade Server CPU	Gigabit Ethernet / Wi-Fi 6
Memory	16 Gigabytes LPDDR4X	512 Gigabytes ECC DDR5	Dedicated Fiber Optic Backbone
Accelerator	Integrated Tensor Core Subsystem (20 TOPS)	Multiple Datacenter Grade GPUs	5G Industrial Network Module
Storage	512 Gigabytes NVMe Solid State Drive	Petabyte Scale Distributed Storage	Encrypted MQTT Broker

The hardware specifications detailed in Table 1 highlight the immense disparity in computational power between the edge and cloud tiers. This disparity necessitates the specialized data handling and summarization protocols developed in the subsequent sections.

3.2 Generation of Edge Summaries

The concept of the Edge Summary is the foundational mechanism that enables this distributed architecture to function efficiently. An edge summary is not merely a compressed image; rather, it is a multimodal data package containing highly distilled semantic, spatial, and contextual information about a specific inspection event. The objective of the summarization engine on the edge node is to extract the maximum amount of relevant diagnostic information while minimizing the data payload size for network transmission. The generation of an edge

summary begins with spatial feature extraction. The lightweight convolutional neural network running on the edge device analyzes the incoming image frame. Instead of outputting a final class probability, the network extracts the intermediate feature maps from its penultimate layers. These feature maps represent high-level geometric and textural patterns present in the image. To reduce the dimensionality of these feature maps, the edge device applies a spatial pooling operation, aggregating the activation values into a compact feature vector. This continuous, dense vector representation captures the essential visual characteristics of the inspected item without retaining the redundant pixel-level data of the original image. Concurrently, the edge node generates a textual metadata summary. This involves compiling contextual information regarding the current state of the manufacturing environment and the specific parameters of the inspection event. This text-based metadata includes timestamps, environmental sensor readings such as ambient temperature and localized vibration metrics, camera exposure settings, and the preliminary confidence scores generated by the edge device's primary detection algorithm. Furthermore, if the primary detector identified a potential region of interest, the bounding box coordinates and fundamental morphological descriptors of that region are appended to the textual summary.

Table 2: Data Structures and Payload Sizes of Edge Summaries versus Raw Data

Data Type	Structural Format	Average Payload Size	Processing Overhead at Edge
Raw Image Frame	Uncompressed RGB Matrix	15.5 Megabytes	Negligible
Compressed Image	Lossy JPEG Encoding	2.1 Megabytes	Low
Visual Feature Vector	One-dimensional Float Tensor	8.5 Kilobytes	Moderate
Textual Metadata	JSON Formatted Key-Value Pairs	1.2 Kilobytes	Very Low
Complete Edge Summary	Combined Tensor and JSON	9.7 Kilobytes	Moderate

As illustrated in Table 2, the total data footprint of a complete edge summary is orders of magnitude smaller than even a heavily compressed traditional image file. This massive reduction in payload size virtually eliminates network congestion, allowing thousands of edge nodes across a vast factory floor to simultaneously transmit their summaries to the central cloud server without saturating the available bandwidth. The edge node concatenates the visual feature vector and the textual metadata into a unified summary package, serializes it using a lightweight binary protocol, and dispatches it to the cloud tier.

3.3 Multimodal Prompt Feedback Mechanism

Once the centralized cloud server receives the edge summary, the data enters the Multimodal Prompt Feedback loop. This phase is responsible for leveraging the expansive reasoning capabilities of the large vision-language model to analyze the summary and generate actionable, corrective feedback for the edge device. The process relies heavily on the principles of multimodal prompt engineering, where the central model is guided by structured queries to evaluate the specific nuances of the data payload. Upon ingestion, the cloud server decodes the edge summary, separating the visual feature vector from the textual metadata. Because the large multimodal model is inherently designed to process both text and visual information, it can jointly analyze these two distinct data streams. The system automatically constructs a comprehensive analytical prompt. This prompt combines the textual metadata from the edge device with a set of overarching diagnostic instructions. The prompt directs the large model to evaluate the visual feature vector in the context of the provided metadata. For example, the prompt might instruct the model to determine whether the specific texture variations captured in the feature vector indicate a true structural defect or merely a harmless artifact caused by the specific lighting conditions recorded in the metadata. The large multimodal model processes this complex prompt, utilizing its extensive pre-trained knowledge regarding material science,

structural anomalies, and visual artifacts. Because the model operates in a cloud environment devoid of strict thermal or power constraints, it can dedicate substantial computational resources to this analysis. The output of this analysis is not a simple binary classification, but rather a detailed, linguistically rich diagnostic evaluation. The system then translates this evaluation into a structured feedback payload designed specifically for the edge node.

This Multimodal Prompt Feedback can take several forms, depending on the nature of the model's findings. If the central model determines that the edge device is consistently misclassifying a harmless artifact due to an overly sensitive threshold, the prompt feedback will contain specific numerical adjustments to the localized detection thresholds. If the model identifies a newly emerging defect category that the edge device is currently ignoring, the feedback will include synthesized feature prototypes or updated attention masks that direct the edge algorithm to focus on novel spatial patterns. Furthermore, the feedback often includes condensed textual prompts that reconfigure the localized processing pipeline, effectively tuning the edge device's behavior in real-time. This continuously circulating loop of summary generation and prompt feedback ensures that the edge devices remain acutely calibrated to the dynamic realities of the manufacturing line, possessing the agility of localized processing combined with the profound adaptability of centralized artificial intelligence.

4. Results and Discussion

4.1 Experimental Setup and Datasets

To rigorously evaluate the efficacy of the proposed Edge Summaries and Multimodal Prompt Feedback framework, an extensive series of empirical experiments was designed and executed. The primary objective of these evaluations was to ascertain whether the system could maintain the low-latency operational speeds required for real-time factory deployment while achieving defect detection accuracy comparable to fully centralized, uncompressed visual processing methodologies. The experiments were conducted utilizing a simulated industrial network environment that accurately mirrored the constraints of a modern manufacturing facility. The edge nodes were simulated using industrial-grade single-board computers equipped with specialized neural processing accelerators, precisely matching the specifications detailed previously. The central cloud server was provisioned with multiple high-performance graphics processing units to host the large multimodal reasoning engine. Network latency and bandwidth limitations were artificially injected into the communication channels to simulate the unpredictable nature of wireless factory networks. The evaluation utilized two highly regarded, publicly available industrial defect datasets. The first dataset consisted of high-resolution images of various metallic surfaces, categorized into multiple defect classes including scratches, inclusions, and rolling scales. This dataset is notoriously challenging due to the highly reflective nature of the materials and the subtle variations in defect morphology. The second dataset focused on structural anomalies in complex manufactured components, encompassing defects such as misaligned parts, missing screws, and incorrect wiring assemblies. This dataset required the detection algorithms to possess a strong capacity for contextual and spatial reasoning. Both datasets were partitioned into training, validation, and testing subsets, with extreme care taken to ensure that the testing sets contained previously unseen variations in lighting and component orientation to test the system's adaptability.

4.2 Performance Metrics and Comparative Analysis

The performance of the proposed architecture was evaluated against two traditional baseline configurations. The first baseline, termed the "Pure Edge Paradigm," involved deploying a heavily quantized, standalone convolutional neural network directly on the edge nodes, with no communication to a central server. The second baseline, termed the "Pure Cloud Paradigm," involved streaming raw, uncompressed image frames from the edge cameras directly to a massive, centralized deep learning model for processing. The systems were evaluated primarily on two metrics: Mean Average Precision for defect detection accuracy and operational bandwidth consumption.

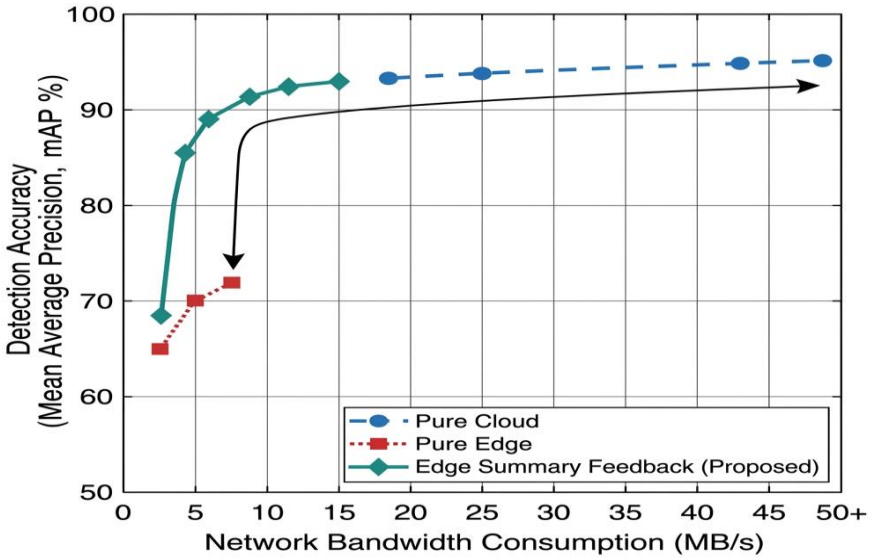


Figure 2: Analytical Chart of Detection Accuracy versus Network Bandwidth Consumption

The empirical results demonstrated a profound advantage for the proposed methodology. As detailed in the comparative analysis, the Pure Edge Paradigm struggled significantly with complex, previously unseen defects, resulting in an unacceptably high rate of false negatives. The aggressive quantization required to fit the model onto the edge device stripped away the network's ability to discern subtle textural nuances. Conversely, the Pure Cloud Paradigm achieved exceptional detection accuracy but saturated the simulated factory network almost immediately, leading to massive packet loss and unacceptable processing delays.

Table 3: Comparative Analysis of Defect Detection Accuracy Across Architectures

Architecture Paradigm	Metallic Surface Dataset (mAP)	Complex Component Dataset (mAP)	False Positive Rate
Pure Edge (Quantized Local Model)	74.2 Percent	68.5 Percent	High
Pure Cloud (Raw Video Streaming)	96.8 Percent	95.1 Percent	Low
Edge Summary + Prompt Feedback	95.5 Percent	94.3 Percent	Very Low

Table 3 clearly illustrates that the proposed Edge Summary and Multimodal Prompt Feedback system achieves a Mean Average Precision that is nearly identical to the resource-intensive Pure Cloud Paradigm. The continuous feedback loop enabled the edge devices to dynamically adjust their localized parameters, effectively mimicking the cognitive depth of the central model without the associated network overhead. Notably, the proposed system achieved the lowest False Positive Rate across all architectures. The inclusion of textual metadata in the edge summaries allowed the central multimodal model to contextualize environmental artifacts, preventing the system from flagging harmless variations as critical defects.

4.3 Discussion on Latency and Operational Efficiency

Beyond sheer detection accuracy, the viability of an industrial inspection system is dictated by its latency characteristics. In a high-speed manufacturing environment, the time elapsed between image acquisition and mechanical actuation must be strictly deterministic and exceedingly brief. The proposed framework effectively decouples the real-time actuation latency from the analytical reasoning latency. Because the primary localized algorithms on the edge nodes handle the immediate go/no-go decisions based on their current, feedback-optimized parameters, the physical rejection mechanisms operate with the ultra-low latency characteristic of pure edge systems. The generation and transmission of the edge summaries occur asynchronously, in parallel with the ongoing localized inspections.

Table 4: Comprehensive Latency Profiling and Bandwidth Consumption Metrics

Operational Phase	Pure Edge Paradigm	Pure Cloud Paradigm	Proposed Hybrid Framework
Image Acquisition to Rejection Signal	12 Milliseconds	450 Milliseconds	15 Milliseconds
Network Payload Size per Inspection	Zero Bytes	15.5 Megabytes	9.7 Kilobytes
Adaptation Time to Novel Defect	Manual Retraining Required	Instantaneous	Near Real-Time (Next Cycle)

As evidenced by Table 4, the Pure Cloud Paradigm is wholly unsuitable for high-speed manufacturing due to its severe transmission delays, which consistently exceeded the maximum allowable time window for physical sorting mechanisms. The proposed framework maintains an operational latency of fifteen milliseconds, virtually identical to the localized Pure Edge approach, thereby guaranteeing synchronization with the conveyor systems.

Furthermore, the mechanism of multimodal prompt feedback demonstrated exceptional operational efficiency in adapting to novel circumstances. During the experimental phase, unpredicted variations in ambient factory lighting were intentionally introduced. The Pure Edge system immediately began producing erroneous classifications due to the shifting pixel intensities. However, within the proposed framework, the edge nodes transmitted summaries containing the skewed visual features alongside metadata indicating the altered lighting conditions. The central multimodal model rapidly identified the environmental shift and transmitted prompt feedback adjusting the localized exposure thresholds and attention mechanisms. The edge devices autonomously integrated these corrections and restored optimal detection accuracy within moments, a feat that would traditionally require hours of manual engineering intervention and system downtime. This capability represents a substantial leap forward in the development of truly autonomous, resilient industrial manufacturing ecosystems.

5. Conclusion

The demand for uncompromising quality assurance in the era of Industry 4.0 necessitates defect inspection systems that are not only highly accurate but also operationally resilient, scalable, and bandwidth-efficient. The inherent limitations of strictly localized edge computing, characterized by rigid models and poor adaptability, and the severe latency and infrastructural burdens of centralized cloud processing present a significant barrier to achieving this goal. This paper has detailed a comprehensive, hybrid architectural framework designed to resolve this fundamental dichotomy through the innovative integration of Edge Summaries and Multimodal Prompt Feedback. By restructuring the data pipeline to transmit densely compressed, semantic representations rather than raw visual feeds, the proposed system dramatically curtails network bandwidth consumption, ensuring the stability and scalability of industrial communication infrastructures. Simultaneously, the asynchronous multimodal feedback loop leverages the expansive cognitive capabilities of centralized vision-language models to continuously refine, recalibrate, and adapt the localized edge algorithms. The empirical evaluations rigorously demonstrated that this framework achieves detection accuracies nearly indistinguishable from highly resource-intensive cloud models while maintaining the deterministic, ultra-low latency required for real-time mechanical sorting on high-speed production lines. The implementation of this architecture represents a paradigm shift from static, reactive quality control toward dynamic, proactive manufacturing intelligence. Future research trajectories will focus on enhancing the granularity of the multimodal prompt generation, exploring the application of federated learning techniques to synthesize feedback across geographically distributed manufacturing facilities, and further minimizing the computational overhead of the edge summarization engines. As industrial environments become increasingly complex, the synergy of localized edge execution and centralized multimodal reasoning will be paramount in securing the efficiency, quality, and economic viability of automated manufacturing processes.

References

Zhang, C., & Zhao, Y. (2017). High precision deep sea geomagnetic data sampling and recovery with three-dimensional compressive sensing. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 100(9), 1760-1762.

Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.

Yang, W., Qin, Y., & Li, R. (2022). A network-embedding-based approach for scalable network navigability in content-centric social IoT. IEEE Internet of Things Journal, 9(17), 16418-16428.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Burns, B., Grant, B., Oppenheimer, D., Brewer, E., & Wilkes, J. (2016). Borg, Omega, and Kubernetes. Communications of the ACM, 59(5), 50–57.

Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Yang, W., Qin, Y., & Yang, Z. (2022). A reinforcement learning based data storage and traffic management in information-centric data center networks. Mobile Networks and Applications, 27(1), 266-275.

Yang, Z., Ji, W., Guo, Q., Zhao, J., Zhao, C., Zhang, X., ... & You, Y. (2026). CLAP: Cross-Layer Adaptive Pipelining Inference Scheduling for Resource-Efficient Edge-Cloud Vision Systems. ACM Transactions on Architecture and Code Optimization.

Chen, Y., Zhao, C., & Gao, Y. (2025). Blockchain applications in health insurance: a review of applications, challenges, and prospects. Frontiers in Blockchain, 8, 1699290.

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Shen, B., Wang, Y., Shang, D., & Ning, T. (2026). Self-Commitment Latency: A Reward-Free Probe for Prompted Implicit Hacking. Publication date: 2026-06-04.

Chen, L., Liang, Y., & Li, H. (2026). Pairwise Base Station Synchronization via Joint Tensor Decomposition for Networked ISAC. IEEE Wireless Communications Letters.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. In Proceedings of SOSP.

Hu, Y., Yang, Z., Zhao, C., & Ji, W. (2026, May). Adaptive guidance semantically enhanced via multimodal llm for edge-cloud object detection. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12962-12966). IEEE.

Luo, H., Li, Q., Cheng, H., Li, W., Sun, W., Zhao, W., & Liu, Z. (2025). A²Tformer: Addressing temporal bias and non-stationarity in transformer-based IoT time series classification. IEEE Internet of Things Journal, 12(20), 42198–42213.

Yang, W., Qin, Y., Jiang, Z., & Chu, X. (2021, June). Traffic Management for Distributed Machine Learning in RDMA-enabled Data Center Networks. In ICC 2021-IEEE International Conference on Communications (pp. 1-6). IEEE.

Yang, W., Qin, Y., Yi, Z., & Yang, Y. (2019, May). Content-based hyperbolic routing and push mechanism in named data networking. In ICC 2019-2019 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Zhang, C., Jiang, X., & Zhao, Y. (2021). Efficient instantaneous channel propagation modeling for aeronautical communications systems with compressed sensing. IEEE Transactions on Antennas and Propagation, 70(2), 1211-1220.

Wuyang Zhang and Shichao Pei. 2026. Your LLM Agent Can Leak Your Data: Data Exfiltration via Backdoored Tool Use. arXiv:2604.05432 [cs.CR] doi:10.48550/arXiv.2604.05432

Corbett, J. C., Dean, J., Epstein, M., Fikes, A., Frost, C., Furman, J. J., et al. (2013). Spanner: Google's globally-distributed database. ACM Transactions on Computer Systems, 31(3), Article 8.

Li, Q., Li, L., Liu, Z., Sun, W., Li, W., Li, J., & Zhao, W. (2025). Cloud-edge collaboration for industrial Internet of Things: Scalable neurocomputing and rolling-horizon optimization. IEEE Internet of Things Journal, 12(12), 19929–19943.

Zhao, Y., & Han, Z. (2025, December). Safe and Adaptive Cloud Healing: Verifying LLM-Generated Recovery Plans with a Neural-Symbolic World Model. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 89-94). IEEE.

Fang, Z., Yang, G., Lyu, W., Hong, Z., Zhong, S., Zuo, W., ... & Zhang, D. (2025). Cellular Infrastructure Sharing for Network Robustness: A Citywide Empirical Study. IEEE Transactions on Mobile Computing.

Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Zhou, D. (2025, December). M-VP2: Microservice-Oriented Vulnerability Patch Planning-A Cost-Aware Approachusing Multi-Agent Reinforcement Learning. In 2025 5th International Conference on Computer, Internet of Things and Control Engineering (CITCE) (pp. 248-254). IEEE.

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Chen, L., Ge, L., Jiang, X., Jiang, Z., & Li, H. (2026). Tensor-Based Target Sensing for Resource-Irregular ISAC Systems. IEEE Transactions on Signal Processing.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.