

Multi-Modal Vision-Language Models for Smart Manufacturing: A Conceptual Synthesis of Perception, Reasoning, and Human-Robot Collaboration in Cognitive Production Systems

Afex Callagher

Ladoke Akintola University of Technology

aoadeyemo49@student.lautech.edu.ng

Abstract

The convergence of computer vision and natural language processing into unified vision-language models (VLMs) has begun to reshape the technological foundations of smart manufacturing. Where earlier generations of industrial artificial intelligence relied on narrow, task-specific classifiers, contemporary VLMs couple visual perception with linguistic reasoning, enabling machines to interpret shop-floor scenes, generate human-readable explanations, and adapt to previously unseen conditions without extensive retraining. This paper undertakes a conceptual and literature-based synthesis of the emerging role of multi-modal VLMs across three interlocking domains of smart manufacturing: human-robot collaboration, quality inspection and anomaly detection, and multimodal perception for decision support. Rather than reporting new empirical measurements, the study integrates and critically compares findings from recent peer-reviewed and preprint literature to identify recurring architectural patterns, persistent technical bottlenecks, and unresolved theoretical questions. The review draws on a structured narrative synthesis method, organizing evidence according to a proposed conceptual framework that positions VLMs as mediating agents between physical production processes and cognitive decision layers. The findings indicate that VLMs consistently improve interpretability, generalization to novel defect types, and natural language interaction with human operators, yet remain constrained by inference latency, dependence on high-fidelity digital twins, and difficulty resolving fine-grained geometric ambiguity. The paper concludes by proposing a research agenda oriented toward hybrid architectures that pair lightweight detectors with language-grounded reasoning modules, and

toward evaluation standards that better reflect real-world deployment constraints in manufacturing environments.

DOI:10.21590/ijhit.06.04.27

1. Introduction

Manufacturing has long served as a proving ground for successive waves of automation, each redefining the boundary between human and machine labor. The transition from programmable logic controllers to networked cyber-physical systems under the banner of Industry 4.0 established the technical infrastructure for pervasive sensing, real-time data exchange, and adaptive control (Oks, Jalowski, Lechner, et al., 2024). Yet despite these advances, a persistent limitation has characterized industrial artificial intelligence: perception systems built on convolutional neural networks and closed-set classifiers excel at narrowly defined tasks but struggle to generalize when confronted with novel defect types, unfamiliar spatial layouts, or ambiguous operator instructions. Traditional object detectors such as Faster R-CNN and the YOLO family (Redmon, Divvala, Girshick, & Farhadi, 2016; Ren, He, Girshick, & Sun, 2017) remain computationally efficient and accurate within their trained distribution, but they lack any inherent capacity for semantic reasoning, contextual explanation, or open vocabulary adaptation.

The recent emergence of large vision-language models, exemplified by contrastive pretraining approaches such as CLIP (Radford et al., 2021) and instruction-tuned multimodal architectures such as LLaVA and its successors, has introduced a qualitatively different paradigm. Rather than mapping pixels directly to a fixed label set, VLMs learn a shared embedding space between visual and textual modalities, allowing them to describe, question, and reason about what they perceive using natural language. This capability has begun to migrate from general-purpose benchmarks into specialized industrial contexts, where researchers have applied VLMs to human-robot collaborative assembly (Liu et al., 2026), cobot navigation in unstructured environments (Wang, Fan, & Zheng, 2024; Wang, Fan, Zheng, Yan, & Wang, 2025), and multimodal perception for production decision-making (Wang, Zhang, Jiang, & Li, 2025). Parallel developments in industrial anomaly detection have demonstrated that multimodal large language models can jointly localize

defects and articulate structured, context-aware inspection reports, a capability unavailable to conventional detectors (Nguyen, Park, Yoo, & Yang, 2026; Zhang, Ruan, Gao, Liu, & Fu, 2025).

Despite this proliferation of application-specific studies, the literature remains fragmented across disciplinary silos: robotics researchers emphasize navigation and manipulation, quality engineers emphasize defect localization, and computer scientists emphasize benchmark performance in isolation from deployment constraints. What is comparatively scarce is a synthesis that positions these strands within a coherent conceptual account of how vision-language models function as a general purpose cognitive layer within smart manufacturing systems. This paper addresses that gap. It asks three interrelated questions. First, what architectural and methodological patterns recur across VLM applications in manufacturing, and how do they differ from earlier deep learning paradigms. Second, what capabilities do VLMs consistently confer upon industrial systems, and under what conditions do those capabilities degrade. Third, what theoretical and practical gaps must be addressed before VLMs can be considered a mature component of industrial artificial intelligence infrastructure rather than an experimental augmentation.

The paper is organized as a conceptual and literature-based inquiry rather than an empirical study reporting newly collected data. This choice reflects both the nascent and rapidly evolving state of the underlying technology and the methodological value of synthesis at a stage where individual studies have proliferated faster than integrative frameworks. The analysis draws on a foundational understanding of multimodal representation learning, in which contrastive and generative pretraining objectives align visual and linguistic embeddings, and on socio-technical perspectives that treat manufacturing systems as networks of human operators, physical assets, and digital cognition layers operating under Industry 4.0 and emerging Industry 5.0 paradigms. By situating recent empirical findings within this dual lens, the paper aims to extend existing knowledge not by adding another isolated case study but by articulating the conceptual scaffolding within which such studies can be interpreted comparatively. The remainder of the paper proceeds as follows. Section two synthesizes the relevant literature across foundational multimodal architectures, human-robot collaboration, quality inspection, and multimodal decision support, while identifying persistent gaps. Section three describes the narrative synthesis methodology and analytical framework employed. Section four presents the comparative findings distilled from the reviewed literature.

Section five discusses the implications, limitations, and future research directions, and section six concludes with a summary of scholarly contribution.

2. Literature Review

2.1 Foundations of Vision-Language Representation Learning

The theoretical basis of contemporary VLMs lies in contrastive language-image pretraining, most influentially articulated by Radford et al. (2021) in the CLIP architecture. Rather than training an image encoder against a fixed taxonomy of labels, CLIP jointly trains an image encoder and a text encoder to predict the correct pairing between images and their natural language descriptions across a very large corpus of noisy internet-scale data. The resulting shared embedding space allows the model to perform zero-shot classification by comparing an image embedding against the embeddings of arbitrary textual prompts, a property that proved unexpectedly robust across distribution shifts and downstream tasks. This architectural insight, that natural language can serve as a flexible and scalable form of supervision, underlies nearly every subsequent generation of vision-language systems, including the encoder-decoder and instruction-tuned models later adapted for manufacturing contexts such as the Qwen-VL, Gemma, and LLaVA families referenced throughout the industrial inspection literature (Nguyen et al., 2026).

Building on contrastive foundations, subsequent architectures introduced generative capabilities that allow models not merely to classify or retrieve but to produce free-form textual output conditioned on visual input. This shift from discriminative to generative multimodal reasoning is what enables a VLM deployed on a factory floor to move beyond binary defect classification toward producing a structured inspection narrative describing the location, severity, and probable cause of an anomaly. It is precisely this generative, explanatory capacity that recurs as the distinguishing value proposition across the manufacturing-specific literature reviewed below, in contrast to the purely discriminative outputs of earlier convolutional architectures.

2.2 From Industry 4.0 Infrastructure to Cognitive Manufacturing

Industry 4.0 established the sensing, connectivity, and data infrastructure necessary for advanced analytics on the factory floor, with cyber-physical systems serving as the conceptual backbone linking physical production assets to digital representations (Oks, Jalowski, Lechner, et al., 2024).

Complementary quality management paradigms, including Six Sigma and Lean methodologies, have increasingly been reframed within a Quality 4.0 lens that integrates machine learning into zero-defect manufacturing initiatives (Gonzalez Santacruz, Romero, Noguez, & Wuest, 2024). These frameworks provide the operational context into which VLMs are now being inserted, but they were largely conceived around statistical process control and supervised learning rather than open-ended multimodal reasoning. The literature increasingly frames the current moment as a transition from data-driven automation toward cognitive manufacturing, in which production systems are expected not only to sense and act but to interpret, explain, and communicate. Fan, Yin, Wang, Dong, Zheng, and Wang (2025), in a systematic review spanning over a hundred studies published between 2020 and 2024, describe this transition explicitly in the context of human-robot collaboration, arguing that large language models and vision-language models have motivated a wave of preliminary exploration that nonetheless remains fragmented across individual technical components rather than integrated into a comprehensive collaborative architecture.

This fragmentation is itself an important finding for the present synthesis. Fan et al. (2025) note that existing scholarship has tended to examine robotic task planning, navigation, and manipulation as separate problems, each augmented independently by VLM capabilities, without a unifying account of how these components interact within a shared human-centric production environment. Their review catalogs the pretraining architectures underlying contemporary VLMs and traces their application into task planning, semantic navigation, and skill transfer through multimodal data integration, providing one of the more comprehensive mappings of the field to date. However, even this systematic account stops short of articulating a general theoretical model of how VLMs function as a cognitive layer that mediates between raw sensory input and production-level decision-making, a gap the present paper seeks to address through conceptual synthesis rather than additional empirical cataloging.

2.3 Vision-Language Models in Human-Robot Collaborative Assembly

A substantial and growing body of work applies VLMs specifically to human-robot collaboration in assembly contexts, where the central technical challenge is translating a human operator's high-level, often ambiguous natural language instruction into a sequence of precise, physically grounded robot actions. Liu, Qian, Tang, Zhu, Pang, and Cai (2026) propose an embodied multi-

agent system that integrates a vision-language model within a digital twin-assisted collaborative assembly pipeline, framing the problem explicitly as one of moving from insight to action. Their architecture distributes cognitive responsibilities across multiple agents, with the VLM component responsible for grounding linguistic instructions in the visual state of the assembly cell as represented within a synchronized digital twin. This approach illustrates a recurring architectural pattern in the literature, in which the VLM does not operate as a standalone controller but as one module within a broader multi-agent or multi-stage pipeline that also incorporates simulation, planning, and low-level motor control.

The dependence on digital twin fidelity introduced by this architecture raises a theoretical tension worth foregrounding. Digital twins are themselves imperfect representations of physical reality, subject to sensor noise, calibration drift, and latency between the physical and virtual states. When a VLM reasons over a digital twin representation rather than raw sensory input, any systematic bias or degradation in the twin's fidelity propagates directly into the model's action recommendations. Liu et al. (2026) do not fully resolve this dependency, though their reported improvements in task completion suggest that current levels of digital twin fidelity are sufficient for many assembly subtasks even if not universally robust. This finding is consistent with the broader pattern identified by Fan et al. (2025), who note that VLM-based human-robot collaboration research has advanced fastest in structured or semi-structured environments where the gap between simulated and physical conditions can be more readily controlled.

Complementing the assembly-focused literature, Wang, Fan, and Zheng (2024) introduce a vision and language cobot navigation approach specifically designed for human-centric smart manufacturing, in which the collaborative robot must move through a shared workspace while interpreting natural language guidance from a human co-worker. Extending this line of inquiry, Wang, Fan, Zheng, Yan, and Wang (2025) present a more elaborate architecture for human-guided mobile robot navigation in unstructured environments, combining three-dimensional scene reconstruction through point cloud techniques with zero-shot semantic segmentation performed by a vision-language model. The zero-shot property is particularly significant for manufacturing environments, which are rarely as static or standardized as the benchmark datasets on which many computer vision systems are originally trained; factory layouts change, new equipment is introduced, and lighting conditions vary across shifts. A navigation system that requires retraining

each time the physical environment changes imposes substantial ongoing engineering cost, whereas a VLM-based system that generalizes from pretraining can, in principle, adapt to layout changes with minimal additional supervision.

Taken together, this cluster of studies demonstrates that VLMs are being positioned less as replacements for classical robotic planning and control stacks and more as an interpretive layer that translates between the unstructured, linguistic, and visually ambiguous inputs characteristic of human communication and the structured, quantitative representations required by downstream planning algorithms. This positioning has important implications for how the technology should be evaluated, a point returned to in the methodology and discussion sections below.

2.4 Vision-Language Models in Quality Inspection and Anomaly Detection

Quality inspection constitutes the most mature and empirically dense application area for VLMs in manufacturing, reflecting both the economic significance of defect detection and the relative tractability of framing inspection as an image understanding task. Historically, automated visual inspection has relied on convolutional architectures such as Faster R-CNN (Ren, He, Girshick, & Sun, 2017) and the YOLO family of real-time detectors (Redmon, Divvala, Girshick, & Farhadi, 2016), which Khanam, Hussain, Hill, and Allen (2024) characterize in a comprehensive review as the dominant paradigm for industrial defect detection owing to their favorable balance of accuracy and inference speed. These architectures, however, share a structural limitation: they are trained under a closed-set, supervised regime that requires large volumes of labeled defect imagery and struggles to generalize when new defect categories emerge or when process conditions shift, a brittleness Sundaram and Zeid (2023) identify as a central barrier to scalable smart quality inspection.

Recent research has sought to address this brittleness by adapting large vision-language models to the inspection task. Nguyen, Park, Yoo, and Yang (2026) provide what is arguably the most methodologically rigorous demonstration to date, fine-tuning seven state-of-the-art large vision-language models drawn from the Qwen, Gemma, and LLaVA families for unified defect localization and structured metadata prediction within liquefied natural gas tank manufacturing, a safety-critical domain characterized by heterogeneous fabrication processes including welding, coating, and cable installation. Their study is notable for directly benchmarking VLM-based

inspection against specialized detectors, finding that the best-performing model, an eight billion parameter variant from the Qwen3 family, achieved localization accuracy statistically comparable to the strongest conventional detector evaluated while additionally producing structured, machine-readable inspection narratives that conventional detectors cannot generate without supplementary post-processing pipelines. This finding is theoretically significant because it demonstrates that the interpretability advantage of VLMs need not come at the cost of core detection accuracy, challenging an implicit assumption in earlier literature that generative, language-capable models would necessarily underperform specialized discriminative architectures on narrow localization tasks.

The same study, however, surfaces a persistent limitation that recurs across the broader anomaly detection literature: degraded performance on thin, elongated, or low-contrast defect geometries such as cracks and scratches, which the authors attribute to the loss of fine spatial detail during patch-based visual tokenization at reduced input resolutions. This limitation is echoed in complementary work on explainable industrial anomaly detection. Zhang, Ruan, Gao, Liu, and Fu (2025) introduce a dedicated multimodal defect localization module designed to decouple dialog generation from core feature extraction, motivated explicitly by the observation that multimodal large language models applied naively to anomaly detection tend to overfit during fine-tuning, sacrificing either question-answering fluency or mask-based grounding precision. Their proposed decoupling strategy, together with a newly constructed defect detection question answering training corpus, represents one of several architectural responses to the tension between generative flexibility and spatial precision that pervades this literature.

A related but distinct line of work targets zero-shot and open-vocabulary anomaly detection, in which the model must identify anomalies in object categories never encountered during training. Deng, Luo, Zhai, Cao, and Kang (2024) propose a visual-enhanced multimodal large language model explicitly designed for this zero-shot setting, arguing that fixed-prompt approaches to anomaly detection remain constrained by closed-world assumptions that limit their applicability to the continuously evolving defect taxonomies characteristic of flexible manufacturing. Gu, Zhu, Zhu, Chen, Tang, and Wang (2024) offer an early and influential contribution in this space, demonstrating that a multimodal large language model can be guided toward substantially improved industrial anomaly detection performance relative to naive prompting, establishing a

template that subsequent studies, including those reviewed above, have iterated upon. Collectively, this cluster of studies traces a coherent trajectory: from demonstrating that VLMs can plausibly perform anomaly detection at all, toward increasingly sophisticated architectural interventions that address the specific failure modes, overfitting during fine-tuning, loss of spatial precision, and closed-world prompting, that separate promising demonstrations from deployment-ready systems.

2.5 Multimodal Perception and Decision-Making Frameworks

A third strand of literature addresses the integration of vision-language models into broader perception and decision-making pipelines that extend beyond either assembly or inspection alone. Wang, Zhang, Jiang, and Li (2025) propose a multimodal large language model framework explicitly designed for intelligent perception and decision-making across smart manufacturing operations, positioning the VLM not as a component dedicated to a single task but as a general reasoning layer capable of fusing heterogeneous sensor and image data into coherent situational assessments that inform operational decisions. This framing is conceptually important because it gestures toward the possibility that VLMs may eventually serve a role analogous to a cognitive supervisory layer within manufacturing execution systems, synthesizing information that would otherwise require separate, purpose-built analytical modules.

This more ambitious framing, however, remains comparatively underdeveloped relative to the assembly and inspection literatures discussed above, both in terms of the number of published studies and the maturity of evaluation methodology. Where inspection studies benefit from well-established metrics such as mean average precision and structured field-level accuracy, and navigation studies can draw on established robotics benchmarks, the evaluation of general-purpose multimodal decision support remains comparatively ad hoc, often relying on case-specific demonstrations rather than standardized benchmarks. This asymmetry constitutes one of the clearer gaps identified through this synthesis and is addressed further in the discussion section.

2.6 Synthesis of Debates, Trends, and Research Gaps

Reading across these four clusters of literature reveals several cross-cutting debates that have not yet been fully resolved. The first concerns the appropriate division of labor between specialized discriminative models and general-purpose generative VLMs. Several studies, most explicitly

Nguyen et al. (2026), find that VLMs can approach but not consistently exceed the raw localization accuracy of specialized detectors, while offering substantial gains in interpretability and multitask flexibility. This suggests that the field is converging, whether explicitly or implicitly, toward hybrid architectures in which lightweight, fast detectors handle high-throughput localization while VLMs are reserved for semantic interpretation, reporting, and edge cases requiring contextual reasoning, a division of labor explicitly proposed as future work by Nguyen et al. (2026) and implicit in the multi-agent architecture of Liu et al. (2026).

A second debate concerns generalization versus specialization. The zero-shot and few-shot capabilities that make VLMs attractive for dynamic manufacturing environments, where defect types and layouts evolve over time, are precisely the capabilities most degraded when models are aggressively fine-tuned on narrow domain-specific corpora, as documented by Zhang et al. (2025) in their discussion of overfitting during fine-tuning. This creates a practical tension for manufacturers: sufficient domain adaptation is necessary to achieve acceptable accuracy on specialized industrial imagery, yet excessive adaptation risks eroding the very generalization advantage that motivated the adoption of VLMs in the first place. Few of the reviewed studies systematically characterize this tradeoff curve, representing a clear opportunity for future methodological contribution.

A third and perhaps more fundamental gap concerns theoretical integration. Individually, the reviewed studies each demonstrate that VLMs can be productively applied to a specific manufacturing subproblem, whether navigation, assembly, or inspection. What remains largely absent from the literature is a unifying conceptual account of why these disparate successes share common underlying mechanisms, namely the capacity of contrastively and generatively pretrained multimodal representations to serve as a flexible interface between unstructured human communication and structured industrial control and reporting systems. The present paper's contribution lies precisely in articulating this unifying account, developed in the methodology and results sections that follow, and in using it to organize a comparative assessment of strengths, limitations, and future priorities across the fragmented literature described above.

3. Methodology

This study adopts a qualitative, theory-building research design grounded in structured narrative literature synthesis, rather than an empirical study involving primary data collection or laboratory experimentation. This methodological choice is deliberate and warrants justification. The application of vision-language models to smart manufacturing is a rapidly evolving field in which individual empirical studies, several published within the past twelve months, already provide detailed quantitative benchmarking of specific architectures against specific datasets. What the field currently lacks is not additional point estimates of model accuracy but an integrative conceptual framework capable of situating these individual findings within a coherent account of how VLMs function across the manufacturing value chain. A narrative synthesis design is therefore methodologically appropriate because it is oriented toward theory development and conceptual integration rather than toward hypothesis testing on a novel dataset, a distinction well established in qualitative and interpretive research traditions concerned with synthesizing heterogeneous evidence into higher order explanatory frameworks.

The analytical procedure proceeded in four stages. The first stage involved a structured search for peer-reviewed articles, conference proceedings, and reputable preprint repositories addressing the intersection of vision-language models, multimodal large language models, and smart manufacturing, encompassing subdomains including human-robot collaboration, cobot navigation, quality inspection, anomaly detection, and multimodal decision support. Search terms combined manufacturing-specific vocabulary, including terms such as smart manufacturing, human-robot collaboration, defect detection, and quality inspection, with vision-language model vocabulary, including terms such as vision-language model, multimodal large language model, and large vision-language model. The second stage applied inclusion criteria requiring that sources either directly apply or critically evaluate vision-language or multimodal large language model architectures within an industrial or manufacturing context, or provide foundational theoretical grounding for the multimodal representation learning techniques underlying such applications. Sources addressing vision-language models exclusively within non-industrial domains, such as general-purpose visual question answering benchmarks unconnected to manufacturing applications, were excluded unless they represented foundational architectural contributions, such

as the original CLIP formulation, that are directly inherited by the manufacturing-specific literature.

The third stage involved thematic coding of the retained literature according to the conceptual framework developed for this study, which positions vision-language models as an interpretive layer mediating three interfaces within a manufacturing system: the interface between human natural language instruction and machine action, corresponding to the human-robot collaboration and cobot navigation literature; the interface between raw visual sensory data and structured quality assessment, corresponding to the inspection and anomaly detection literature; and the interface between heterogeneous multimodal sensor streams and higher order operational decisions, corresponding to the perception and decision support literature. Each retained source was coded according to which of these three interfaces it primarily addressed, which underlying VLM architecture or family it employed or discussed, what specific capability gain was reported relative to prior non-multimodal baselines, and what specific limitation or failure mode was identified by the study's own authors. The fourth stage synthesized these codes across sources to identify recurring patterns, points of convergence, and unresolved tensions, the results of which are presented in section four.

This framework-driven coding procedure functions analogously to a qualitative content analysis, in which the unit of analysis is the individual study's reported capability or limitation rather than a numerical measurement, and in which the analytical objective is pattern identification and theoretical synthesis rather than statistical inference. Reliability in this context was pursued through explicit, auditable coding criteria tied directly to the three-interface framework described above, and through triangulation across multiple independently authored studies addressing overlapping subproblems, for example the multiple independent treatments of cobot navigation by Wang, Fan, and Zheng (2024) and Wang, Fan, Zheng, Yan, and Wang (2025), which allowed convergent findings to be distinguished from findings idiosyncratic to a single study or dataset.

Several validity considerations merit acknowledgment. First, the rapid publication pace in this field, with several central sources published or made available as preprints within the twelve months preceding this analysis, means that the synthesized picture reflects a moving target rather than a settled body of knowledge; conclusions drawn here should be understood as provisional and subject to revision as the underlying architectures continue to mature. Second, because the review

draws predominantly on studies that report positive or partially positive results for VLM adoption, an inherent risk of publication bias must be acknowledged; studies in which VLM-based approaches failed outright to improve upon established baselines are less likely to have been published or indexed, and the present synthesis cannot fully correct for this asymmetry beyond noting it explicitly. Third, the conceptual framework employed, organizing the literature around three interfaces, is itself an interpretive construction developed for this study rather than a framework independently validated in prior scholarship; its value should be assessed by its capacity to generate coherent and useful comparative insight, which is the standard applied in section four, rather than by claims of objective discovery. Ethical considerations for a literature-based study of this kind are comparatively limited relative to studies involving human participants, but the paper nonetheless adheres to standard scholarly norms of accurate attribution, avoidance of selective misrepresentation of cited findings, and transparent acknowledgment of the boundaries of the evidence base.

4. Results and Findings

Applying the three-interface coding framework described in section three to the reviewed literature yields a coherent comparative picture, summarized in Table 1 and elaborated narratively below. Across all three interfaces, a consistent pattern emerges: vision-language models confer their most distinctive advantage not in raw quantitative performance relative to specialized baselines, but in the qualitative expansion of what the system can do, specifically its capacity to generalize to previously unseen conditions and to communicate its reasoning in natural language rather than in opaque numerical scores alone.

Within the human-robot collaboration interface, the reviewed studies converge on the finding that VLMs substantially improve a system's capacity to interpret ambiguous or high-level natural language instructions and to ground those instructions in the immediate visual context of the shop floor, whether through direct scene interpretation, as in the cobot navigation work of Wang, Fan, and Zheng (2024), or through mediation via a digital twin, as in the multi-agent assembly architecture of Liu et al. (2026). The consistent limitation identified across this cluster is the dependence of system performance on the fidelity of the underlying spatial representation, whether a reconstructed point cloud or a synchronized digital twin, and the associated computational latency introduced by real-time multimodal inference, a constraint that becomes particularly

salient in mobile robotics applications requiring rapid response to a dynamically changing environment (Wang, Fan, Zheng, Yan, & Wang, 2025).

Table 1: *Comparative Synthesis of Vision-Language Model Applications across the Three Manufacturing Interfaces*

Application domain	Representative systems	Primary vision-language function	Reported strength	Reported limitation
Human-robot collaborative assembly	Embodied multi-agent VLM systems for digital twin-assisted assembly (Liu et al., 2026)	Task grounding and action planning from visual scene and instruction	Improved translation of high-level instructions into executable robot actions	Dependence on digital twin fidelity and simulation-to-reality transfer
Cobot navigation and guidance	LLM-based vision and language cobot navigation (Wang, Fan, & Zheng, 2024); human-guided mobile robot navigation (Wang, Fan, Zheng, Yan, & Wang, 2025)	Zero-shot semantic segmentation and language-conditioned path planning	Operates in unstructured shop-floor environments without extensive retraining	Latency and computational cost of onboard multimodal inference
Quality inspection and defect detection	Resource-efficient LVLM adaptation for LNG tank inspection (Nguyen, Park, Yoo, & Yang, 2026); EIAD (Zhang, Ruan, Gao, Liu, & Fu, 2025); AnomalyGPT (Gu et al., 2024); VMAD (Deng, Luo, Zhai, Cao, & Kang, 2024)	Joint defect localization and natural language report generation	Approaches or matches specialized detector accuracy while adding interpretability	Weaker performance on fine, low-contrast, or elongated defect geometries
Perception and decision support	Multimodal LLM framework for intelligent perception and decision making (Wang, Zhang, Jiang, & Li, 2025)	Sensor and image fusion for situational reasoning	Consolidates heterogeneous data streams into a single reasoning interface	Limited transparency regarding failure modes under sensor noise

Within the quality inspection interface, the most direct and quantitatively grounded finding in the reviewed literature comes from Nguyen, Park, Yoo, and Yang (2026), whose head-to-head comparison found that a fine-tuned large vision-language model achieved localization accuracy essentially on par with the strongest conventional detector evaluated, while requiring substantially longer per-image inference time due to the additional cost of autoregressive text generation. This finding, considered alongside the recurring reports of degraded performance on fine, low-contrast defect geometries across multiple independent studies (Nguyen et al., 2026; Zhang et al., 2025), supports a nuanced conclusion: VLMs are not yet categorically superior detectors in a narrow accuracy sense, but they consistently offer a distinctive value proposition in structured, explainable, multitask output that specialized detectors cannot natively produce, a value proposition that appears sufficiently compelling to justify continued architectural investment despite the latency and fine-grained accuracy tradeoffs identified.

Within the perception and decision support interface, the evidence base remains comparatively thin, with Wang, Zhang, Jiang, and Li (2025) providing the clearest articulation of the conceptual ambition, namely a general reasoning layer fusing heterogeneous multimodal production data, but without the extensive comparative benchmarking available in the inspection literature. This asymmetry across interfaces is itself an important finding of the present synthesis, indicating that VLM adoption in manufacturing is proceeding unevenly, with quality inspection representing the most empirically mature application and general decision support representing the most conceptually ambitious but least validated.

5. Discussion

The comparative synthesis presented above supports the paper's central conceptual claim: vision-language models function across human-robot collaboration, quality inspection, and multimodal decision support as a common interpretive layer that translates between unstructured, often linguistically expressed human intent or ambiguous sensory context and the structured representations required by downstream industrial control and reporting systems. This reframing carries several implications. For manufacturing organizations considering VLM adoption, it suggests that evaluation should not proceed solely along the axis of raw detection or classification

accuracy, where specialized discriminative models often remain competitive, but should weight interpretability, generalization to novel conditions, and reduction in system engineering overhead as first order criteria, consistent with the explicit tradeoffs documented by Nguyen et al. (2026) between VLM and conventional detector deployment.

The findings also lend empirical support to the hybrid architecture proposition advanced across multiple strands of the reviewed literature, in which lightweight specialized models handle high-throughput, latency-sensitive tasks while VLMs are reserved for semantic interpretation, novel condition handling, and human-facing communication. This division of labor addresses the latency limitation consistently identified in both the navigation literature (Wang, Fan, Zheng, Yan, & Wang, 2025) and the inspection literature (Nguyen et al., 2026), while preserving the interpretability and generalization advantages that motivate VLM adoption in the first place. Situated within the broader Industry 4.0 to Industry 5.0 transition, in which manufacturing paradigms increasingly emphasize human-centricity alongside efficiency, this hybrid architectural pattern also aligns with a socio-technical vision of production systems in which machines communicate their reasoning to human operators rather than functioning as opaque black boxes.

Several limitations of the present synthesis should be acknowledged. Because the underlying literature is dominated by studies reporting successful or partially successful VLM deployments, the picture assembled here may understate the frequency and severity of outright failures or negative results that remain underrepresented in the published record. The conceptual three-interface framework, while analytically productive for organizing the reviewed evidence, was developed specifically for this study and has not been independently validated against alternative organizing schemes; future scholarship might productively test whether alternative framings yield different or complementary insights. Additionally, because this is a conceptual and literature-based synthesis rather than a primary empirical study, it cannot resolve open quantitative questions, such as the precise tradeoff curve between fine-tuning depth and generalization retention, that require dedicated experimental investigation.

These limitations point toward a concrete future research agenda. Priority areas include the systematic characterization of the fine-tuning generalization tradeoff identified in section two; the development of standardized, cross-study benchmarks for multimodal decision support analogous to the mean average precision standards already established in the inspection literature; and

empirical investigation of hybrid detector-VLM architectures under realistic production throughput constraints rather than offline evaluation alone. Extending this inquiry to the broader Industry 5.0 discourse, future work might also examine how VLM-mediated communication between machines and human operators shapes worker trust, cognitive workload, and skill development over sustained periods of deployment, questions that lie beyond the scope of the technical literature reviewed here but that bear directly on the human-centric ambitions increasingly associated with next-generation manufacturing paradigms.

6. Conclusion

This paper has synthesized a rapidly growing but fragmented body of literature on the application of multi-modal vision-language models within smart manufacturing, organizing recent contributions in human-robot collaboration, cobot navigation, quality inspection, anomaly detection, and multimodal perception around a unifying conceptual framework that positions VLMs as an interpretive layer mediating between unstructured human and sensory input and structured industrial action. The synthesis finds that VLMs consistently expand system capability along dimensions of generalization, interpretability, and natural language interaction, even where they do not categorically surpass specialized discriminative baselines in narrow accuracy terms, and that persistent limitations, including inference latency, dependence on high-fidelity digital or spatial representations, and degraded precision on fine-grained visual detail, recur across otherwise disparate application domains. The paper's principal scholarly contribution lies in this integrative reframing, which extends prior domain-specific reviews by articulating the shared theoretical mechanism underlying their disparate empirical successes. Future research should prioritize standardized cross-domain evaluation, systematic study of the fine-tuning generalization tradeoff, and empirical investigation of hybrid architectures that pair the speed of specialized detectors with the reasoning and communicative capacity of vision-language models, positioning the manufacturing sector to more deliberately harness this emerging technological layer as it continues to mature.

References

1. Deng, H., Luo, H., Zhai, W., Cao, Y., & Kang, Y. (2024). VMAD: Visual-enhanced multimodal large language model for zero-shot anomaly detection (arXiv:2409.20146). arXiv. <https://arxiv.org/abs/2409.20146>
2. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 10088–10115.
3. Fan, J., Yin, Y., Wang, T., Dong, W., Zheng, P., & Wang, L. (2025). Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. *Frontiers of Engineering Management*, 12(1), 177–200. <https://doi.org/10.1007/s42524-025-4136-9>
4. Gonzalez Santacruz, E., Romero, D., Noguez, J., & Wuest, T. (2024). Integrated quality 4.0 framework for quality improvement based on six sigma and machine learning techniques towards zero-defect manufacturing. *The TQM Journal*, 37, 1115–1155. <https://doi.org/10.1108/TQM-11-2023-0361>
5. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., & Wang, J. (2024). AnomalyGPT: Detecting industrial anomalies using large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3), 1932–1940. <https://doi.org/10.1609/aaai.v38i3.27963>
6. Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., & Addepalli, P. V. (2022). A Review of Frameworks, and Impact on the Software Development Life Cycle with Artificial Intelligence in Software Engineering. *American International Journal of Computer Science and Technology*, 4(1), 44-53.
7. Manoharan, P., Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., & Shyam, D. (2022). Understanding the Use of Agile Development Practices Alongside Quantum Computing in Modern Engineering Projects. *International Journal of Emerging Research in Engineering and Technology*, 3(1), 136-143.
8. Routhu, K. K. (2023). AI-driven succession planning in Oracle HCM Cloud: Building resilient leadership pipelines through predictive analytics. *International Journal of Science, Engineering and Technology*, 11(5).
9. Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., & Syed, M. H. (2022). Intelligent Predictive System of Cloud Resource Utilization Forecasting with Advanced Deep Learning. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(1), 143-151.
10. Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., & Manoharan, P. (2022). A Survey of Statistical Process Control and Its Applications in Manufacturing Quality Assurance. *International Journal of Emerging Trends in Computer Science and Information Technology*, 3(1), 127-135.
11. Padur, S. K. R. (2023). AI-Augmented Enterprise ERP Modernization: Zero-Downtime Strategies for Oracle E-Business Suite R12. 2 and Beyond. *Available at SSRN 5605510*.
12. Routhu, K. K. (2019). Hybrid machine learning architecture for absence forecasting within Oracle Cloud HCM. *KOS Journal of AIML, Data Science, and Robotics*, 1(1), 1-5.
13. Padur, S. K. R. (2022). AI augmented platform engineering, transforming developer experience through intelligent automation and self optimizing internal platforms. *International Journal of Science, Engineering and Technology*, 10(5), 10-5281.
14. Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., & Ragireddy, U. K. (2022). Survey on Blockchain Integration with Sap S/4hana for Transparent and Secure Supply

- Chains. *International Journal of AI, BigData, Computational and Management Studies*, 3(1), 161-170.
15. Routhu, K. K. (2023). Embedding fairness into the digital enterprise: Data-driven dei strategies with oracle hcm analytics. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 9(8), 266-274.
 16. Mylavarapu, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., & Panchali, S. H. (2022). Comparative Performance of Machine Learning Techniques for Sentiment and Insight Analysis in Amazon Product Reviews. *American International Journal of Computer Science and Technology*, 4(1), 54-64.
 17. Mohammad, J. A., Panchali, S. H., Mylavarapu, K. B., Pilli, J., & Jannu, P. K. (2022). An Ensemble Machine Learning Approach for Reliable Fake News Detection on Social Media Platforms. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(1), 152-161.
 18. Pilli, J., Jannu, P. K., Mohammad, J. A., Panchali, S. H., & Mylavarapu, K. B. (2022). Review of Supply Chain-Integrated CRM Systems for Business Agility. *International Journal of Emerging Research in Engineering and Technology*, 3(1), 155-163.
 19. Padur, S. K. R. (2022). Intelligent resource management: AI methods for predictive workload forecasting in cloud data centers. *J. Artif. Intell. Mach. Learn. & Data Sci*, 1(1), 2936-2941.
 20. Panchali, S. H., Mylavarapu, K. B., Pilli, J., Jannu, P. K., & Mohammad, J. A. (2022). Predictive Maintenance in Smart Manufacturing: A Review of Machine Learning and Digital Twin Approaches. *International Journal of AI, BigData, Computational and Management Studies*, 3(1), 150-160.
 21. Jannu, P. K., Mohammad, J. A., Panchali, S. H., Mylavarapu, K. B., & Pilli, J. (2022). Heat Dissipation Strategies for High-Performance Computing: A Review of Air and Liquid Cooling Efficiency. *International Journal of Emerging Trends in Computer Science and Information Technology*, 3(1), 145-153.
 22. Routhu, K. K. (2021). AI-augmented benefits administration: A standards-driven automation framework with Oracle HCM Cloud. *International Journal of Scientific Research and Engineering Trends*, 7(3).
 23. Taluri, R. (2022). Cloud Data Engineering Strategies for Large-Scale Financial Data Integration and Intelligent Corporate Performance Reporting. *International Journal of Emerging Research in Engineering and Technology*, 3(4), 176-188. <https://doi.org/10.63282/3050-922X.IJERET-V3I4P119>
 24. Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., Chilakapati, V. K., & Keshireddy, S. R. (2022). Machine Learning-Driven Threat Analysis for Enhancing Security in Data Center Networking. *American International Journal of Computer Science and Technology*, 4(1), 65-76.
 25. Padur, S. K. R. (2021). Bridging Human, System, and Cloud Integration through RESTful Automation and Governance. *the International Journal of Science, Engineering and Technology*, 9(6).
 26. Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2022). Towards Secure and Reliable IoT: A Review on Anomaly Detection and Predictive Analytics in Sensor Networks. *International Journal of AI, BigData, Computational and Management Studies*, 3(1), 140-149.
 27. Padur, S. K. R. (2021). Deep learning and process mining for ERP anomaly detection: Toward predictive and self-monitoring enterprise platforms. *Available at SSRN 5605531*.

28. Routhu, K. K. (2020). Strategic compensation equity and rewards optimization: A multi-cloud analytics blueprint with Oracle Analytics Cloud. *Available at SSRN 5737266*.
29. Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., & Kavuluri, H. V. R. (2022). Effective Monitoring of Sensor-Actuator Integration Methods for Real-Time Controlling in Smart Device Networks. *International Journal of Emerging Research in Engineering and Technology*, 3(1), 144-154.
30. Padur, S. K. R. (2021). From Control to Code: Governance Models for Multi-Cloud ERP Modernization. *International Journal of Scientific Research & Engineering Trends*, 7(3).
31. Routhu, K. K. (2019). AI-enhanced payroll optimization: Improving accuracy and compliance in Oracle HCM. *KOS Journal of AIML, Data Science, and Robotics*, 1(1), 1-5.
32. Padur, S. K. R. (2021). From Scripts to Platforms-as-Code: The Role of Terraform and Ansible in Declarative Infrastructure Rollouts. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 7(1), 621-628.
33. Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., & Dasari, A. (2022). Review on Network Virtualization and SDN in Industrial IoT (IIoT). *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(1), 162-172.
34. Dasari, S., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., & Pathani, A. K. (2022). Networking Paradigms for Large-Scale IoT Deployments: A Survey on Reliability, Latency, and Scalability. *International Journal of Emerging Trends in Computer Science and Information Technology*, 3(1), 136-144.
35. Padur, S. K. R. (2020). AI-augmented disaster recovery simulations: From chaos engineering to autonomous resilience orchestration. *Int J Sci Res Sci Eng Technol*, 7(6), 367-378.
36. Taluri, R. (2021). Cloud-Native Architectures for Enterprise Financial Data Management, Analytics, and Regulatory Reporting Compliance. *International Journal of Emerging Trends in Computer Science and Information Technology*, 2(2), 101-111. <https://doi.org/10.63282/3050-9246.IJETCSIT-V2I2P112>
37. Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., & Addepalli, P. V. (2023). Data-Driven Approaches for Forecasting Cost Overruns in Infrastructure Projects. *European Journal of Technology*, 7(5), 23-34.
38. Dhuli, S., Prabu, M., Muzaffer, H. S., Uday Kumar, R., & Prasanth Varma, A. (2023). A Survey on Critical Success Factors Influencing Project Success Across Multiple Industries. *J Contemp Edu Theo Artif Intel: JCETAI-106*.
39. Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., & Syed, M. H. (2023). An Overview of Different Automation Strategies Used in Cloud-Based CI/CD Pipelines for Software Deployment. *International Journal of Technology, Management and Humanities*, 9(01), 116-127.
40. Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., & Ragireddy, U. K. (2023). Performance Analysis of Machine Learning Models for Cybersecurity Risk Classification in IT Projects. *Universal Library of Engineering Technology*, (Issue).
41. Mylavarapu, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., & Panchali, S. H. (2023). Artificial Intelligence for Continuous Quality Assurance in Cloud-Native Applications: A Review. *European Journal of Technology*, 7(5), 35-49.
42. Routhu, K. K. (2017). The evolution of HR from on-premise to Oracle Cloud HCM: Challenges and opportunities. *International Journal of Scientific Research & Engineering Trends*, 3(1).

43. Padur, S. K. R. (2020). From centralized control to democratized insights: Migrating enterprise reporting from IBM Cognos to Microsoft Power BI. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 6(1), 218-225.
44. Jannu, P. K., Mohammad, J. A., Panchali, S. H., Mylavarapu, K. B., & Pilli, J. (2023). A Survey on Energy-Efficient Microarchitecture Design Strategies for Mobile CPUs and GPUs: Techniques and Future Directions. *International Journal of Technology, Management and Humanities*, 9(04), 137-151.
45. Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2023). An AI-Based Framework for Detecting IoT Botnets Through Network Traffic Analysis and Modeling. *Universal Library of Engineering Technology*, (Issue).
46. Kavuluri, H. V. R., Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., & Nagumotu, V. T. (2023). Cloud Data-Center Network Verification: Approaches, Algorithms and Toolchains. *American Journal of Computing and Engineering*, 6(2), 18-32.
47. Routhu, K. K. (2022). From RFID to Geofencing: IoT-Enabled Smart Time Tracking in Oracle HCM Cloud. *International Journal of Science, Engineering and Technology*, 10(4).
48. Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., & Dasari, A. (2023). A Survey on 5G/6G Radio Access Networks: Technologies, Protocols, and Performance Metrics. *International Journal of Technology, Management and Humanities*, 9(02), 79-93.
49. Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2023). An AI-Based Framework for Detecting IoT Botnets Through Network Traffic Analysis and Modeling. *Universal Library of Engineering Technology*, (Issue).
50. Routhu, K. K. (2022). From Case Management to Conversational HR: Redefining Help Desks with Oracle's AI and NLP Framework. *International Journal of Science, Engineering and Technology*, 10(6).
51. Padur, S. K. R. (2019). Machine learning for predictive capacity planning: Evolution from analytical modeling to autonomous infrastructure. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 5(5), 285-293.
52. Khanam, R., Hussain, M., Hill, R., & Allen, P. (2024). A comprehensive review of convolutional neural networks for defect detection in industrial applications. *IEEE Access*, 12, 94250–94295. <https://doi.org/10.1109/ACCESS.2024.3425166>
53. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning, PMLR 202*, 19730–19742.
54. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., & Shi, G. (2025). A survey of state of the art large vision language models: Benchmark evaluations and challenges. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1587–1606.
55. Liu, C., Qian, Y., Tang, D., Zhu, H., Pang, J., & Cai, Q. (2026). From insight to action: Embodied multi-agent system integrating vision language model for digital twin-assisted human-robot collaborative assembly. *Journal of Manufacturing Systems*, 85, 531–556.
56. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 34892–34916.

57. Nguyen, H. V., Park, H., Yoo, N., & Yang, J. (2026). Resource-efficient adaptation of large vision–language models for multimodal defect inspection in LNG tank manufacturing. *The International Journal of Advanced Manufacturing Technology*, 144, 5135–5152. <https://doi.org/10.1007/s00170-026-17818-y>
58. Oks, S. J., Jalowski, M., Lechner, M., Möslin, K. M., et al. (2024). Cyber-physical systems in the context of Industry 4.0: A review, categorization and outlook. *Information Systems Frontiers*, 26, 1731–1772. <https://doi.org/10.1007/s10796-022-10252-x>
59. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 8748–8763.
60. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.
61. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
62. Sundaram, S., & Zeid, A. (2023). Artificial intelligence-based smart quality inspection for manufacturing. *Micromachines*, 14(3), 570. <https://doi.org/10.3390/mi14030570>
63. Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, 94, 3563–3576.
64. Wang, T., Fan, J., & Zheng, P. (2024). An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing. *Journal of Manufacturing Systems*, 75, 299–305.
65. Wang, T., Fan, J., Zheng, P., Yan, R., & Wang, L. (2025). Vision-language model-based human-guided mobile robot navigation in an unstructured environment for human-centric smart manufacturing. *Engineering*. <https://doi.org/10.1016/j.eng.2025.04.028>
66. Wang, T., Zhang, B., Jiang, D., & Li, D. (2025). A multimodal large language model framework for intelligent perception and decision-making in smart manufacturing. *Sensors*, 25(10), 3072.
67. Zhang, Z., Ruan, J., Gao, X., Liu, T., & Fu, Y. (2025). EIAD: Explainable industrial anomaly detection via multi-modal large language models (arXiv:2503.14162). *arXiv*. <https://arxiv.org/abs/2503.14162>