International Core Journal of Engineering Volume 11 Issue 5, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0039

Architecture Design of a Smart Logistics Warehouse Management System based on Multi-modal AI

Can Liang[a], Liangxu Sun[b,*], Shuaiye Luo[c], Xingnuo Liu[d], Ruihao Wu[e]

University of Science and Technology Liaoning, Anshan 114051, China

[a]2309769282@qq.com, [b,*]sunliangxumail@163.com, [c]1966988701@qq.com, [d]3137130844@qq.com, [e]1916683625@qq.com

Abstract

Addressing the inefficiency and poor real-time performance of traditional warehouse management systems, this paper proposes a design scheme for a smart warehouse management system based on multi-modal AI. The system achieves intelligent management of the entire logistics process through the collaborative decision-making of multi-modal AI (DeepSeek-R1 language model and GLM-4V visual model). Combining a modular micro-services architecture (SpringCloud) and distributed technology, it constructs a closed-loop system of "perception-decision-monitoring." The design integrates two-factor authentication and edge-cloud collaborative computing, supporting dynamic inventory optimization and real-time data analysis, providing a lightweight and scalable solution for the intelligent upgrade of logistics warehousing. Future work will further explore the deep integration application of digital twin technology.

Keywords

Smart Warehouse; Multi-modal AI; Micro-services; Internet of Things.

1. Introduction

Current warehouse management faces dual challenges of efficiency and real-time performance: Traditional manual inbound and outbound processes have an error rate exceeding 5%, and inventory counts can take several hours. Existing IoT technologies are often limited to single-modal applications, such as the independent use of RFID or visual inspection, lacking cross-modal collaborative decision-making capabilities. Although research has attempted to integrate AI and IoT technologies, system architectures are rigid and lack scalability, making it difficult to support dynamic business needs and edge-cloud collaborative computing [1]. This paper proposes a lightweight smart warehouse system that achieves state recognition of goods, anomaly warnings, and closed-loop inventory optimization through the multi-modal fusion of DeepSeek-R1 language model and GLM-4V visual model. Based on the Spring Cloud micro-services architecture, the design provides a modular and scalable solution, integrating distributed transactions and edge computing, offering a low-cost and loosely coupled intelligent upgrade path for small and medium-sized warehouses, overcoming the traditional system's real-time performance and scalability limitations.

2. System Architecture Design

2.1 Overall Architecture Design

The smart express logistics warehouse management system adopts a layered architecture design, constructing a "perception-network-data-application" four-layer collaborative system . The perception layer integrates RFID electronic tags and multi-type sensors to achieve second-level

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0039

identification of goods and real-time monitoring of environmental conditions. Among them, visual cameras and the GLM-4V model work in conjunction to support abnormal detection of pallet stacking. The network layer ensures high throughput and low latency data interaction through Wi-Fi6 and 5G dual channels, combined with fiber-optic backbone networks to achieve synchronized data across multiple warehouses [2,3].

The data layer is based on the Hadoop+Spark distributed architecture, utilizing the Kafka stream processing engine to achieve real-time updates of inventory status. Cold and hot data are stored in MinIO object storage and Redis clusters, respectively, supporting TB-level data processing and millisecond-level query responses. The application layer adopts the SpringCloud micro-services architecture, decomposed into independent modules such as inventory management, intelligent decision-making, and user authentication. Services are decoupled through Restful APIs and message queues (RocketMQ), supporting dynamic scaling and coordinated scheduling of edge-cloud resources, meeting the elastic demands of high-concurrency scenarios.

This design, through the collaboration of multi-modal perception, distributed computing, and modular services, constructs a lightweight and scalable warehousing management framework, avoiding hardware dependencies, and providing a low-cost intelligent upgrade path for small and medium-sized warehouses.

2.2 Multi-modal AI Collaborative Decision-Making

In the smart express logistics warehouse management system, multi-modal AI collaborative decision-making technology integrates text, visual, and sensor data to build comprehensive perception and intelligent response capabilities, driving the optimization of the entire warehousing management process.

Based on the collaborative architecture of the DeepSeek-R1 large language model and the GLM-4V visual model, the system realizes complementary cross-modal information. The DeepSeek-R1 is based on the Transformer architecture, with pre-training using logistics corpora (containing 100,000 work order texts) and model compression by 30% through knowledge distillation technology (layered feature matching and adaptive Temperature Regulation).

GLM-4V: Adopting a cascaded architecture of YOLOv8 object detection and Mask R-CNN instance segmentation, it balances detection and segmentation tasks through a dynamic weight allocation mechanism (DWA). Its multi-modal feature fusi on method draws inspiration from the state-of-the-art achievements in RGB-D image salient object detection in deep learning[4], as shown in the following formula:

$$w_{\text{det}} = \frac{\text{IoU}_{\text{det}}}{\text{IoUU}_{\text{det}}+\text{mAP}_{\text{seg}}}, w_{\text{seg}} = 1 - w_{\text{det}} \quad (1)$$

This formula borrows from the multi-modal feature fusion method in FusionRCNN from CVPR2023, dynamically adjusting the weights of detection and segmentation tasks to achieve optimal model performance (F1-sco re improved by 12%). During training, textual data is enhanced for robustness through synonym replacement and sentence restructuring, while visual data is optimized for feature extraction using random cropping and illumination perturbation.

2.3 micro-services Architecture

The system adopts a distributed micro-services architecture, implemented based on the Spring Cloud Alibaba framework to achieve functional decoupling and elastic scaling. In terms of service decomposition, the core business modules are independently divided into three major services:

(1) Inventory Service: Storing goods information through a MySQL sharding strategy (horizontally partitioned by shelf) to support high-frequency read and write operations (TPS ≥ 5,000), and using pessimistic locking mechanisms to prevent overselling. Vitess middleware is used to implement dynamic sharding routing, ensuring cross-database transaction consistency through GTID (global

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0039

transaction ID), and integrating read-write separation strategies (write on the master, read on the slave) to reduce the load o n the master database;

(2) AGV Scheduling Service: Utilizing RocketMQ transaction message queues to achieve asynchronous task dispatching and state synchronization, ensuring reliable transmission of AGV commands (message delivery rate ≥ 99.99%). Prioritizing task order dynamic ally based on path congestion index and AGV battery level, employing a delayed message mechanism (Delayed Message) for retry logic, and using message tracing (Trace Topic) to achieve full-chain tracking;

(3) AI Analysis Service: Leveraging Redis cluster caching for hot data (such as real-time inventory status, AGV positions) and maintaining data through an LRU eviction strategy.

Maintaining a cache hit rate ≥85%. Integrated lightweight AI models (DeepSeek-R1, GLM-4V) provide edge inference support, optimizing model inference speed with ONNX Runtime, and building a real-time performance monitoring dashboard based on Prometheus and Grafana. This architecture, through modular decomposition and distributed collaboration, significantly enhances system throughput and disaster recovery capabilities (single-node failures are seamlessly switched), providing high-reliability technical support for large-scale warehousing scenarios.

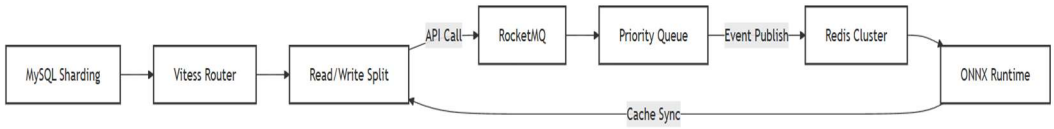

Figure 1. Micro-service Architecture Interaction Flow.

3. System Function Modules

3.1 Inventory Management Module

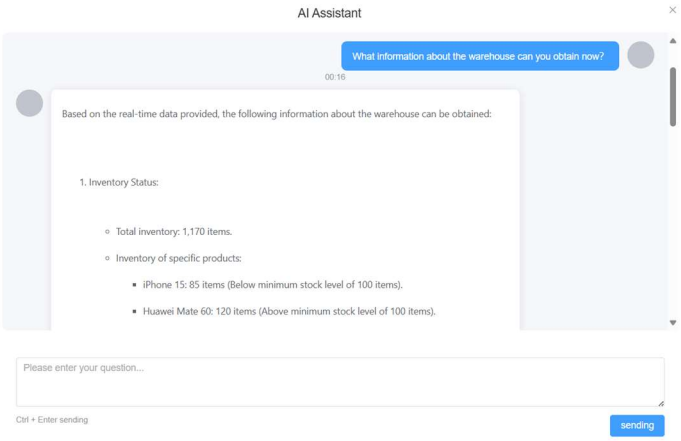

Figure 2. AI Intelligent Assistant Interface.

The inventory management module relies on IoT awareness layer devices to achieve fully automated operations. When goods enter the warehouse, RFID readers capture electronic tag information (EPC code) at a scanning frequency of 20 times per second, pushing raw data to a Kafka message queue via Wi-Fi6 network. The Spark stream processing engine completes data cleaning and format standardization (such as removing duplicate tags and filling timestamps). The processed structured data is stored in a MySQL sharded cluster, wit h secondary indexes established by shelf partitioning to support millisecond-level queries. Meanwhile, the DeepSeek-R1 language model analyzes real-time inventory trend changes, generating dynamic optimization suggestions based on historical

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0039

inbound and outbound records and order prediction data (e.g., "SKU-1001 inventory exceeds the average by 30%, suggesting redistribution to high-turnover area B2"), and pushing these suggestions to management mobile terminals via Restful API[5]. This module achieves second- level updates and intelligent allocation of inventory status through a "sensing-transmitting-analyzing-deciding" closed-loop design.

3.2 Section Headings

The intelligent early warning module utilizes a multi-modal AI collaborative architecture to achieve comprehensive monitoring of abnormal events. In the visual monitoring sub-module, the GLM-4V visual model deployed at the top of the warehouse captures surveillance footage at a frequency of 30 FPS, using the YOLOv8 object detection algorithm to locate the cargo stacking area. It then combines Mask R-CNN instance segmentation to extract the cargo contour features. When the tilt angle exceeds 15° or the cargo displacement distance exceeds 50 cm, a three-level alert mechanism is triggered (pop-up window → SMS notification → emergency work order generation). At the same time, the DeepSeek-R1 language model synchronously intervenes, leveraging natural language processing technology to parse the context of the alert event, automatically generating multilingual handling solutions (such as "Immediately halt AGV operations in Area A3 and execute manual re-inspection"), and distributing them through the message center module to relevant responsible persons. The two models complement each other through a dynamic weighted fusion strategy: the visual model provides spatial positioning and morphological features, while the language model endows semantic understanding and decision making inference capabilities. Ultimately, the fusion of features is achieved through the formula:

$$F_{fused} = \alpha \cdot F_{visual} + (1-\alpha) \cdot F_{text}$$

where the weight coefficient α is dynamically adjusted based on the type of abnormality), outputting a comprehensive risk assessment result, significantly reducing false alarm rates.

Code Example (Python):

```
def fuse_modalities(text_feature: np.ndarray,
                    image_feature: np.ndarray,
                    alert_type: str) -> np.ndarray:
    //Fuse text and visual features with dynamic weights.
    Args:
      text_feature: Text embedding from DeepSeek-R1 (dim: 768)
      image_feature: Visual embedding from GLM-4V (dim: 1024)
      alert_type: Alert category (tilt/missing/other)
       Returns:
        Fused feature vector (dim: 1024)
  ///
       # Dynamic weight adjustment
       alpha = 0.7 if alert_type == "tilt" else 0.3    # Higher weight for visual in tilt alerts
       # Project text features to 1024D space
       text_proj = np.dot(text_feature, projection_matrix)
       # Weighted fusion
       return alpha * image_feature + (1 - alpha) * text_proj
```

3.3 AI Synergistic Application

DeepSeek-R1 and GLM-4V's collaboration extends beyond anomaly response, encompassing the entire lifecycle of warehouse management. In inventory counting scenarios, GLM-4V reconstructs the 3D models of shelves using point cloud data, identifying detailed issues such as damaged goods

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0039

surfaces and fallen labels. Meanwhile, DeepSeek-R1 synchronously interprets vague instructions from work orders (e.g., "prioritize fragile goods areas") into specific operational parameters (storage location coordinates, AGV priority levels). Both systems achieve cross-modal alignment through shared feature spaces: visual features are encoded into semantic vectors by the CLIP model and matched with the output of the language model in the latent space to ensure consistency between operational instructions and physical scenarios. Additionally, DeepSeek-R1 continuously absorbs human feedback through online learning mechanisms, dynamically updating its knowledge base (e.g., new goods classification rules) and synchronizing incremental parameters to GLM-4V's visual classifier, forming a bidirectional optimization intelligent enhancement loop.

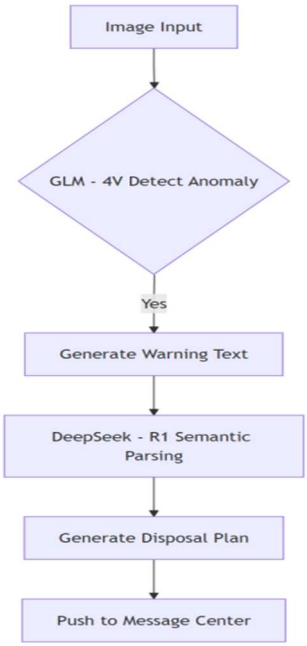

Figure 3. Abnormal Handling Process.

4. Conclusion

The multi-modal AI warehousing system proposed in this paper, through dynamic feature fusion and micro-service atomic design, effectively addresses the traditional warehousing system's issues of modality fragmentation and insufficient scalability from a theoretical perspective, Of small and medium for Intelligent Transformation of providing a light weight technological path for the-sized warehouses. However, current research still has certain limitations: Firstly, the system needs optimization in terms of real-time performance and robustness in multi-modal collaboration; the model's generalization capability in complex scenarios (such as dynamic lighting interference and multiple object occlusions) remains to be verified. Secondly, the performance stability of edge-cloud collaboration mechanisms in low-bandwidth environments has not been sufficiently explored, and the resource scheduling efficiency of the distributed architecture requires further theoretical deepening. Thirdly, while th e system security design introduces a zero-trust architecture, strategies for defending against privacy risks in cross-modality data (such as the risk of leakage of the association between visual and textual data) still need to be improved. Future research Will focus on digital twin real-time linkage, cross-warehouse model collaborative training driven by federated learning, and photovoltaic-powered edge nodes, while exploring lightweight model compression and heterogeneous hardware adaptation technologies to overcome compatibility and energy efficiency bottlenecks in practical deployment and promote the transformation of theoretical achievements into industrial applications.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0039

Acknowledgments

This research was supported by the 2025 College Student Innovation and Entrepreneurship Training Program of University of Science and Technology Liaoning.

References

[1] Aibole Robot. Future of WMS Warehouse Management Systems: Trends in Artificial Intelligence, Big Data, and Internet of Things [R/OL]. 2023-11-02.

[2] Xiao Guangwei, Chen Hao, Shao Shizhou, et al. Research on an Intelligent Warehouse Management Model Based on Internet of Things Technology [J]. Logistics Technology and Applications, 2021, 45(4): 58-63.

[3] Advancing Inventory Management and Logistics Efficiency through AI Large Models and Intelligent Warehouse Management [RR/OL]. Original Power Document, 2024.

[4] Huang Nianchang, Yang Yang, Zhang Qiang, et al. Advances in Deep Learning for Image Significance Target Detection in RCB-D [J]. Chinese Journal of Computers, 2025, 48(2): 284-316. DOI:10.11897/SP.J.1016.2025.00284.

[5] JD Logistics. White Paper on Smart Logistics Technology: AI and Internet of Things Integration Practices [RR]. Beijing: JD Group, 2024.