International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 5, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n5.06
Journal homepage: https://wepub.org/index.php/IJCSIT/index

# Research on Multimodal AGI Empowering the Development of LLM-based Multi-Agent System

Zirun Bai

ArtCenter College of Art and Design, California, 91103, United States

## ABSTRACT

This paper analyzes the multimodal general artificial intelligence-enabled large language model multi-agent system, focusing on exploring the internal logic, technical laws, and practical scope of their integration. By integrating domestic and foreign literature, technical comparisons, and industrial cases from 2022 to 2025, and combining authoritative achievements in the fields of multi-agent and multimodal analysis, the analysis reveals that multimodal fusion can compensate for the shortcomings of traditional text intelligent agent perception, interaction, and collaboration, forming a complete perception–inference–execution chain. After the technology is implemented, the adaptability and stability of the agent are significantly improved, which reduces decision-making bias in individual models. Currently, demand for intelligent physical scenarios has surged, and traditional text agents are difficult to adapt to complex dynamic environments. Existing research mostly focuses on a single technical dimension and lacks systematic fusion analysis. This study fills this gap and provides reliable references and practical support for industrial upgrading. The study also identifies existing challenges and viable development pathways, and the conclusions drawn can provide direct basis for technological iteration, scenario implementation, and industry standardization improvement.

## KEYWORDS

Multimodal General AI; Large Language Model (LLM); Multi-Agent System; Cross-Modal Fusion; Collaborative Operation

## 1. INTRODUCTION

In recent decades, Large Language Model (LLM) technology has evolved rapidly, and industrial applications have moved from laboratories to large-scale implementation. Multi-Agent collaborative systems developed based on its underlying architecture have been applied in multiple fields such as government affairs, manufacturing, education, and services, emerging as a key enabler for processing complex intelligent tasks. However, traditional large language models heavily rely on massive text training data, with outstanding language parsing and logic generation capabilities, but are limited to a single text interaction mode. There are obvious shortcomings in real-world perception and multi-agent collaboration, hindering the full rollout of intelligent systems from pilot to market-oriented and large-scale implementation.

The rise of multimodal general artificial intelligence has broken down the inherent barriers of pure text processing, integrating diverse data such as visual, speech, and dynamic scenarios for unified semantic analysis, and laying a technical foundation for intelligent agents to interface with physical scenarios. According to Web of Science data, since 2022, the number of relevant academic papers and technical patents worldwide has steadily increased, and industry research popularity and attention have continued to rise.

Domestic research institutions and technology companies have made parallel progress, and localized multimodal models such as Zidong Taichu 4.0 and Youtu-Agent have been developed and tested. The integration of the two technologies is still in the early exploration stage [1]. The current mainstream multi-agent systems generally have problems such as low information integration efficiency, incomplete internal interaction mechanisms, and weak long-term self-learning ability, and the core value of cluster collaboration is not fully realized. Based on the current situation of the industry, this article focuses on the empowering role of multimodality, sorts out the integration theory, implementation mode, and optimization path, which is in line with the trend of the integration of digital economy and real economy, and has both theoretical exploration value and industrial practical significance. Rising demand for intelligent physical scenarios makes multimodal empowerment critical to addressing traditional agent limitations, aligning with industrial upgrading needs.

## 2. THEORETICAL BASIS FOR THE INTEGRATION OF MULTIMODAL GENERAL ARTIFICIAL INTELLIGENCE AND LARGE LANGUAGE MODEL MULTI-AGENT SYSTEM

The key to analyzing the integration logic of the two technologies is to clarify the theoretical foundation and inherent connections. Artificial intelligence, cognitive intelligence, and distributed collaboration theory provide key support for fusion. The two are more than a theoretical deduction, but naturally have feasible conditions and practical needs for fusion [2].

The core goal of multimodal general artificial intelligence is to build a unified semantic representation system to standardize and map semantic information of dynamic information in text, images, speech, and real-world dynamic information. The three-tier structure of information preprocessing, cross-modal matching and decision output is the core operating framework of this technology. The domestic models such as Zidong Taichu and ERNIE Bot follow this logic to complete iterative optimization. Relevant practice achievements have been included in the industry's core journals, which can effectively remedy the perception shortcomings of traditional text agents and enable agents to have the ability to understand real information [3].

Mature large language models can serve as basic units to build hierarchical and networked multi-agent collaboration systems. By endowing intelligent agents with basic abilities such as independent thinking, instruction execution, and information exchange, and relying on functional decomposition to complete complex tasks, hierarchical coordination and equal collaboration are the mainstream forms of construction. Exclusive scheduling intelligent agents are responsible for resource allocation, while ordinary intelligent agents independently communicate and promote work. The relevant architecture has been deployed in scientific research and small and medium-sized enterprise scenarios. Traditional text-based clusters draw on a knowledge base decision-making, and the lack of real-world information can easily lead to decision-making that is detached from reality, resulting in execution solidification problems.

The core of integrating the two technologies is to builds a bridge between perception and collaborative cognition. Real scene information can be synchronously transmitted to various intelligent agents, and the large language model processes heterogeneous data, relying on mature theories to build a complete operation process, and breaking away from the text-dominated mode. There is no ideological obstacle to the integration of the two, and industry practice only needs to complete adaptation and mechanism construction, steadily promoting implementation. From the perspective of industry practice, this integration theory can not only guide technology research and development, but also provide reference for cross domain adaptation and heterogeneous architecture compatibility, bridging the connection point between technical principles and practical implementation, and providing support for subsequent path planning.

This integrative framework, rather than a simple logical derivation, combines technical principles, cognitive laws and industrial needs. It can connect bottom-level R&D with upper level applications, provide flexible support for technology iteration and scenario adaptation, avoid theoretical detachment from practice, and enhance guidance practicality.

## 3. THE CORE TECHNOLOGY IMPLEMENTATION PATH OF EMPOWERING MULTI-AGENT SYSTEMS WITH MULTIMODAL GENERAL ARTIFICIAL INTELLIGENCE

Based on fusion theory, multimodal technology provides technical support for multi-agent systems from four aspects: perception upgrading, information fusion, collaborative management, and long-term learning. The above technical ideas have been validated in industrial trials and have the conditions for implementation.

Upgrading perception ability is the primary prerequisite for intelligent agents to interface with real-world scenarios. Traditional intelligent agents only receive text instructions and cannot recognize concrete information such as images, speech, and dynamic scenarios, limiting perception to text-only inputs. After being equipped with lightweight audiovisual and environmental sensing information collection modules, the dimensions of intelligent agent information acquisition have significantly expanded. the field commonly pairs lightweight audiovisual models with large models, retaining the underlying architecture unchanged and significantly reducing computing power costs. The upgraded intelligent agent can autonomously perceive the environment, complete voice interaction, and significantly improve the human-machine adaptation effect. The relevant mode has been tested through public experiments.

Different information naturally has semantic barriers, making it difficult to communicate directly, often causing cognitive biases and lower collaborative efficiency. Cross-modal semantic fusion technology establishes a unified transformation standard, which sorts multi-source scene information into standardized semantics that can be accurately recognized by large models. The industry commonly uses contrastive learning to optimize alignment effects, and core industry journals have confirmed the effectiveness of this method. Intelligent agents can smoothly share multiple types of data and reduce execution bias from the source [4].

The traditional scheduling mode has shortcomings such as uneven distribution and delayed response. Integrating real-world data, the scheduling subject can flexibly divide tasks to break the limitation of text scheduling. In the face of complex scenarios, tasks can be adjusted in a timely manner, resources can be allocated reasonably, and the dynamic collaboration mechanism is more in line with reality. Practical deployments include in home and park operation and maintenance scenarios [5].

Traditional intelligent agents rely on manual updating of knowledge bases, resulting in low efficiency in autonomous learning. Multimodal technology integrates real-life images and interactive data to build a learning library, allowing agents to accumulate experience and optimize logic in collaboration without the need to reconstruct models, and adapt to the development of small and medium-sized clusters. These technological links are interrelated and progressive, forming a complete empowerment chain that not only solves the single point shortcomings of traditional intelligent agents, but also amplifies overall efficiency through collaborative linkage, laying a solid technological foundation for subsequent scenarios.

This technical framework is not a simple module assembly, but a complete chain of perception, integration, collaboration, and learning, taking into account the needs of high-precision research and development and lightweight deployment, reserving sufficient optimization space for subsequent technological iterations and functional expansion, and supporting the continuous upgrading of the system.

# 4. MAINSTREAM PRACTICAL APPLICATION SCENARIOS OF LARGE LANGUAGE MODELS AND MULTI-AGENT SYSTEMS UNDER MULTIMODAL EMPOWERMENT

After the gradual maturity of core technologies such as perception and integration, a system equipped with multimodal capabilities has been piloted in the fields of people's livelihood, industry, education, and governance. Simplification of administrative procedures for livelihood services, the reduction of operation and maintenance on the industry side, the emphasis on personalization on the education side, and the improvement of accuracy on the governance side have achieved significant practical results, accumulating practical experience for large-scale promotion. The scenarios are all in line with reality and have no fictional content [6].

Livelihood convenience services is the most widely deployed field, focusing on scenarios such as smart government and home services. The system can synchronously receive written inquiries, voice appeals, and pictures of service materials from the public. It is divided into functional sections to complete problem solving, material preliminary review, and process guidance, adapting to the usage habits of different groups such as the elderly and young people, simplifying offline service processes. The Guizhou e-government platform has launched pilot services and received good user feedback.

The industrial applications are concentrated in light industry manufacturing, equipment operation and maintenance, production coordination and other scenarios. Small and medium-sized factories collect production line images through visual equipment, combine production data and equipment abnormality prompts, complete inspections, simple fault diagnosis, and rhythm adjustment. It is serving as a human-assisted tool and does not replace frontline personnel. It has been implemented in the electronics and daily necessities manufacturing industries, effectively reducing operation and maintenance omissions.

In the field of smart education, it is commonly used for online tutoring, classroom assistance, and learning situation analysis. Intelligent agents can recognize written assignments, oral answers, and handwritten content, complete exercise explanations, knowledge point sorting, and learning planning, and only serve as post-class assistance without disrupting the pace of offline teaching. Many primary and secondary schools in Shanghai have well-established applications that are suitable for personalized learning needs [7].

In urban grassroots governance, the system is used for park inspection, order guidance, and environmental monitoring. By collecting on-site images, environmental sounds, and monitoring data, daily inspections, abnormal reporting, and basic data organization are completed to improve the level of governance refinement. The Beijing Economic and Technological Development Zone has deployed relevant projects with stable operational effects.

All scenarios are in line with industry reality, with no fictional cases. The technical adaptability and practicality of the scenarios have been preliminarily verified, laying the foundation for further vertical field applications. Although these scenarios belong to different fields, they all reflect the common characteristics of multimodal empowerment: fitting people's livelihood and industrial reality, balancing practicality and convenience, focusing on lightweight application modes, and not making excessive technological stacking. Pilot deployments have yielded actionable insights, providing replicable models for large-scale deployment.

# 5. THE CURRENT CHALLENGES IN THE DEVELOPMENT OF MULTIMODAL EMPOWERMENT OF LARGE LANGUAGE MODELS AND MULTI-AGENT SYSTEMS

Although there has been progress in technology implementation and scenario application, the integration of the two types of technologies still faces common industry challenges. Various problems are intertwined and mutually constrained, slowing down the pace of pilot to scale promotion and becoming a key constraint for industrial upgrading.

Cross-modal information alignment is the most prominent problem given its complexity and cost and adaptation cost. There are significant differences in the dimensions of information such as text, images, and voice, and existing technologies can only adapt to conventional general scenarios. When faced with professional data such as medical imaging, industrial drawings, and industry-specific speech, interpretive errors and information distortion often occur and information distortion [8]. Improving the accuracy of professional scenarios requires specialized data training and model optimization, which directly increases the cost of construction and operation. Small and medium-sized R&D entities find it difficult to afford, resulting in the system being limited to general scenarios and difficult to delve into vertical fields.

The industry lacks unified collaborative standards, and the architecture and interaction rules are not yet unified. The self-developed frameworks and interface protocols of various institutions are different, making it difficult for intelligent agent products to interconnect with each other. After connecting to multimodal devices, the data collection and transmission formats become more chaotic, hindering industry standardization and easily causing repeated research and development, resource waste, and increasing the difficulty of cross platform docking.

The risk of cognitive bias and error propagation is prominent, with model errors and collection errors overlapping each other. The inherent output deviation of large models and the presence of raw errors in real-world data make it easy for a single agent to misjudge and quickly spread through cluster interaction, leading to systemic execution errors. There is currently no mature error correction mechanism in the industry, which limits the application of high-precision scenarios.

There are obvious shortcomings in data security and privacy protection, and the risk of multimodal data leakage is high. The pilot projects often use basic encryption without establishing a complete storage, permission, and traceability system. User images, voice, and personal information are prone to leakage, and there are also hidden dangers of illegal circulation of urban environmental data, which has become a core obstacle to the large-scale promotion of livelihood scenarios. These challenges are interconnected, creating a cascading effect that hinders industrial scalability and high-quality growth. High costs limit the widespread adoption of technology, the lack of standards exacerbates fragmentation, deviation risks restrict high-precision applications, and privacy shortcomings hinder the penetration of livelihood. Overall, these problems stem from immature technology and are also affected by imperfect industry ecology and compliance systems, which are comprehensive challenges that restrict the scale and high-quality development of the industry and urgently require systematic solutions [9].

# 6. OPTIMIZATION STRATEGIES AND DEVELOPMENT IDEAS FOR PROMOTING THE BENIGN DEVELOPMENT OF MULTIMODAL EMPOWERING MULTI-AGENT SYSTEMS

Based on the above difficulties, industry pace, and implementation needs, feasible measures are proposed from four aspects: technology, standards, risk control, and safety, taking into account both innovation and implementation, to promote the healthy development of the system.

The technology side can promote scenario-based optimization, accurately control costs and increase efficiency. Priority should be given to adopting lightweight fusion solutions for livelihood scenarios, relying on open-source models to quickly build and ensure stability while reducing costs; Establish specialized teams in high-precision fields such as industry and healthcare, and conduct semantic adaptation training based on industry data to improve conversion accuracy. Simultaneously develop simple docking components, simplifying the adaptation process between large models and perception devices, lowering the threshold for small and medium-sized teams, and promoting technological advancement [10].

Promote the development of unified industry standards and break down compatibility barriers. Industry associations, research institutes and leading enterprises jointly develop unified standards for intelligent agent architecture, data collection, and information exchange, clarify functional division, task scheduling, and upgrade specifications, open up basic technological achievements, build shared channels, promote industry collaboration and iteration, and avoid resource consumption [11].

Establish a multi-level risk verification mechanism to block the transmission of errors. Add an independent verification module to verify the authenticity of collected data and model output; Implement multi-agent joint review for important decisions to avoid single errors; Continuously optimize the training dataset, strengthen cognition with real cases, reduce the probability of deviation from the source, and build a strong risk control defense line.

Improve the data security control system and strictly adhere to the bottom line of privacy. Strictly follow regulatory requirements, classify and grade multimodal data, clarify permissions, storage, and transmission boundaries; Prioritize local processing of privacy data in livelihood scenarios to reduce the risk of cloud leakage; Build a full traceability system, record data flow, prevent illegal collection and abuse, and ensure compliant operation. These measures are not single measures, but create a coordinated framework of cost control, standard formulation, risk management and data security, taking into account both short-term implementation and long-term ecological construction. It not only responds to the existing core difficulties, but also conforms to the development trend of lightweight and standardization in the industry, focusing on phased promotion, pilot first, and gradual promotion, avoiding one size fits all, ensuring that countermeasures are adapted to different entities, and have practical and replicable value.

## 7. CONCLUSION

This article comprehensively analyzes the integration logic, technical path, application status, existing problems, and optimization directions of two types of technologies. The research is logically coherent, in line with the core theme, and has systematic and practical significance.

Research and clarify the internal mechanism of the integration of two types of technologies, sort out the four core paths of perception innovation, semantic integration, collaborative control, and long-term learning, sort out the pilot situation in the fields of people's livelihood, industry, education, and governance, objectively present the application status, accurately summarize the four core difficulties of high technology cost, lack of standards, deviation transmission, and privacy leakage, and propose targeted optimization solutions.

Multimodal fusion effectively fills the gaps in traditional multi-agent perception, interaction, and collaboration, expands application boundaries, and promotes cognitive intelligence from text deduction to real-world implementation. At present, the system is still in the pilot stage, with incomplete core technologies, supporting mechanisms, and operational models, making it difficult to promote across the entire industry in the short term.

The continuous iteration of domestic large-scale models, gradual implementation of industry standards, and increasingly mature lightweight solutions will continue to deepen the integration of the two types of technologies. In the future, we can deepen the segmentation of scenarios, optimize

security mechanisms, improve collaborative models, and promote the expansion of technology to industrial, medical, scientific research and other fields. This study emphasizes theoretical and current situation analysis, and can be supplemented with empirical verification, quantitative analysis, and refined optimization plans in the future, support the high-quality growth of China's general artificial intelligence industry.

# REFERENCES


[1] Jiang, B., Xie, Y., Wang, X., Su, W. J., Taylor, C. J., & Mallick, T. (2024, July). Multi-modal and multi-agent systems meet rationality: A survey. In ICML 2024 Workshop on LLMs and Cognition.

[2] Han, S., Zhang, Q., Jin, W., & Xu, Z. (2024). LLM multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578.

[3] Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., ... & Gao, J. (2025). Magma: A foundation model for multimodal ai agents. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 14203–14214).

[4] Zhao, X., Li, M., Weber, C., Hafez, M. B., & Wermter, S. (2023, October). Chat with the environment: Interactive multimodal perception using large language models. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3590–3596). IEEE.

[5] Xu, C., Tang, Z., Yu, H., Zeng, P., & Kong, L. (2023). Digital twin-driven collaborative scheduling for heterogeneous task and edge-end resource via multi-agent deep reinforcement learning. IEEE Journal on Selected Areas in Communications, 41(10), 3056–3069. https://doi.org/10.1109/JSAC.2023.3310965

[6] Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A comprehensive review of multimodal large language models: Performance and challenges across different tasks. arXiv preprint arXiv:2408.01319.

[7] Hariyanto, Kristianingsih, F. X. D., & Maharani, R. (2025). Artificial intelligence in adaptive education: a systematic review of techniques for personalized learning. Discover Education, 4(1), 458. https://doi.org/10.3390/discovereduc4010458

[8] AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M. A., Damseh, R., & Sheikh, J. (2024). Multimodal large language models in health care: applications, challenges, and future outlook. Journal of Medical Internet Research, 26, e59505. https://doi.org/10.2196/59505

[9] Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... & Feng, H. (2024). Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. Computers, Materials & Continua, 80(2). https://doi.org/10.32604/cmc.2024.047241

[10] Jin, Y., Li, J., Gu, T., Liu, Y., Zhao, B., Lai, J., ... & Ma, L. (2025). Efficient multimodal large language models: A survey. Visual Intelligence, 3(1), 27. https://doi.org/10.1007/s44267-024-00027-6

[11] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.