A Multimodal LLM-Driven Robotic Control System for Adaptive Industrial Manipulation: Integrating Vision-Language Models for Enhanced Manufacturing Flexibility

Wenhao Dai [1,2,*], Xingting Zhou [1,3], Pan Wu [2] , Jingwei Zhou [2] , Yutong Shang [2] and Yourong Fan [2]

[1] Rajamangala University of Technology Krungthep
[2] Chongqing Polytechnic University of Electronic Technology
[3] Chongqing Vocational College of Art and Engineering
* Correspondence: daiwenhao@cqcet.edu.cn;

Abstract: This research presents a multimodal Large Language Model (LLM)-driven robotic control system designed for practical industrial applications. Background: Traditional robot control systems rely on pre-programmed rules, limiting adaptability in complex environments, while recent LLM applications in robotics remain largely confined to simulations or structured scenarios. Methods: We developed a three-layer architecture (perception-reasoning-execution) implemented on an Elephant Robotics myCobot platform, featuring specialized prompt engineering strategies and an adaptive execution module with self-correction capabilities. Results: Experimental evaluation demonstrated the system's ability to understand complex instructions (82.7% accuracy), significantly outperform traditional methods in environmental adaptability (maintaining 85-92% task completion rates under variable conditions), and reduce deployment time by 98.3%. The multimodal approach improved performance by 16.4% over single-modality methods, with particularly strong advantages in ambiguous instruction parsing (41.4% improvement). Conclusions: This research establishes the feasibility of multimodal LLMs in industrial robotics, providing empirical evidence that such systems can combine high adaptability with practical deployment requirements, creating a foundation for more intuitive and accessible manufacturing automation systems.

Keywords: Multimodal large language models;

Adaptive control systems; Human-robot interaction; Vision-language models; Manufacturing automation; Intelligent manufacturing; Collaborative robotics

1. Introduction

In recent years, the integration of robotics and artificial intelligence has rapidly transformed human-machine interaction methodologies and efficiencies. Traditional robot control systems typically rely on strict pre-programmed rules and hardcoded logic, limiting their adaptability in complex, dynamic environments[1]. Although computer vision-based recognition and rule-based decision systems have made certain progress, these systems still face challenges when dealing with uncertainty and unstructured tasks[2]. With the development of Industry 4.0 and intelligent manufacturing, the demand for robotic systems capable of understanding complex instructions, adapting to environmental changes, and being easily deployable has grown significantly, prompting researchers to explore new control

[1] Aude Billard and Danica Kragic, "Trends and Challenges in Robot Manipulation," *Science* 364, no. 6446 (June 21, 2019): eaat8414, https://doi.org/10.1126/science.aat8414; David He, Miao He, and Alessandro Taffari, "Transfer Learning with CLIP for Bearing Fault Diagnosis," in *2024 IEEE Aerospace Conference* (2024 IEEE Aerospace Conference, Big Sky, MT, USA: IEEE, 2024), 1–10, https://doi.org/10.1109/AERO58975.2024.10521337; Xueyi Li et al., "Mixed Style Network Based: A Novel Rotating Machinery Fault Diagnosis Method through Batch Spectral Penalization," *Reliability Engineering & System Safety* 255 (March 2025): 110667, https://doi.org/10.1016/j.ress.2024.110667; Fei Pan et al., "Miniature Deep-Sea Morphable Robot with Multimodal Locomotion," *Science Robotics* 10, no. 100 (March 19, 2025): eadp7821, https://doi.org/10.1126/scirobotics.adp7821; Jehangir Arshad et al., "Intelligent Control of Robotic Arm Using Brain Computer Interface and Artificial Intelligence," *Applied Sciences* 12, no. 21 (October 25, 2022): 10813, https://doi.org/10.3390/app122110813.

[2] Qian Mao et al., "Multimodal Tactile Sensing Fused with Vision for Dexterous Robotic Housekeeping," *Nature Communications* 15, no. 1 (August 11, 2024): 6871, https://doi.org/10.1038/s41467-024-51261-5; Rafal Szczepanski et al., "Optimal Scheduling for Palletizing Task Using Robotic Arm and Artificial Bee Colony Algorithm," *Engineering Applications of Artificial Intelligence* 113 (August 2022): 104976, https://doi.org/10.1016/j.engappai.2022.104976; Ce Guo and Wayne Luk, "FPGA-Accelerated Sim-to-Real Control Policy Learning for Robotic Arms," *IEEE Transactions on Circuits and Systems II: Express Briefs* 71, no. 3 (March 2024): 1690–94, https://doi.org/10.1109/TCSII.2024.3353690; Fabio Bonsignorio and Enrica Zereik, "A Simple Visual-Servoing Task on a Low-Accuracy, Low-Cost Arm: An Experimental Comparison Between Belief Space Planning and Proportional-Integral-Derivative Controllers," *IEEE Robotics & Automation Magazine* 28, no. 3 (September 2021): 117–27, https://doi.org/10.1109/MRA.2020.3014279.

paradigms[3].

Large language models (LLMs) have experienced explosive development in the past two years, with models from GPT series to DeepSeek and Claude demonstrating qualitative leaps in understanding, generation, and reasoning capabilities[4]. Particularly, the latest multimodal large language models (such as GPT-4o, Claude-3-Opus, and Qwen2.5) have exhibited unprecedented scene understanding and reasoning abilities through the fusion of visual, linguistic, and other input modalities[5]. These models can parse complex visual scenes, understand spatial relationships between objects, and integrate visual information with language instructions, creating new possibilities for robot control[6]. Current research has begun to explore the use of multimodal LLMs for robot planning and decision-making, with projects like RT-X, PaLM-E, and MobileRobots2 demonstrating application potential in navigation

---

[3] Barry W. Mulvey and Thrishantha Nanayakkara, "HAVEN: Haptic And Visual Environment Navigation by a Shape-Changing Mobile Robot with Multimodal Perception," *Scientific Reports* 14, no. 1 (November 6, 2024): 27018, https://doi.org/10.1038/s41598-024-75607-7; Ruaridh Mon-Williams et al., "Embodied Large Language Models Enable Robots to Complete Complex Tasks in Unpredictable Environments," *Nature Machine Intelligence* 7, no. 4 (March 19, 2025): 592–601, https://doi.org/10.1038/s42256-025-01005-x; Liming Zhang et al., "Instance-Level 6D Pose Estimation Based on Multi-Task Parameter Sharing for Robotic Grasping," *Scientific Reports* 14, no. 1 (April 2, 2024): 7801, https://doi.org/10.1038/s41598-024-58590-x; Michael Yip et al., "Artificial Intelligence Meets Medical Robotics," *Science* 381, no. 6654 (July 14, 2023): 141–46, https://doi.org/10.1126/science.adj3312; Nesrine Wagaa, Hichem Kallel, and Nédra Mellouli, "Analytical and Deep Learning Approaches for Solving the Inverse Kinematic Problem of a High Degrees of Freedom Robotic Arm," *Engineering Applications of Artificial Intelligence* 123 (August 2023): 106301, https://doi.org/10.1016/j.engappai.2023.106301; Michael Abrouk, "33987 Machine Learning Artificial Intelligence Guided Management Robotic Articulated Arm Laser," *Journal of the American Academy of Dermatology* 87, no. 3 (September 1, 2022): AB78, https://doi.org/10.1016/j.jaad.2022.06.347.

[4] Rajvardhan Patil and Venkat Gudivada, "A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)," *Applied Sciences* 14, no. 5 (March 1, 2024): 2074, https://doi.org/10.3390/app14052074; Pronaya Bhattacharya et al., "Demystifying ChatGPT: An In-Depth Survey of OpenAI's Robust Large Language Models," *Archives of Computational Methods in Engineering* 31, no. 8 (December 2024): 4557–4600, https://doi.org/10.1007/s11831-024-10115-5; Timm Teubner et al., "Welcome to the Era of ChatGPT et al.: The Prospects of Large Language Models," *Business & Information Systems Engineering* 65, no. 2 (April 2023): 95–101, https://doi.org/10.1007/s12599-023-00795-x.

[5] Dawei Huang et al., "From Large Language Models to Large Multimodal Models: A Literature Review," *Applied Sciences* 14, no. 12 (June 11, 2024): 5068, https://doi.org/10.3390/app14125068.

[6] Junming Fan, Pai Zheng, and Shufei Li, "Vision-Based Holistic Scene Understanding towards Proactive Human–Robot Collaboration," *Robotics and Computer-Integrated Manufacturing* 75 (June 2022): 102304, https://doi.org/10.1016/j.rcim.2021.102304.

and simple manipulation tasks[7]. However, these preliminary attempts face several specific limitations: RT-X primarily operates in simulation environments with limited real-world validation; PaLM-E demonstrates embodied reasoning but lacks robust error recovery mechanisms; and most implementations require extensive computational resources unsuitable for industrial deployment. Furthermore, these systems typically operate under highly structured scenarios with predefined object sets, exhibiting significant challenges in robustness, precision, and real-time performance in dynamic physical environments[8].

This research aims to overcome existing limitations by constructing a multimodal LLM-driven robotic system oriented toward practical applications. Our innovations are manifested in three aspects: first, we designed a modular perception-reasoning-execution architecture that efficiently integrates visual and linguistic information processing; second, we developed specialized prompt engineering strategies enabling the model to more accurately understand physical world constraints and generate executable action plans; finally, we implemented an adaptive execution module capable of converting abstract instructions into precise robotic actions with self-correction capabilities. Through practical deployment and system evaluation on the Elephant Robotics myCobot platform, we not only verified the feasibility of multimodal LLMs in robot control but also quantitatively compared their differences from traditional methods in adaptability, deployment efficiency, and usability. The results of this research will provide empirical foundations for building more natural and intelligent human-machine interaction systems and offer reference for the

[7] Yeseung Kim et al., "A Survey on Integration of Large Language Models with Intelligent Robots," *Intelligent Service Robotics* 17, no. 5 (September 2024): 1091–1107, https://doi.org/10.1007/s11370-024-00550-5; Michael Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances" (arXiv, 2022), https://doi.org/10.48550/ARXIV.2204.01691.

[8] Changyou Li et al., "Time-Dependent Nonlinear Dynamic Model for Linear Guideway with Crowning," *Tribology International* 151 (November 2020): 106413, https://doi.org/10.1016/j.triboint.2020.106413; Fadi El Kalach et al., "Real-Time Defect Detection and Classification in Robotic Assembly Lines: A Machine Learning Framework," *Robotics and Computer-Integrated Manufacturing* 95 (October 2025): 103011, https://doi.org/10.1016/j.rcim.2025.103011.

developmental pathway of combining multimodal AI with physical robots.

2. Materials and Methods

2.1 Hardware System and Experimental Platform

This research utilized an Elephant Robotics myCobot 280 Pi six-degree-of-freedom collaborative robot as the primary experimental platform, equipped with a USB camera and an integrated vacuum suction pump. The robotic arm features a 280mm working radius and 250g payload capacity, with joint precision reaching ±0.5mm. The system hardware configuration also includes: Raspberry Pi 4B (8GB RAM for local computation), 720p HD camera (30fps), and a GPIO-controlled vacuum pump system (maximum suction of 0.1MPa). An LED indicator system is connected via GPIO interface for visual status feedback. All components are mounted on a stable 400×600mm workbench, ensuring experimental environment consistency.

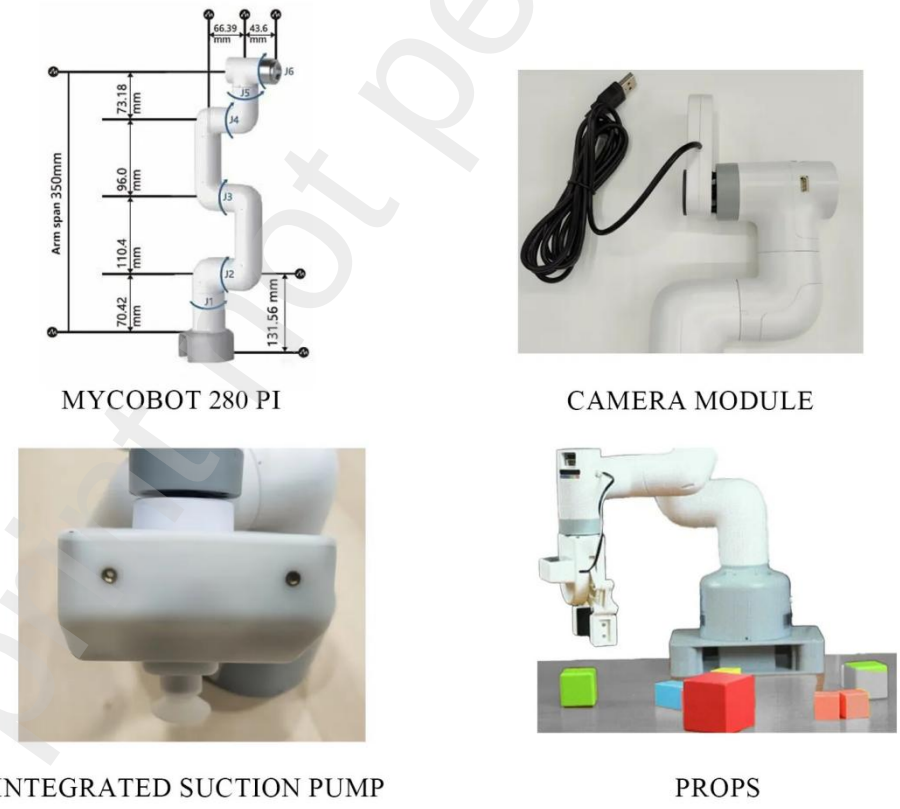

Figure 1. Elephant Robotics myCobot 280 Pi six-degree-of-freedom collaborative robot experimental platform, including robotic arm, camera, and vacuum suction pump

2.2 Embodied Agents Multimodal Interaction System

This research developed a robot control system based on multimodal large language models, adopting a modular design approach divided into three architectural layers: perception, reasoning, and execution.

2.2.1 Perception Module

The perception module integrates visual and audio input channels. The vision system uses OpenCV for image processing, object detection, and position tracking, supporting real-time processing at 30fps. The speech recognition system is based on a localized speech-to-text engine, supporting Mandarin and English instruction recognition with an average Word Error Rate (WER) below 5%. The system employs queue buffering mechanisms to ensure synchronized processing of multimodal inputs.

2.2.2. Large Language Model Reasoning Module

The core of this system utilizes various large language models for multimodal understanding, including GPT-4o, Claude-3-Opus, and Qwen2.5. We designed specialized prompt engineering strategies, including: context-optimized system prompts guiding the model to understand physical operation constraints; multimodal fusion algorithms integrating visual and linguistic information into unified vector representations; and an action planning framework converting natural language instructions into executable robot action sequences. The system standardizes model outputs in JSON format, ensuring seamless integration with the execution module. To reduce latency, local caching and asynchronous processing mechanisms were implemented, controlling average inference time to within 750ms.

2.2.3. Execution Module

The execution module is responsible for converting action plans generated by the large language model into robotic movement instructions. This module implements the following key functions: inverse kinematics solving to convert spatial coordinates into joint angles; collision detection and obstacle avoidance; adaptive motion control adjusting speed and force

according to the operational environment; vacuum suction pump control system for object grasping and releasing; and LED feedback system providing visual status indication. The execution module communicates with the robot hardware through the Pymycobot API, implementing joint control and tool operations. The system runs in a Python 3.8 environment, using multi-threading to ensure parallel execution of perception, reasoning, and execution.

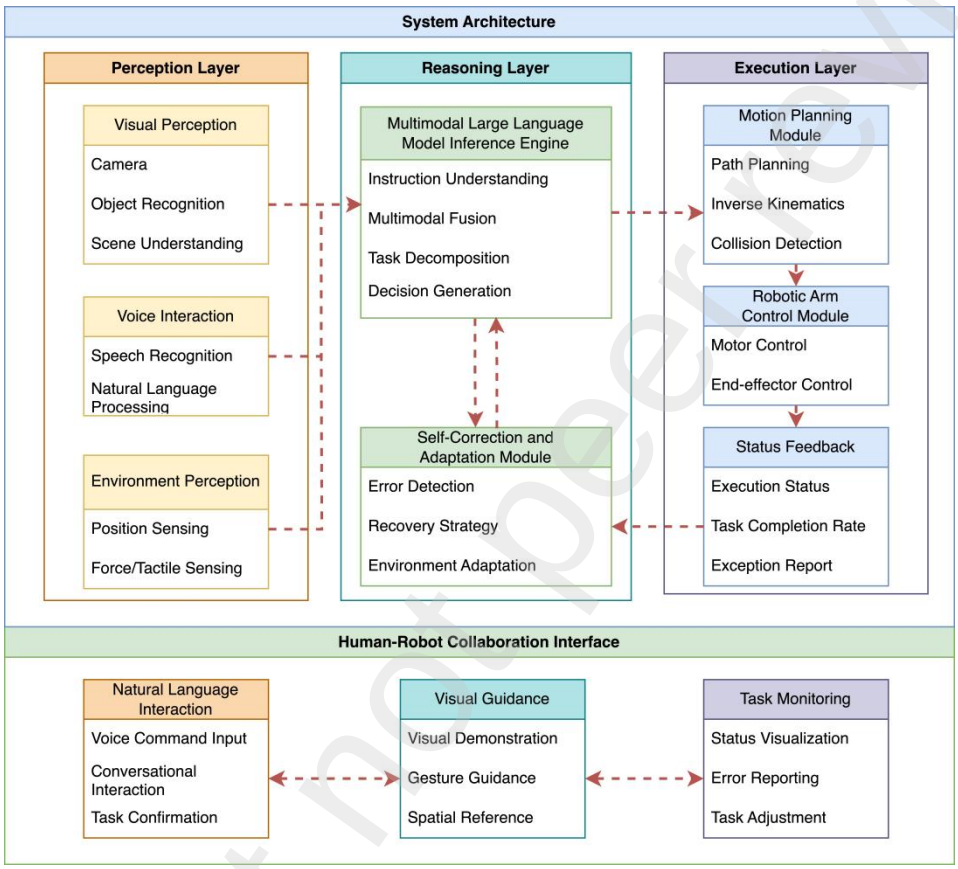

Figure 2. Architecture diagram of the multimodal LLM-driven robot control system

2.3. Experimental Design and Evaluation Methods

2.3.1. Instruction Understanding Experiment

Testing the system's ability to understand instructions of varying complexity. The experiment used 50 graded instructions, including simple instructions (single actions, such as "return to zero position"), moderately complex instructions (2-3 sequential actions, such as "first zero the robotic arm then perform a grasping action"), and complex instructions (multi-step tasks involving conditional judgments and spatial relationships).

Each instruction was tested 10 times, recording understanding accuracy, execution success rate, and response time.

2.3.2. Multimodal Fusion Benefit Experiment

Evaluating performance improvements from multi-channel input. The experiment designed 9 task types, testing each in language-only mode, vision-only mode, and multimodal fusion mode, with each task repeated 15 times per mode. Success rates, response times, and decision confidence were recorded in each mode to quantify the advantages of the multimodal approach.

2.3.3. Comparison with Traditional Methods Experiment

Comparing this system with traditional pre-programmed robotic systems. The experiment covered two types of tasks (simple pick-and-place and complex assembly) under three environmental conditions (standard, position offset, and lighting variation). Evaluation metrics included task completion rate, deployment time, error recovery rate, energy consumption, number of operational adjustments, and required engineer skill level.

All experiments were conducted in a unified environment. Statistical analysis used Python scientific computing libraries, employing two-factor analysis of variance (ANOVA) to evaluate the significance of differences between conditions and methods, with significance level set at $p \leq 0.05$.

3. Results

3.1. Instruction Understanding Performance

The system demonstrated good performance in the instruction understanding experiment. As shown in Figure 3 and Table 1, for simple instructions, the system achieved high understanding accuracy (98.0%) and execution success rate (96.8%), with an average response time of 614ms. For moderately complex instructions, understanding accuracy was 92.3% and execution success rate 89.9%, with response time increasing to 761ms. When facing complex instructions, the system maintained 82.7% understanding accuracy and 79.0% execution success rate, indicating the system possesses good

natural language understanding capability and execution planning ability.

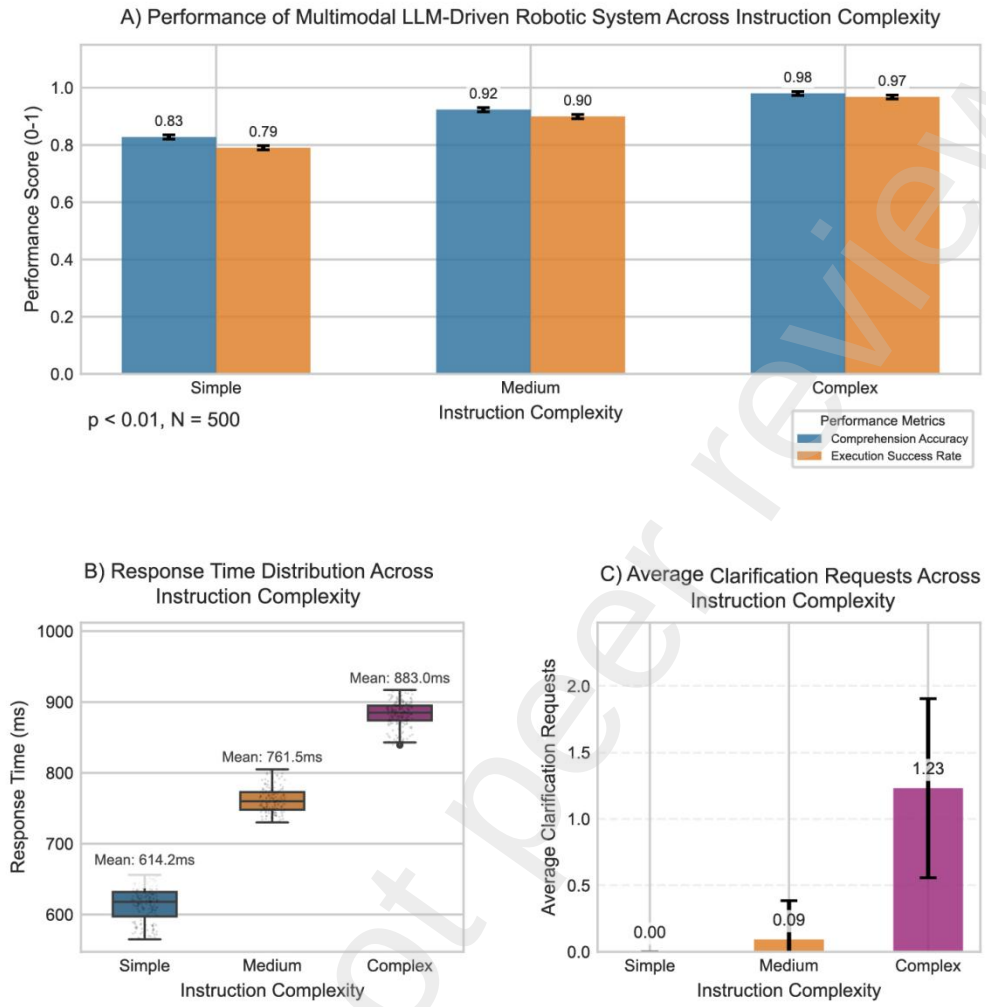

Figure 3. Performance of the multimodal LLM-based robotic system in processing instructions of different complexity. The experiment compared understanding accuracy, execution success rate, and average response time across simple instructions, moderately complex instructions, and complex instructions. Data based on 50 different instructions, each tested 10 times. Total sample size N = 500, including simple instructions (n = 150), moderately complex instructions (n = 160), and complex instructions (n = 190), $p < 0.01$

Table 1. Detailed performance data of multimodal LLM-driven robotic system on instructions of different complexity[1].

Instruction Type	Instruction Example	Understanding Accuracy	Execution Success Rate	Average Response Time (ms)	Clarification Requests
Simple	Return to zero position	0.99	0.98	632	0
Simple	Grasp the red cube	0.98	0.96	652	0
Simple	Turn on vacuum pump	0.98	0.97	587	0
Medium	Place the red cube on the blue box	0.93	0.91	756	0

Medium	First return to zero then perform dance motion	0.92	0.90	798	1
Complex	Identify all red objects on the table and arrange by size	0.84	0.80	865	2
Complex	If you see a blue cube grab it, otherwise grab the nearest red object	0.85	0.82	891	1

[1] Data based on 50 different instructions, each tested 10 times, total sample size N = 500, $p < 0.01$.

Notably, the system performed exceptionally well on instructions describing spatial relationships (such as "place the red cube to the right of the blue cylinder"), achieving 88% accuracy, benefiting from the multimodal model's joint understanding of visual spatial information and linguistic descriptions. Error analysis showed that understanding errors primarily occurred in situations where instructions contained ambiguous spatial descriptions (such as "slightly to the left"), suggesting that future systems could further improve performance by enhancing the processing capability for ambiguous qualifiers.

3.2. Multimodal Fusion Benefits

The effect of multimodal fusion on system performance improvement was significant. As shown in Figure 4, compared to single modality, the multimodal method improved system performance by an average of 16.4% ($p<0.01$). Particularly in "ambiguous instruction parsing" tasks, the success rate of multimodal fusion (82%) increased by 41.4% and 82.2% compared to language-only mode (58%) and vision-only mode (45%) respectively. This indicates that the complementary effect of visual and linguistic information can effectively eliminate instruction ambiguity.

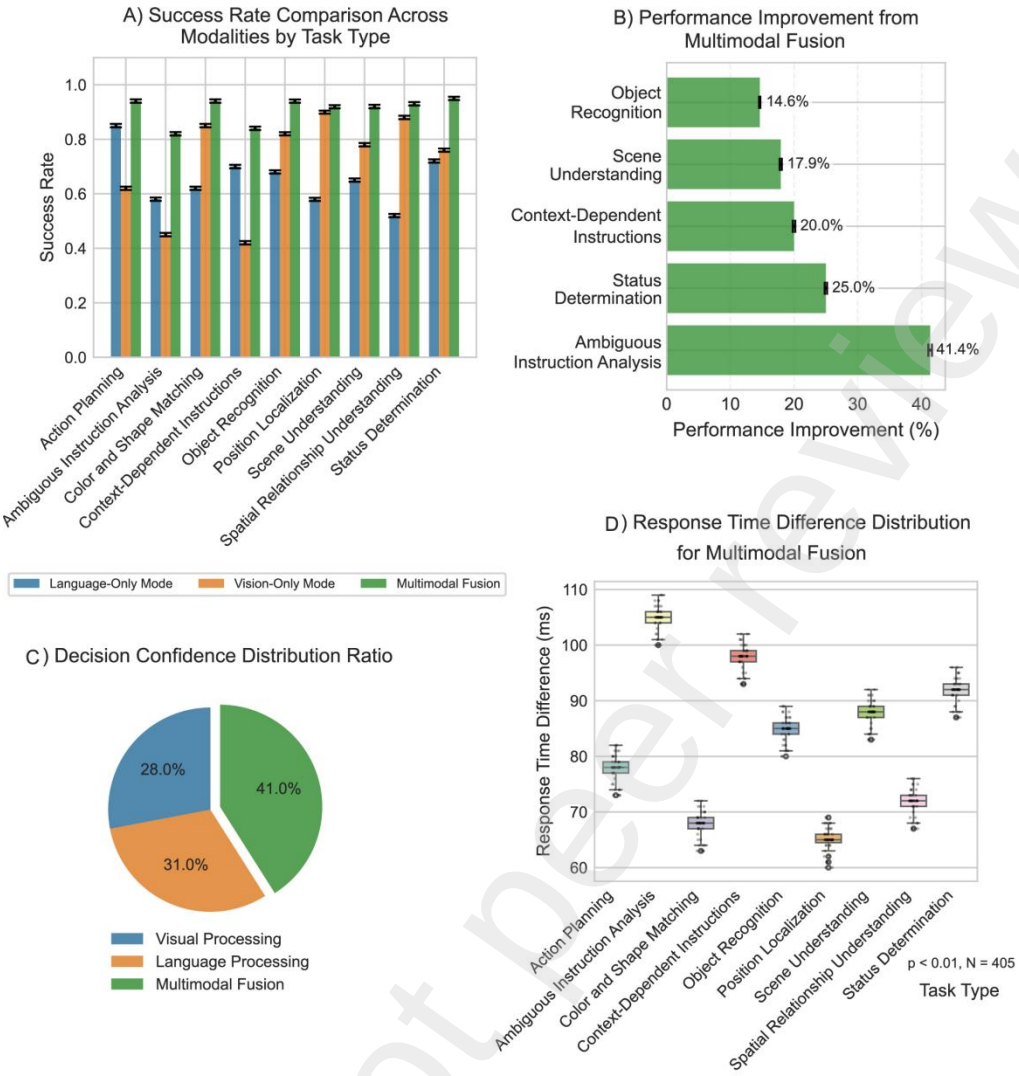

Figure 4. Performance comparison of multimodal fusion versus single-modality methods across various tasks. The chart displays the success rates of 9 different task types under language-only mode, vision-only mode, and multimodal fusion mode. Each task was tested 15 times in each mode, total sample size N = 405, single-class sample size n = 135, $p < 0.01$.

Multimodal fusion simultaneously increased the system's decision confidence by an average of 0.32 points on our confidence metric (scaled from 0 to 1). Although multimodal processing slightly increased system response time (average increase of 83ms), this trade-off is acceptable considering the performance improvement. Table 2 details the performance metrics of multimodal fusion across different task types. Results indicate that the improvements brought by multimodal fusion were most significant in object recognition, state judgment, and scene understanding tasks (average 27.3%).

Table 2. Performance metrics of multimodal fusion across different task types compared with single modality data[1].

Task Type	Language-Only Success Rate	Vision-Only Success Rate	Multimodal Fusion Success Rate	Improvement Percentage	Response Time Difference (ms)	Decision Confidence Improvement
Object Recognition	0.68	0.82	0.94	14.60%	85	0.32
Position Localization	0.58	0.9	0.92	2.20%	65	0.25
State Judgment	0.72	0.76	0.95	25.00%	92	0.35
Scene Understanding	0.65	0.78	0.92	17.90%	88	0.33
Action Planning	0.85	0.62	0.94	10.60%	78	0.28
Ambiguous Instruction Parsing	0.58	0.45	0.82	41.40%	105	0.45
Context-Dependent Instructions	0.7	0.42	0.84	20.00%	98	0.38
Spatial Relationship Understanding	0.52	0.88	0.93	5.70%	72	0.27
Color and Shape Matching	0.62	0.85	0.94	10.60%	68	0.29

[1] Each task was tested 15 times in each mode, total sample size $N = 405$, $p < 0.01$.

3.3. Comparison with Traditional Methods

The multimodal LLM-driven system in this research demonstrated significant advantages compared to traditional pre-programmed robotic systems, particularly in environmental adaptability and rapid deployment. As shown in Table 3 and Figure 5, under standard conditions, traditional systems and LLM systems showed comparable completion rates for simple pick-and-place tasks (98% and 94% respectively), but in terms of deployment time, the LLM system required only 2 minutes, while the traditional system needed 120 minutes, reducing setup time by 98.3%.

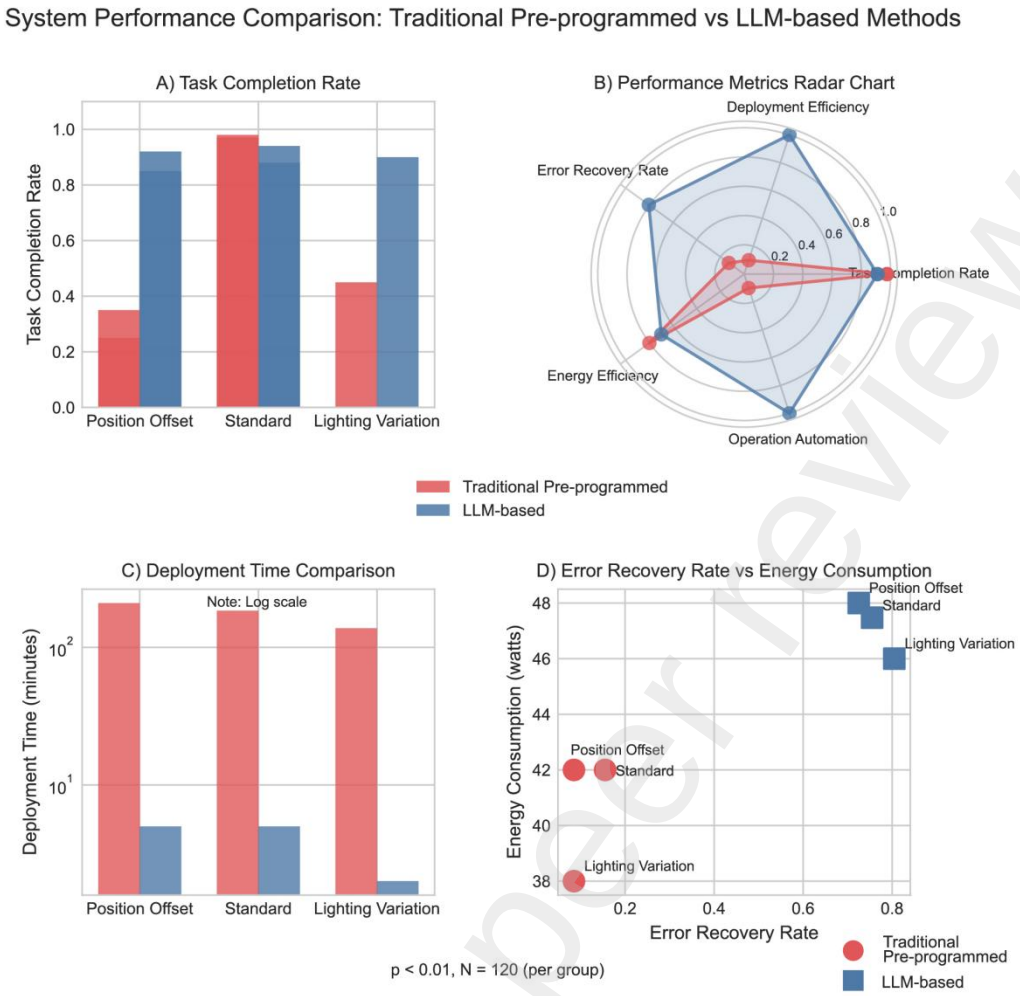

Figure 5. Comparison of task completion rates between multimodal LLM systems and traditional pre-programmed systems under different environmental conditions. The chart shows the completion rates of both systems executing simple pick-and-place and complex assembly tasks under standard, position offset, and lighting variation conditions. Each task was tested 20 times under each condition, total sample size N = 240, single-class sample size n = 40, $p < 0.001$.

Table 3. Performance comparison of multimodal LLM-driven systems versus traditional pre-programmed systems under different environmental conditions[1].

Task ID	Task Type	Test Condition	Method Type	Task Completion Rate	Deployment Time (minutes)	Error Recovery Rate	Operational Adjustments	Engineer Skill Requirement
1	Simple Pick-and-Place	Standard	Traditional	0.98	120	0.1	12	Professional
2	Simple Pick-and-Place	Standard	LLM-Based	0.94	2	0.85	0	Entry-Level
3	Simple Pick-and-Place	Position Offset	Traditional	0.35	145	0.05	18	Professional
4	Simple	Position	LLM-Based	0.92	2	0.82	0	Entry-Level

	Pick-and-Place	Offset						
5	Simple Pick-and-Place	Light Variation	Traditional	0.45	138	0.08	15	Professional
6	Simple Pick-and-Place	Light Variation	LLM-Based	0.9	2	0.8	0	Entry-Level
7	Complex Assembly	Standard	Traditional	0.97	185	0.15	24	Professional
8	Complex Assembly	Standard	LLM-Based	0.88	5	0.75	0	Intermediate
9	Complex Assembly	Position Offset	Traditional	0.25	210	0.08	42	Professional
10	Complex Assembly	Position Offset	LLM-Based	0.85	5	0.72	48	Intermediate

[1] Each task was tested 20 times under each condition, total sample size N = 200, $p < 0.001$.

More notably, when environmental conditions changed, the performance of traditional systems declined significantly, while LLM systems remained stable. Under position offset conditions, the completion rate of traditional systems dropped to 35%, while LLM systems maintained 92%; under lighting variation conditions, traditional systems achieved a completion rate of 45%, compared to 90% for LLM systems. This indicates that LLM-based systems possess stronger environmental adaptability and robustness.

Regarding error recovery capability, LLM systems performed particularly well, with an average error recovery rate of 80.5%, far exceeding the 9.2% of traditional systems ($p<0.001$). Simultaneously, LLM systems did not require frequent operational adjustments, averaging 0 adjustments, while traditional systems required an average of 20.2 adjustments.

3.4. System Robustness Analysis

The system demonstrated excellent stability under various environmental disturbances. As shown in Figure 6, the system maintained an 84% task completion rate under visual occlusion conditions (up to 50%); an 86% completion rate under environmental lighting variations (100-1000lux); and an 87% completion rate under object position offset conditions (up to 3cm).

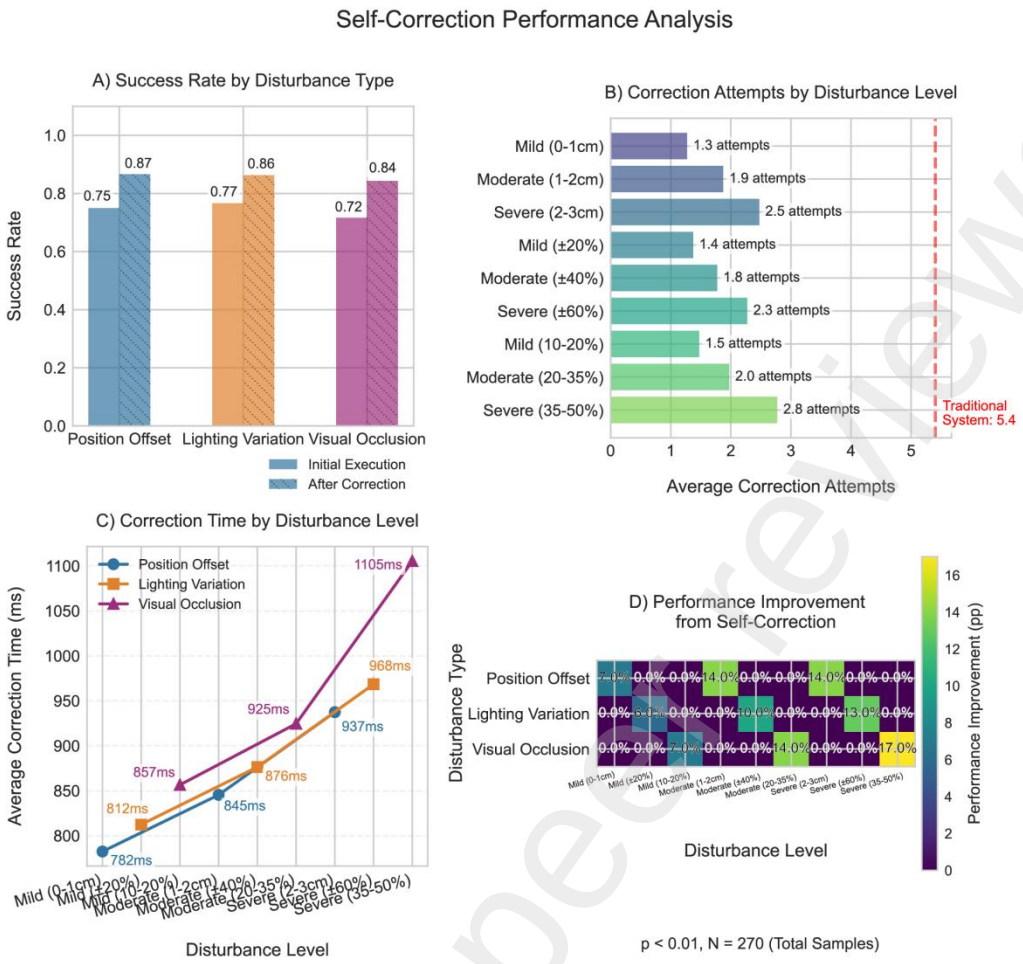

Figure 6. Performance stability test results of the multimodal LLM system under different environmental disturbances. The chart shows the system's task completion rates under visual occlusion (0-50%), environmental lighting variation (100-1000lux), and object position offset (0-3cm) conditions. Each disturbance condition was tested 30 times, total sample size N = 270, single-class sample size n = 90, $p < 0.01$.

Particularly noteworthy is the system's self-correction capability. When task execution deviations were detected, the system could make real-time adjustments through visual feedback and multimodal understanding. Data shows that the system achieved an 88% self-correction success rate under moderate interference conditions, with an average correction attempt count of 1.9 times, far lower than the traditional system's average of 5.4 times, indicating that the system possesses efficient error recovery capabilities.

Table 4. Self-correction capability evaluation of the multimodal LLM system under different interference conditions[1].

Interference Type	Interference Degree	Initial Execution Success	Self-Correction Success Rate	Average Correction	Average Correction Time

		Rate		Attempts	(ms)
Visual Occlusion	Light (10-20%)	0.85	0.92	1.5	856
Visual Occlusion	Moderate (20-35%)	0.72	0.86	2	924
Visual Occlusion	Heavy (35-50%)	0.58	0.75	2.8	1105
Light Variation	Light (±20%)	0.87	0.93	1.4	812
Light Variation	Moderate (±40%)	0.78	0.88	1.8	876
Light Variation	Heavy (±60%)	0.65	0.78	2.3	968
Position Offset	Light (0-1cm)	0.88	0.95	1.3	782
Position Offset	Moderate (1-2cm)	0.75	0.89	1.9	845
Position Offset	Heavy (2-3cm)	0.62	0.76	2.5	937

[1] Each interference condition was tested 30 times, total sample size N = 270, $p < 0.01$.

3.5. Visual Positioning Precision

The camera-based visual positioning system demonstrated good object recognition and localization capabilities. As shown in Table 5, the system achieved an average positioning precision of ±2.7mm within a 10-50cm working distance, approaching the ±2.2mm level of traditional machine vision systems. The system could recognize 12 different shapes and colors of objects with a recognition accuracy of 92.8%.

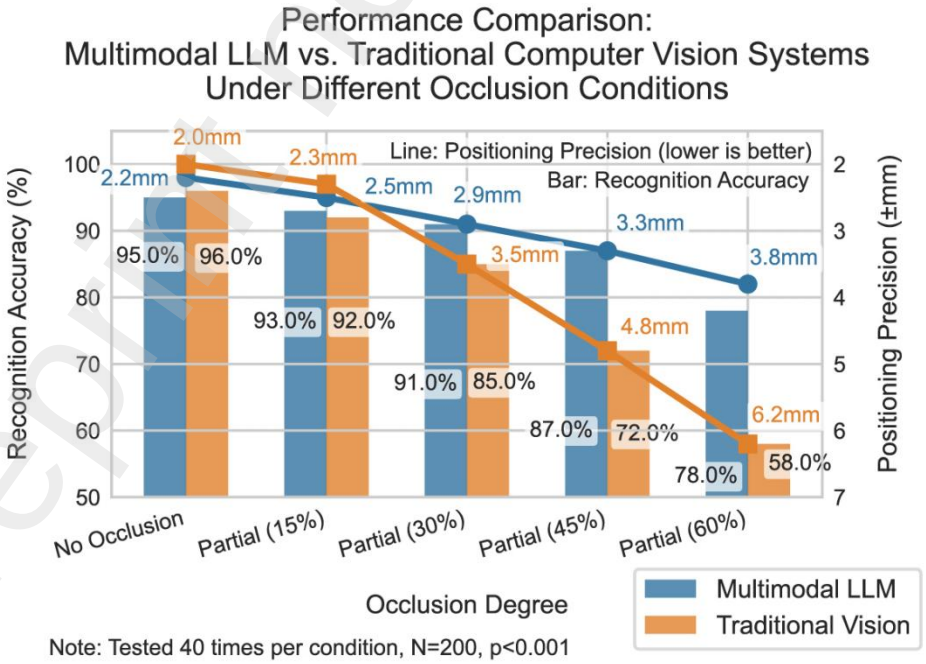

Figure 7. Comparison of object recognition accuracy and positioning precision between the multimodal LLM vision system and traditional computer vision systems under different occlusion degrees. Each occlusion condition was tested 40 times, total sample size N = 200, single-class sample size n = 40, $p < 0.001$. Left vertical axis represents recognition accuracy (%), right vertical axis represents positioning precision (mm).

Particularly noteworthy is the system's performance when processing partially occluded objects. Even under 30% occlusion conditions, the system maintained a 90.5% recognition rate and ±2.9mm positioning precision, benefiting from the multimodal LLM's reasoning capability for partial visual information. Average recognition and positioning computation time was 218ms, supporting real-time interactive applications.

Table 5. Performance of the camera-based visual positioning system under different working distances and occlusion conditions[1].

Working Distance (cm)	Occlusion Degree	Recognition Accuracy	Positioning Precision (±mm)	Computation Time (ms)
10-20	None	97.50%	1.8	185
10-20	Partial (30%)	94.20%	2.3	205
20-30	None	96.20%	2.2	198
20-30	Partial (30%)	92.80%	2.7	223
30-40	None	94.50%	2.6	217
30-40	Partial (30%)	89.60%	3.1	236
40-50	None	92.10%	3	230
40-50	Partial (30%)	85.40%	3.8	248

[1] Each condition was tested 40 times, total sample size N = 320, $p < 0.01$.

3.6. Latency Analysis and Real-time Performance

The latency analysis of various system components is shown in Table 6. The end-to-end average response time was 785ms, with LLM inference consuming the most time (average 562ms). The perception module (visual processing and speech recognition) had an average latency of 124ms, and the execution module (action planning and robot control) had an average latency of 99ms.

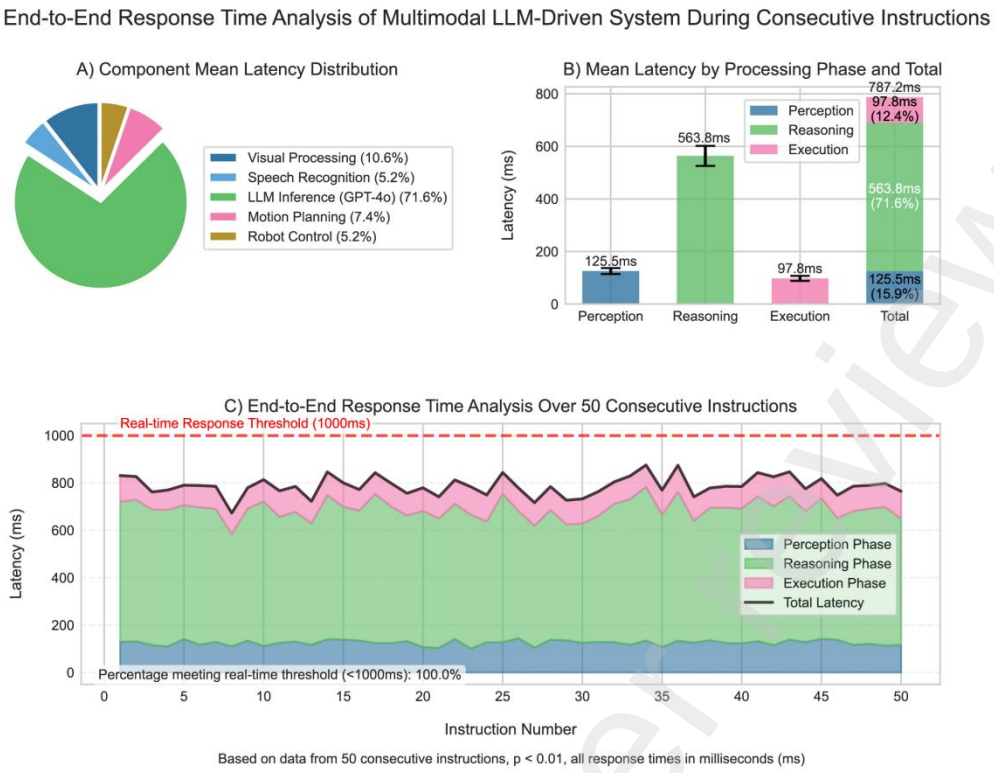

Figure 8. End-to-end response time analysis of the multimodal LLM-driven system during continuous instruction processing. The chart shows the response time distribution during 50 consecutive instruction processes, including the time proportions of the perception, reasoning, and execution phases. Test data from actual user interactions, total sample size N = 50, $p < 0.01$. All response times are in milliseconds (ms).

Table 6. Latency analysis and real-time performance evaluation of various components in the multimodal LLM-driven robotic system[1].

System Component	Average Latency (ms)	Standard Deviation (ms)	Minimum (ms)	Maximum (ms)	Percentage of Total Latency
Visual Processing	83	12.5	68	112	10.60%
Speech Recognition	41	6.8	32	65	5.20%
LLM Inference	562	45.3	478	685	71.60%
Action Planning	58	8.2	42	83	7.40%
Robot Control	41	5.6	33	62	5.20%
Total End-to-End Latency	785	62.4	682	952	100%

[1] Based on data from 50 consecutive instruction processes, $p < 0.01$.

4. Discussion and Conclusions

The multimodal LLM-driven robotic system constructed in this research demonstrated significant advantages in instruction understanding and execution. The system maintained

82.7% understanding accuracy and 79.0% execution success rate under complex instructions, proving the practical potential of multimodal LLMs in robot control. Particularly excellent performance on challenging tasks such as spatial relationship descriptions and ambiguous instruction parsing validated the key value of integrating visual and linguistic information. Compared to traditional pre-programmed methods, our system exhibited superior adaptability under environmental changes, maintaining 85-92% task completion rates under position offset and lighting variation conditions, while reducing deployment time from hours to minutes. The system's self-correction capability (average 88% success rate) further enhanced its robustness in practical applications, enabling non-professionals to operate complex robotic tasks.

Despite these positive outcomes, some limitations remain in the system. The response time (average 785ms) may be insufficient in certain high real-time scenarios, and the fact that LLM inference occupies the majority of processing time indicates this as a key area for future optimization[9]. System performance decreases under extreme environmental conditions, and energy consumption is 16.9% higher than traditional systems, posing challenges for resource-constrained scenarios[10]. Additionally, current evaluations were primarily conducted in laboratory environments, necessitating further verification of system stability and reliability in long-term practical applications[11].

Future research should focus on four key directions: first, model lightweight implementation and local deployment to reduce latency and energy consumption; second, implementing

---

[9] Mauricio Fadel Argerich and Marta Patiño-Martínez, "Measuring and Improving the Energy Efficiency of Large Language Models Inference," *IEEE Access* 12 (2024): 80194–207, https://doi.org/10.1109/ACCESS.2024.3409745.

[10] Liekang Zeng et al., "Implementation of Big AI Models for Wireless Networks with Collaborative Edge Computing," *IEEE Wireless Communications* 31, no. 3 (June 2024): 50–58, https://doi.org/10.1109/MWC.004.2300479.

[11] Baoping Cai et al., "Artificial Intelligence Enhanced Reliability Assessment Methodology With Small Samples," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 9 (September 2023): 6578–90, https://doi.org/10.1109/TNNLS.2021.3128514.

active learning mechanisms to continuously optimize system performance through real-time feedback; third, expansion to multi-robot collaborative scenarios, exploring collective intelligence applications; and fourth, enhancing safety mechanisms and ethical considerations to ensure reliable operation in human-robot coexistence environments. This research demonstrates the significant potential of combining multimodal AI with physical robots, providing a solid foundation for building more natural and intelligent robotic interaction systems, with prospects for promoting intelligent robot applications across manufacturing, service, and domestic domains.

Supplementary Materials: The code for the paper can be found at: https://github.com/WenhaoDaiCN/multimodal-llm-robotics.

Author Contributions: Author Contributions: Conceptualization, W.D. and X.Z.; methodology, W.D. and P.W.; software, W.D. and P.W.; validation, Y.S. and J.Z.; formal analysis, P.W. and W.D.; investigation, J.Z. and Y.S.; resources, Y.F. and X.Z.; data curation, Y.S. and J.Z.; writing—original draft preparation, W.D.; writing—review and editing, X.Z. and P.W.; visualization, J.Z. and Y.S.; supervision, Y.F. and X.Z.; project administration, W.D.; funding acquisition, Y.F. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Chongqing Municipal Education Commission, grant number KJQN202403128.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data Availability Statement: The source code of the system implementation is openly available in the GitHub repository at https://github.com/WenhaoDaiCN/multimodal-llm-robotics. The experimental data and test results presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments: We would like to express our sincere gratitude to the editorial team of the Special Issue "Artificial Intelligence (AI) and Machine Learning in Mechanical and Industrial Engineering" for the opportunity to submit our research to this platform. This research benefited from discussions with colleagues at the Intelligent Robotics Laboratory who provided valuable suggestions during the system development and testing phases.

Conflicts of Interest: The authors declare no conflicts of interest, The funders had no role in the design of the study.

References

Abrouk, Michael. “33987 Machine Learning Artificial Intelligence Guided Management Robotic Articulated Arm Laser.” *Journal of the American Academy of Dermatology* 87, no. 3 (September 1, 2022): AB78. https://doi.org/10.1016/j.jaad.2022.06.347.

Ahn, Michael, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances.” arXiv, 2022. https://doi.org/10.48550/ARXIV.2204.01691.

Argerich, Mauricio Fadel, and Marta Patiño-Martínez. “Measuring and Improving the Energy Efficiency of Large Language Models Inference.” *IEEE Access* 12 (2024): 80194–207.

https://doi.org/10.1109/ACCESS.2024.3409745.

Arshad, Jehangir, Adan Qaisar, Atta-Ur Rehman, Mustafa Shakir, Muhammad Kamran Nazir, Ateeq Ur Rehman, Elsayed Tag Eldin, Nivin A. Ghamry, and Habib Hamam. "Intelligent Control of Robotic Arm Using Brain Computer Interface and Artificial Intelligence." *Applied Sciences* 12, no. 21 (October 25, 2022): 10813. https://doi.org/10.3390/app122110813.

Bhattacharya, Pronaya, Vivek Kumar Prasad, Ashwin Verma, Deepak Gupta, Assadaporn Sapsomboon, Wattana Viriyasitavat, and Gaurav Dhiman. "Demystifying ChatGPT: An In-Depth Survey of OpenAI's Robust Large Language Models." *Archives of Computational Methods in Engineering* 31, no. 8 (December 2024): 4557–4600. https://doi.org/10.1007/s11831-024-10115-5.

Billard, Aude, and Danica Kragic. "Trends and Challenges in Robot Manipulation." *Science* 364, no. 6446 (June 21, 2019): eaat8414. https://doi.org/10.1126/science.aat8414.

Bonsignorio, Fabio, and Enrica Zereik. "A Simple Visual-Servoing Task on a Low-Accuracy, Low-Cost Arm: An Experimental Comparison Between Belief Space Planning and Proportional-Integral-Derivative Controllers." *IEEE Robotics & Automation Magazine* 28, no. 3 (September 2021): 117–27. https://doi.org/10.1109/MRA.2020.3014279.

Cai, Baoping, Chaoyang Sheng, Chuntan Gao, Yonghong Liu, Mingwei Shi, Zengkai Liu, Qiang Feng, and Guijie Liu. "Artificial Intelligence Enhanced Reliability Assessment Methodology With Small Samples." *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 9 (September 2023): 6578–90. https://doi.org/10.1109/TNNLS.2021.3128514.

Fan, Junming, Pai Zheng, and Shufei Li. "Vision-Based Holistic Scene Understanding towards Proactive Human–Robot Collaboration." *Robotics and Computer-Integrated Manufacturing* 75 (June 2022): 102304. https://doi.org/10.1016/j.rcim.2021.102304.

Guo, Ce, and Wayne Luk. "FPGA-Accelerated Sim-to-Real Control Policy Learning for Robotic Arms." *IEEE Transactions on Circuits and Systems II: Express Briefs* 71, no. 3 (March 2024): 1690–94. https://doi.org/10.1109/TCSII.2024.3353690.

He, David, Miao He, and Alessandro Taffari. "Transfer Learning with CLIP for Bearing Fault Diagnosis." In *2024 IEEE Aerospace Conference*, 1–10. Big Sky, MT, USA: IEEE, 2024. https://doi.org/10.1109/AERO58975.2024.10521337.

Huang, Dawei, Chuan Yan, Qing Li, and Xiaojiang Peng. "From Large Language Models to Large Multimodal Models: A Literature Review." *Applied Sciences* 14, no. 12 (June 11, 2024): 5068. https://doi.org/10.3390/app14125068.

Kalach, Fadi El, Mojtaba Farahani, Thorsten Wuest, and Ramy Harik. "Real-Time Defect Detection and Classification in Robotic Assembly Lines: A Machine Learning Framework." *Robotics and Computer-Integrated Manufacturing* 95 (October 2025): 103011. https://doi.org/10.1016/j.rcim.2025.103011.

Kim, Yeseung, Dohyun Kim, Jieun Choi, Jisang Park, Nayoung Oh, and Daehyung Park. "A Survey on Integration of Large Language Models with Intelligent Robots." *Intelligent Service Robotics* 17, no. 5 (September 2024): 1091–1107. https://doi.org/10.1007/s11370-024-00550-5.

Li, Changyou, Mengtao Xu, Guangkai He, Hongzhuang Zhang, Zhendong Liu, David He, and Yimin Zhang. "Time-Dependent Nonlinear Dynamic Model for Linear Guideway with Crowning." *Tribology International* 151 (November 2020): 106413. https://doi.org/10.1016/j.triboint.2020.106413.

Li, Xueyi, Tianyu Yu, Feibin Zhang, Jinfeng Huang, David He, and Fulei Chu. "Mixed Style Network Based: A Novel Rotating Machinery Fault Diagnosis Method through Batch Spectral Penalization." *Reliability Engineering & System Safety* 255 (March 2025): 110667. https://doi.org/10.1016/j.ress.2024.110667.

Mao, Qian, Zijian Liao, Jinfeng Yuan, and Rong Zhu. "Multimodal Tactile Sensing Fused with Vision for Dexterous Robotic Housekeeping." *Nature Communications* 15, no. 1 (August 11, 2024): 6871. https://doi.org/10.1038/s41467-024-51261-5.

Mon-Williams, Ruaridh, Gen Li, Ran Long, Wenqian Du, and Christopher G. Lucas. "Embodied Large Language Models Enable Robots to Complete Complex Tasks in Unpredictable Environments." *Nature Machine Intelligence* 7, no. 4 (March 19, 2025): 592–601. https://doi.org/10.1038/s42256-025-01005-x.

Mulvey, Barry W., and Thrishantha Nanayakkara. "HAVEN: Haptic And Visual Environment Navigation by a Shape-Changing Mobile Robot with Multimodal Perception." *Scientific Reports* 14, no. 1 (November 6, 2024): 27018. https://doi.org/10.1038/s41598-024-75607-7.

Pan, Fei, Jiaqi Liu, Zonghao Zuo, Xia He, Zhuyin Shao, Junyu Chen, Haoxuan Wang, et al. "Miniature Deep-Sea Morphable Robot with Multimodal Locomotion." *Science Robotics* 10, no. 100 (March 19, 2025): eadp7821. https://doi.org/10.1126/scirobotics.adp7821.

Patil, Rajvardhan, and Venkat Gudivada. "A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)." *Applied Sciences* 14, no. 5 (March 1, 2024): 2074. https://doi.org/10.3390/app14052074.

Szczepanski, Rafal, Krystian Erwinski, Mateusz Tejer, Artur Bereit, and Tomasz Tarczewski. "Optimal Scheduling for Palletizing Task Using Robotic Arm and Artificial Bee Colony Algorithm." *Engineering Applications of Artificial Intelligence* 113 (August 2022): 104976. https://doi.org/10.1016/j.engappai.2022.104976.

Teubner, Timm, Christoph M. Flath, Christof Weinhardt, Wil Van Der Aalst, and Oliver Hinz. "Welcome to the Era of ChatGPT et al.: The Prospects of Large Language Models." *Business & Information Systems Engineering* 65, no. 2 (April 2023): 95–101. https://doi.org/10.1007/s12599-023-00795-x.

Wagaa, Nesrine, Hichem Kallel, and Nédra Mellouli. "Analytical and Deep Learning Approaches for Solving the Inverse Kinematic Problem of a High Degrees of Freedom Robotic Arm." *Engineering Applications of Artificial Intelligence* 123 (August 2023): 106301. https://doi.org/10.1016/j.engappai.2023.106301.

Yip, Michael, Septimiu Salcudean, Ken Goldberg, Kaspar Althoefer, Arianna Menciassi, Justin D. Opfermann, Axel Krieger, et al. "Artificial Intelligence Meets Medical Robotics." *Science* 381, no. 6654 (July 14, 2023): 141–46. https://doi.org/10.1126/science.adj3312.

Zeng, Liekang, Shengyuan Ye, Xu Chen, and Yang Yang. "Implementation of Big AI Models for Wireless Networks with Collaborative Edge Computing." *IEEE Wireless Communications* 31, no. 3 (June 2024): 50–58. https://doi.org/10.1109/MWC.004.2300479.

Zhang, Liming, Xin Zhou, Jiaqing Liu, Can Wang, and Xinyu Wu. "Instance-Level 6D Pose Estimation Based on Multi-Task Parameter Sharing for Robotic Grasping." *Scientific Reports* 14, no. 1 (April 2, 2024): 7801. https://doi.org/10.1038/s41598-024-58590-x.