Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.128134
Published by: Zeus Press

LLM-Based Zero-Code Control for Robotic Arms: Representative Methods, Challenges, and Future Trends

Jiajun He*

School of Data Science and Artificial Intelligence (Aberdeen), South China Normal University, Guangzhou 528225, China

**Corresponding author: Jiajun He.*

Abstract

With the rapid evolution of Large Language Models (LLMs), significant advancements have emerged in natural language processing, cross-modal reasoning, and complex task planning, heralding a paradigm shift in robotic manipulation. Traditional robotic arm control demands specialized programming and extensive engineering expertise, creating a high barrier to entry. This paper systematically reviews recent advancements in LLM-based zero-code implementation technologies for robotic arms, which enable direct control through natural language and substantially reduce development costs and operational complexity. We establish a clear technical classification framework, analyzing the performance characteristics and applicability of various methods such as direct mapping approaches (SayCan, Code as Policies, ProgPrompt) and feedback-driven architectures (Inner Monologue). We also explore multimodal fusion techniques (RT-2, PaLM-E, VIMA, VoxPoser) and their application across diverse scenarios including education, home services, and manufacturing. Furthermore, we identify key challenges—including natural language parsing accuracy, robustness of multimodal fusion, long-horizon planning, and safety—and propose future research directions aimed at fostering the integration of embodied intelligence with natural language interaction. This study clarifies the research trajectory in this field, offering theoretical support for the broader popularization and deployment of robotic technology.

Keywords

Large Language Models (LLMs), robotic arms, zero-code control, natural language interaction, embodied AI

1. Introduction

The rapid advancement of Artificial Intelligence, particularly Large Language Models (LLMs), has enabled advanced capabilities in natural language processing, cross-modal reasoning, and complex task planning. These advances are reshaping robotic manipulation paradigms. Traditionally, controlling robotic arms has demanded specialized robotics software frameworks and toolchains, such as ROS and MoveIt!, as well as intricate control frameworks, requiring users to possess deep engineering and programming expertise for instruction coding and debugging. This high barrier to entry significantly restricts robot adoption in non-professional sectors such as education, home services, and small-to-medium enterprise (SME) manufacturing.

Furthermore, traditional methods often rely on manually predefined rules or state machines, which lack flexibility and adaptive capacity in dynamic and multi-step environments.

In response, both academia and industry are increasingly integrating LLMs into robotic control systems. The goal is to empower robots to execute complex, multi-step tasks directly through natural language instructions. In this paper, "zero-code control" refers to user-facing robot control in which non-expert users specify tasks through natural language without manually writing robot control programs, although the underlying system may still generate intermediate code, policies, or action representations. These "zero-code" implementation methods are designed to reduce development costs and operational complexity, making robotic technology more accessible to non-professionals. Recent exploratory achievements demonstrate the potential of combining LLMs with multimodal perception, advanced task planning, and intuitive human-robot interaction to support autonomous operation in complex environments.

Despite this progress, the field is still in a phase of rapid iteration, with several core issues awaiting resolution. These include the precise and robust mapping of natural language to robotic actions, the reliable fusion of multimodal information, ensuring safety during task execution, and maintaining adaptability across diverse application scenarios. Much existing research primarily focuses on implementation, often lacking a comprehensive categorization of the underlying technical frameworks or a systematic discussion of future trends.

This study addresses these gaps by systematically reviewing recent LLM-based zero-code implementation technologies for robotic arms. We aim to establish a clear technical classification framework, analyze the performance characteristics and applicability of different methods, and propose key challenges alongside promising research directions for the future. This work will not only help to clarify the research trajectory in this burgeoning field but also provide theoretical support and practical reference for the broader popularization and deployment of advanced robotic technology.

2. Classification and Analysis of Technical Routes

2.1 Models Based on Direct Mapping from Natural Language to Action Sequences

Direct mapping from natural language to action sequences is an important route for zero-code intelligent control. This approach leverages the semantic understanding, task decomposition, and logical reasoning capabilities of LLMs to translate natural language inputs into executable action steps, skill sequences, or control programs. Compared to traditional methods reliant on manual programming and symbolic planning, this route significantly lowers the operational threshold and enhances the system's adaptability to complex, multi-step tasks.

In terms of specific research, SayCan [1] is highly representative. It integrates the high-level language planning of LLMs with robotic skill affordances, allowing the system to select and combine appropriate action sequences from a skill library based on the instruction's context. Code as Policies [2] extends this mapping by enabling LLMs to generate executable policy code directly, shifting control from "action selection" to "program generation." ProgPrompt [3] emphasizes using prompt engineering to guide LLMs in generating structured task plans. Finally, Inner Monologue [4] introduces environmental feedback during execution, allowing the model to dynamically revise plans based on perception, thereby enhancing closed-loop adaptability.

2.2 Control Architectures Integrating Vision-Language-Action (VLA) Multimodality

The fusion of Vision-Language-Action (VLA) into a multimodal control architecture is a major development trend. Unlike text-only methods, this approach emphasizes the unification of language understanding, visual perception, and action execution. Its core lies in achieving effective alignment between task semantics, target locations, object attributes, and manipulation actions through multimodal joint modeling.

Key research includes RT-1 [5], which built a Transformer-based model mapping images and instructions to actions, validating the benefit of large-scale robotic data. This was followed by RT-2 [6], which transferred semantic knowledge from vision-language models into robotic control. PaLM-E [7] embeds visual signals and robot states into a unified framework. VIMA [8] utilizes multimodal prompts for general manipulation, while

VoxPoser [9] combines language instructions with 3D spatial representations for fine-grained tasks under spatial constraints. Additionally, CLIPort [10] laid the groundwork for vision-language-guided manipulation.

2.3 Affordance-Based, Feedback-Driven, and Data-Driven Control Methods

The synergy between RL and LLMs focuses on combining the high-level reasoning of LLMs with the low-level optimization and stability of RL. In this "High-level Planning – Low-level Control" framework, the LLM decomposes instructions into sub-tasks, while the RL policy evaluates action feasibility and executes the movement.

SayCan [1] exemplifies this direction by combining language intent with affordance-based action evaluation. Inner Monologue [4] further incorporates environmental feedback to support iterative control adjustments. Although RT-1 and RT-2 [5, 6] are not primarily reinforcement-learning frameworks, they provide scalable paradigms for integrating semantic knowledge with action learning. The Open X-Embodiment dataset [11] further supports cross-embodiment robot learning by providing large-scale data for training and evaluating generalist robotic models. However, this route still faces challenges regarding high training costs and limited policy generalization.

2.4 Emerging Extensions Toward Digital-Twin-Assisted Human-Robot Collaboration

Cognitive Digital Twin systems represent the evolution toward intelligent, interactive, and systemic control. Unlike traditional digital twins that focus on physical mapping, "Cognitive Digital Twins" integrate LLMs and multimodal perception to "understand intent" and "predict outcomes," facilitating collaborative decision-making.

Research such as Inner Monologue [4] provides useful inspiration for feedback-driven state updating in human-robot interaction. PaLM-E [7] offers a potential basis for multimodal state perception through its unified embedding of visual and robotic states. VIMA and VoxPoser [8, 9] may also provide methodological references for task transfer and spatial planning in virtual-physical synchronized environments. While still in the exploratory phase, these directions may support future research on risk assessment, remote operation, and human-robot collaboration.

3. Application Scenario Adaptability of Different Technical Routes

The transition from traditional programming to natural language interaction in robotics represents a fundamental paradigm shift. At the core of LLM-based zero-code control is the ability to leverage LLMs' semantic understanding, task decomposition, and logical reasoning to translate natural language inputs into executable robotic actions. This section classifies current technical routes and discusses their adaptability across various application scenarios.

3.1 Classification of Technical Routes

Current LLM-based zero-code technologies for robotic arms can be broadly categorized into three main routes, each with distinct characteristics:

3.1.1 Direct Mapping from Natural Language to Action Sequences

This foundational approach directly translates natural language commands into robot-executable action steps, skill sequences, or control programs. Compared to manual programming, it significantly lowers the operational threshold and enhances adaptability for multi-step tasks.

SayCan [1] is a highly representative method that integrates the high-level language planning capabilities of LLMs with robotic skill affordances. It allows the system to select and combine appropriate action sequences from a predefined skill library based on the instruction's context and the robot's current capabilities.

Code as Policies [2] extends this mapping by enabling LLMs to directly generate executable policy code. This shifts control from mere "action selection" to "program generation," offering high flexibility. However, its effectiveness heavily depends on the quality of the generated code.

ProgPrompt [3] emphasizes the use of carefully designed prompt engineering to guide LLMs in generating structured, coherent task plans. This method provides clear task structures for complex tasks.

3.1.2 Feedback-Driven and Closed-Loop Control Architectures

While direct mapping is useful, real-world environments demand adaptability. These architectures incorporate environmental feedback during execution, allowing for dynamic plan revision.

Inner Monologue [4] introduces environmental perception feedback during task execution. This enables the model to dynamically revise its plans based on real-time sensory input, improving closed-loop adaptability and robustness to unexpected changes. Compared to ProgPrompt, Inner Monologue offers stronger real-time adaptability due to its feedback mechanism.

3.1.3 Multimodal Information Fusion Architectures

Complex robotic tasks often require understanding not only language but also visual and proprioceptive information. Multimodal architectures integrate these diverse sensory inputs.

RT-2 and PaLM-E [6, 7] are examples of large-scale multimodal architectures that demonstrate the potential for generalization across different tasks and environments. They effectively combine visual perception with language understanding to enable robust robot control.

VoxPoser [9] is particularly relevant to precision-oriented manipulation under spatial constraints because it integrates language instructions with 3D spatial information.

VIMA [8] is mainly characterized by its multimodal prompt design and task transfer capabilities, allowing robots to adapt to new task specifications and environments in a more structured way.

3.2 Application Scenario Classification and Adaptability

The “zero-code” approach, while broadly applicable, often requires scenario-specific optimization to improve its effectiveness.

- **Education and Research:** In these environments, the priority is often on interactive learning and low barriers to entry. Direct mapping methods such as Code as Policies and ProgPrompt [2, 3] are suitable for demonstrations, teaching fundamental robotic concepts, and rapid prototyping due to their intuitive nature and ease of use.
- **Home Services and Assistance:** These environments are typically unstructured, dynamic, and unpredictable. Multimodal architectures such as RT-2 and PaLM-E [6, 7] may be better suited here. Their ability to integrate object recognition, environmental perception, and language understanding makes them promising for domestic tasks.
- **SME Manufacturing and Flexible Production:** These sectors demand high precision, quick reconfiguration, and robustness. Methods like VIMA and VoxPoser [8, 9] may be particularly relevant because VIMA supports multimodal task specification and transfer, while VoxPoser emphasizes fine-grained spatial constraint handling.
- **Human-Robot Collaboration (HRC) and Remote Operation:** Scenarios requiring high safety, real-time feedback, and immediate human oversight can benefit from frameworks incorporating closed-loop control. Inner Monologue [4], with its dynamic plan revision mechanism, provides a useful reference for real-time adaptability in collaborative and remotely operated tasks, while cognitive digital twin frameworks may further support safety monitoring and human oversight.

4. Key Challenges and Future Research Directions

Despite significant progress in LLM-based zero-code control for robotic arms, transitioning from experimental validation to stable, robust industrial and domestic applications presents several formidable challenges. Addressing these issues is crucial for the widespread adoption of this technology.

4.1 Key Challenges

Accuracy of Natural Language Parsing and Grounding: Natural language is inherently ambiguous, context-dependent, and highly variable. Current LLMs must evolve beyond mere semantic understanding to achieve "grounded" execution. This requires precisely mapping abstract language to concrete physical actions, accounting for spatial coordinates, object properties, environmental states, and operational constraints. Failures in grounding can lead to misinterpretations and unsafe robot behaviors.

Robustness of Multimodal Fusion: The integration of diverse sensory data—vision, language, touch, and proprioception—in real time remains a significant hurdle. Aligning these heterogeneous data streams, especially in complex and dynamic scenes, is difficult. Sensory errors, noise, and semantic mismatches across modalities can severely degrade system performance. There is a pressing need for more robust cross-modal alignment techniques and lightweight, real-time feedback mechanisms that can effectively handle imperfect and partial information.

Long-Horizon Planning and Closed-Loop Execution: Many current systems struggle with long-sequence tasks where intermediate failures can cascade and derail the entire operation. Integrating LLMs with sophisticated memory mechanisms, symbolic planning, and more advanced feedback control loops is essential for dynamic error detection, recovery, and continuous adaptation over extended task sequences. The ability to learn from errors and adapt plans in real-time is critical for handling unforeseen circumstances.

Safety and Reliability in Physical Environments: The "black-box" nature of large language models poses inherent risks, particularly in physical environments where errors can have tangible and potentially harmful consequences. Ensuring the safety and reliability of LLM-driven robotic systems is paramount. This necessitates developing rigorous safety-constrained prompting techniques, formal verification methods for generated code, and robust human-in-the-loop safeguards that allow for immediate intervention and oversight. Explainability of LLM decisions is also crucial for building trust and diagnosing failures.

4.2 Future Research Directions

Based on the identified challenges, several promising research directions emerge that could drive the field forward:

- **Enhanced Grounded Language Understanding:** Future research should focus on developing LLMs with deeper embodied understanding, capable of directly translating high-level goals into low-level motor commands while adhering to physical laws and environmental constraints. This includes integrating common-sense reasoning and physics-based simulations into the training process.
- **Advanced Multimodal Architectures with Causal Reasoning:** Developing next-generation multimodal fusion architectures that not only integrate sensory data but also perform causal reasoning will be vital. This would allow robots to understand why certain actions lead to particular outcomes, improving their ability to predict, adapt, and recover from unexpected situations. Research into efficient, low-latency multimodal processing for real-time applications is also critical.
- **Self-Correction and Continual Learning for Long-Horizon Tasks:** Future systems need improved capabilities for self-correction, dynamic replanning, and continual learning directly from interaction and experience. This includes sophisticated memory architectures that can store and retrieve task-specific knowledge, as well as meta-learning approaches that enable robots to quickly adapt to new tasks and environments without extensive retraining.
- **Provably Safe and Explainable AI for Robotics:** Research into safety will remain a top priority. This involves exploring formal methods for verifying LLM-generated policies, developing robust human-robot interaction (HRI) protocols for safety, and creating more transparent and explainable LLMs whose decision-making processes can be audited and understood. The goal is to build robotic systems that are not only capable but also reliably safe and trustworthy for deployment in sensitive environments.
- **Foundation Models for Embodied AI:** The development of large-scale foundation models specifically pre-trained for embodied AI, capable of understanding both language and rich sensorimotor experiences across diverse robot platforms, may provide an important basis for future

research. Such models could provide a unified framework for learning a wide range of robotic skills and transferring them efficiently.

5. Conclusion

This study has systematically reviewed the technical landscape of LLM-based zero-code control for robotic arms. The shift from traditional programming to intuitive natural language interaction marks an important paradigm shift, lowering the operational barrier for robotic applications across diverse sectors. We have established a technical classification framework, analyzing various methods from direct mapping (e.g., SayCan, Code as Policies, ProgPrompt) to feedback-driven architectures (e.g., Inner Monologue) and multimodal fusion systems (e.g., RT-2, PaLM-E, VIMA, VoxPoser). We further explored their applicability across education, home services, and manufacturing.

The current technical trajectory is clearly moving beyond isolated language processing towards integrated perception-planning-execution systems. While significant progress has been made, substantial challenges remain, particularly concerning the accuracy of natural language grounding, the robustness of multimodal fusion, the capability for long-horizon planning with closed-loop execution, and ensuring safety and reliability in physical environments. By addressing these key challenges and pursuing the identified future research directions, this field can achieve the deep integration of embodied intelligence and natural language interaction, reshaping how robots are deployed and utilized in industrial, service, and educational domains. This survey aims to provide a solid foundation for subsequent research, accelerating the development and widespread adoption of intelligent robotic technology.

References

[1] Ichter, B., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., Kalashnikov, D., Levine, S., Lu, Y., Parada, C., Rao, K., Sermanet, P., Toshev, A. T., Vanhoucke, V., ... Fu, C. K. (2023). Do as I can, not as I say: Grounding language in robotic affordances. In K. Liu, D. Kulic, & J. Ichnowski (Eds.), Proceedings of the 6th Conference on Robot Learning (Vol. 205, pp. 287–318). PMLR.

[2] Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., & Zeng, A. (2023). Code as policies: Language model programs for embodied control. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9493–9500). IEEE. https://doi.org/10.1109/ICRA48891.2023.10160591

[3] Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., & Garg, A. (2023). ProgPrompt: Generating situated robot task plans using large language models. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11523–11530). IEEE. https://doi.org/10.1109/ICRA48891.2023.10161317

[4] Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Jackson, T., Brown, N., Luu, L., Levine, S., Hausman, K., & Ichter, B. (2023). Inner Monologue: Embodied reasoning through planning with language models. In K. Liu, D. Kulic, & J. Ichnowski (Eds.), Proceedings of the 6th Conference on Robot Learning (Vol. 205, pp. 1769–1782). PMLR.

[5] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., ... Zitkovich, B. (2023). RT-1: Robotics Transformer for real-world control at scale. In Proceedings of Robotics: Science and Systems. Robotics: Science and Systems Foundation. https://doi.org/10.15607/RSS.2023.XIX.025

[6] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P. R., Salazar, G., ... Han, K. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In J. Tan, M. Toussaint, & K. Darvish (Eds.), Proceedings of the 7th Conference on Robot Learning (Vol. 229, pp. 2165–2183). PMLR.

[7] Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., ... Florence, P. (2023). PaLM-E: An embodied multimodal language model. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 8469–8488). PMLR.

[8] Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., & Fan, L. (2023). VIMA: Robot manipulation with multimodal prompts. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 14975–15022). PMLR.

[9] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., & Fei-Fei, L. (2023). VoxPoser: Composable 3D value maps for robotic manipulation with language models. In J. Tan, M. Toussaint, & K. Darvish (Eds.), Proceedings of the 7th Conference on Robot Learning (Vol. 229, pp. 540–562). PMLR.

[10] Shridhar, M., Manuelli, L., & Fox, D. (2022). CLIPort: What and where pathways for robotic manipulation. In A. Faust, D. Hsu, & G. Neumann (Eds.), Proceedings of the 5th Conference on Robot Learning (Vol. 164, pp. 894–906). PMLR.

[11] Open X-Embodiment Collaboration. (2024). Open X-Embodiment: Robotic learning datasets and RT-X models. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6892–6903). IEEE. https://doi.org/10.1109/ICRA57147.2024.10611477

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights