Position Paper: A New Paradigm for Robot Multimodal Understanding and Decision-Making with Large Language Models as the Cognitive Core

Yulai Zhang[1], Yinrong Zhang[1], Ting Wu[1], and Linqi Ye[1,*]
[1]*School of Future Technology, Shanghai University, Shanghai, China*
[*]*Corresponding author: yelinqi@shu.edu.cn*

***Abstract*—This paper presents a systematic exploration of the application of large language models (LLMs) as cognitive cores in robotics, focusing on multimodal understanding and intelligent decision-making. While traditional robotic architectures face inherent limitations in environmental modeling precision and task-specific data dependency—struggling with open-ended instruction comprehension, dynamic environment adaptation, and cross-task knowledge transfer—the emergence of LLMs offers a transformative solution. By positioning LLMs as the central cognitive core, robotic systems can achieve deeply integrated perception, reasoning, and decision-making capabilities. This paradigm empowers robots to interpret multimodal inputs more effectively, perform commonsense reasoning, and generate executable action sequences. The paper delineates key implementation pathways, including unified semantic interfaces, commonsense reasoning engines, and metacognitive coordination mechanisms. Furthermore, it examines advanced techniques for enhancing multimodal understanding through vision-language-action models, cross-modal commonsense comprehension, and open-vocabulary semantic construction. The discussion extends to how LLMs facilitate efficient intelligent decision-making processes. Finally, the paper outlines future research directions and proposes mechanisms for achieving long-term adaptability and continuous learning within this new paradigm.**

***Index Terms*—Large Language Models, Robot Cognitive Core, Multimodal Understanding, Intelligent Decision-Making**

I. Introduction

Developing robots capable of operating autonomously, safely, and efficiently in complex, unstructured human environments remains a core objective in AI and robotics [1]. Traditional robotics paradigms—whether model-based planning or data-driven reinforcement learning—rely on precise environmental modeling, finite state spaces, or vast task-specific demonstration data [2], [3]. This limits their ability to handle open-ended commands, dynamic adaptation, and cross-task knowledge transfer [4]. For instance, a robot programmed for "pick-and-place" tasks struggles with tasks like "tidy up the living room," lacking the semantic understanding to map abstract goals to actions in a dynamic environment. Meanwhile, large language models (LLMs) based on Transformer architectures, pre-trained on vast text corpora, excel in common-sense reasoning and code generation [5]. This raises the question: How can we integrate the symbolic cognitive capabilities of LLMs with a robot's embodied perception and action capabilities?

Recent research confirms a profound paradigm shift: establishing "LLM-as-Cognitive-Core" [6], [7]. This fundamentally differs from prior approaches treating LLMs as peripheral "task planners" or "natural language interfaces." Previously, LLMs served as high-level instruction parsers for traditional robotic systems. The emerging paradigm positions LLMs at the center of robotic cognition, responsible for constructing unified world models, interpreting multimodal sensor data, initiating complex reasoning, and generating executable plans [8], [9]. This shifts robot intelligence research from "how to better perceive and control" to "how to understand and think like humans."

Currently, embodied intelligence has seen dual breakthroughs in "cognitive intelligence" (the brain) and "physical intelligence" (the body). LLMs enable robots to "perceive, think, and interact" [10], while reinforcement and imitation learning enhance motor and manipulation skills [11]. These parallel advancements create a unique convergence opportunity. However, large-scale deployment remains nascent, with core debates over model architectures, data paradigms, and optimal robotic forms [12]. Key questions persist: Should LLMs be fine-tuned for robotics or remain general-purpose? How to ensure safety and reliability of LLM-generated plans? What are effective ways to ground linguistic knowledge in sensory-motor experience? Thus, articulating a new paradigm centered on LLMs as cognitive cores is crucial for consensus and future directions.

This paper aims to systematically elaborate on this new paradigm, with the central thesis that the application of large language models as the cognitive core of robotic systems will fundamentally transform traditional approaches to robot perception, planning, and control. To substantiate this claim, we will first discuss the inevitability of this paradigm shift and its core conceptual essence. We will then delve into how this new architecture enables unprecedented multimodal understanding capabilities and intelligent decision-making mechanisms in robots. Subsequently, we will analyze how such systems can achieve continuous learning and adaptive evolution. Finally, we will outline critical future research directions and con-

This work was supported by the Science and Technology Commission of Shanghai Municipality (24511103304).

clude. Through this systematic analysis, we seek to provide a clear theoretical framework and developmental roadmap that demonstrates how this paradigm shift redefines the foundations of robotic intelligence.

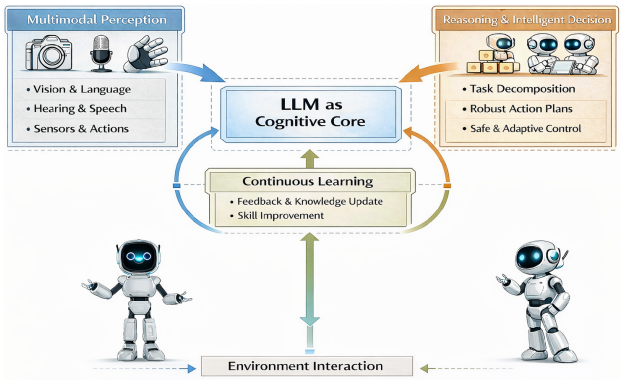

Fig. 1. Overview of the LLM-as-Cognitive-Core Paradigm for Robotics.

II. LLMs as Cognitive Cores: Paradigm Shift

Traditional robotic architectures follow sequential modular paradigms with discrete perception, planning, and control stages. These rigid structures limit abstraction, reasoning, and contextual interpretation [13]. Large language models (LLMs) offer a transformative solution as a central cognitive substrate, redefining robotic intelligence [14]. LLMs act as core orchestrators for multimodal sensorimotor integration, bridging the semantic-symbolic divide through a unified cognitive framework. This role manifests through three synergistic dimensions advancing embodied intelligence.

A. Unified semantic interface and contextual cognition

The LLM acts as a *unified semantic interface*, mapping multimodal inputs—language, vision, and proprioception—into a shared semantic space. This enables cross-modal alignment and contextual understanding, providing a common substrate for perception, reasoning, and action [15]. For instance, the command "retrieve the textbook from the upper shelf" is dynamically linked to real-time visual detections, spatial relationships, and grasp hypotheses [16]. This semantic binding resolves referential ambiguity, infers implicit constraints, and adapts to perceptual uncertainty.

Architecturally, this is realized through end-to-end Vision-Language-Action (VLA) models, which learn direct mappings from pixels and language commands to motor policies [17]. Trained on large-scale, language-annotated teleoperation datasets, these models acquire grounded representations that support zero-shot generalization to novel objects, scenes, and instructions [18]. By bypassing hand-engineered representations, VLA models reduce system bottlenecks and enable tighter coupling between perception, cognition, and action.

B. Common-Sense Reasoning and Planning Engines

Beyond semantic mapping, LLMs function as *commonsense reasoning and planning engines*. Pre-trained on corpora that encapsulate physical, social, and procedural knowledge, LLMs internalize rich priors about object properties, causal relationships, and typical action sequences [19]. This knowledge enables them to perform non-trivial inference, such as inferring that "fragile" objects necessitate gentle manipulation, or decomposing abstract goals like "set the table" into structured sequences of sub-actions [20]. Such reasoning is critical for handling unexpected situations, interpreting underspecified instructions, and generating robust, long-horizon plans [21].

In this capacity, the cognitive core acts as a high-level task planner, translating abstract user intents into executable task networks. The planning problem can be formulated as finding an optimal policy π^* that maximizes expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right] \tag{1}$$

$$\text{s.t.} \quad a_t \sim \pi(\cdot|s_t, h_t), \quad s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t) \tag{2}$$

where s_t represents the state at time t, a_t the action, h_t the history of observations and actions, $\gamma \in [0, 1]$ the discount factor, and $\mathcal{P}$ the state transition dynamics. This planning process can be iterative, incorporating feedback from the environment or the user to refine and adapt the plan dynamically. The integration of LLM-based reasoning with classical symbolic planners or reinforcement learning frameworks further enhances the reliability and optimality of generated plans, addressing the limitations of purely neural or purely symbolic approaches [22], [23].

C. Metacognitive and Learning Coordinator

The third dimension of the cognitive core is *metacognitive and learning coordination*. LLMs allow robots to monitor performance, diagnose failures, and orchestrate learning for continuous improvement [24]. For instance, after a task failure, the LLM can analyze sensor data to generate a diagnosis (e.g., "grasp failed due to slippage") and trigger learning mechanisms like human feedback, experience retrieval, or targeted exploration [25]. This turns the robot into an adaptive, lifelong learner.

This role is evident in recent works. AutoRT uses LLMs/VLMs as central decision-makers for autonomous task proposal and data collection. HiAgent employs LLMs to manage working memory and dynamically update sub-goals, improving long-term task efficiency [26]. The MetaWorld framework builds hierarchical world models, separating semantic planning from physical control via a cognitive core to bridge the semantic-physical gap [27]. Together, these demonstrate LLMs' critical role as a cognitive core.

D. Case Study: Robotic Soccer Paradigms

To concretize the paradigm shift, we contrast three approaches to a canonical task: robotic soccer. This domain requires real-time perception, strategic planning, and precise control, making it ideal for comparing paradigms.

- Traditional Modular Approach: Separates perception, planning, and control into isolated modules. This leads to brittle systems where errors cascade and adaptation to new commands (e.g., "play defensively") requires manual re-engineering [28].
- End-to-End Reinforcement Learning (RL): Trains a monolithic policy mapping raw inputs to actions. While capable, it requires massive task-specific data, lacks interpretability, and struggles to generalize or incorporate new knowledge without retraining [29].
- LLM-as-Cognitive-Core: An LLM (or VLA model) serves as the central cognitive unit. It processes multimodal inputs (vision, language) to generate high-level plans (e.g., "pass to the open player on the left"), which are executed by low-level controllers. Crucially, the LLM's pre-trained knowledge enables zero-shot strategic reasoning and natural language understanding, directly replacing the traditional perception-planning pipeline without task-specific retraining [30].

This case illustrates our thesis: the LLM-as-Cognitive-Core paradigm fundamentally redefines robot perception and planning by introducing a flexible, knowledge-driven semantic layer, overcoming the limitations of both modular rigidity and RL opacity.

III. Multimodal Understanding: From Fusion to Cognition

Under the new paradigm where LLMs serve as the cognitive core, robots' multimodal understanding transcends traditional data-layer or feature-layer fusion, elevating to contextualized cognition at the semantic level. This signifies that robots can not only "see" pixels and "hear" words but also comprehend their underlying meanings, relationships, and contexts, forming holistic, inferable representations of their environment [31], [32]. This transformation is primarily achieved through three key directions:

A. Joint modeling and end-to-end learning of vision-language-action

Next-generation vision-language-action (VLA) models epitomize this trend. Key aspects include:

- Architectural Integration: Models like ChatVLA [33] directly couple visual perception, language understanding, and action generation at the architectural level, enabling seamless multimodal fusion.
- Formal Foundation: The core objective can be formalized as learning a conditional probability distribution:

$$P(A \mid V, L) = \prod_{t=1}^{T} P(a_t \mid v_{\leq t}, l_{\leq t}, a_{<t}) \tag{3}$$

  where V denotes the visual input sequence, L represents the language instruction, $A = \{a_1, a_2, \ldots, a_T\}$ is the action sequence, and t is the time step.
- Training and Generalization: Through pre-training on internet-scale multimodal data, these models achieve robust visual-semantic generalization capabilities, enabling robots to perform numerous open-word "zero-shot" tasks such as folding clothes or assembling boxes according to verbal instructions [34].
- Training Challenges: VLA training faces core challenges including catastrophic forgetting and task interference [35]. Advanced methods such as staged alignment training and hybrid expert networks have been proposed to address these issues while preserving multimodal understanding capabilities [36].

B. Physical Property Inference and Cross-Modal Common Sense Understanding

The cognitive core enables robots to comprehend deep physical properties and functionalities of objects beyond their appearance. This capability encompasses several key aspects:

- Visual Property Inference: Fine-tuning Vision-Language Models allows robots to infer physical concepts like material and rigidity from a single image, applying this knowledge for safe operation planning [37].
- Cross-Modal Sensing: By analyzing force/torque sensor data through models like GPT-4V to generate temporal images, robots can indirectly "perceive" liquid viscosity, achieving cross-modal reasoning from non-visual modalities to physical understanding [38].
- Foundation for Manipulation: This physical intuition forms the foundation for safe and dexterous robotic manipulation, enabling more nuanced and adaptive interaction with the physical world [39].

C. Open-Vocabulary Semantic Context Construction and Spatial Intelligence

LLM-driven cognitive systems support open-vocabulary perception and scene understanding. The HOV-SG method [40] utilizes open-vocabulary vision foundation models to construct hierarchical 3D scene semantic graphs online, enabling robots to comprehend complex referents like "the drawer to the left of the desk in the study" and support long-range, cross-floor navigation tasks based on natural language. This exemplifies spatial intelligence—the ability to understand and act in physical space by integrating geometric, semantic, and functional knowledge—which is crucial for embodied cognition [41].

Current research trends are propelling VLA toward greater efficiency and intelligence [42]: Discrete diffusion model architectures enable efficient parallel generation of action sequences [43]; embodied thought chain techniques enable robots to reason through intermediate steps before acting, enhancing planning interpretability and complexity [44]; while action segmenter research focuses on discretizing continuous actions into "lexemes" better processed by VLM, addressing a key challenge in modality alignment [45].

IV. Intelligent Decision-Making: Layered, Collaborative, and Resilient

With LLMs as the cognitive core, robotic decision intelligence exhibits new characteristics of hierarchical struc-

ture, collaborative capabilities, and high resilience. Decision-making is no longer a single feedforward process but rather an evaluable, correctable closed-loop system based on the cognitive core [46]. This closed-loop is primarily achieved through the following mechanisms:

A. Hierarchical task decomposition and neuro-symbolic grounding

LLMs excel at decomposing abstract instructions into concrete steps. To ensure precision and physical feasibility, neuro-symbolic fusion has emerged as a key trend [47]. Its key aspects include:

- Hybrid Planning Framework: Systems like CLMASP [48] and TwoStep [49] integrate LLMs with symbolic planners (e.g., ASP, PDDL). LLMs handle high-level decomposition and common-sense reasoning to produce a plan skeleton.
- Symbolic Refinement and Validation: Symbolic systems then ground and validate the skeleton into an executable plan, ensuring logical consistency and physical feasibility.
- Performance Gain: This synergy raises success rates from under 2% (pure neural) to over 90%, balancing neural generalization with symbolic reliability [50].

B. Verifiable Decisions and Enhanced Robustness

To prevent cumulative errors from biased single decisions, new frameworks introduce verification mechanisms [51]. A prevalent advanced approach constructs a verifiable confidence scoring function for decision plans. This function comprehensively integrates the model's confidence in the candidate plan, the estimated uncertainty, and its relative advantage over alternatives. A robust decision criterion can be formulated as:

$$\frac{P(y_{\text{plan}} \mid x)}{P(y_{\text{alt}} \mid x)} \cdot \exp(-\lambda \cdot \sigma(y_{\text{plan}})) > \tau, \tag{4}$$

where y_{plan} denotes the current plan, y_{alt} represents the alternative plan, $\sigma(y_{\text{plan}})$ quantifies the uncertainty of the current plan (e.g., predictive variance or entropy), $\lambda \geq 0$ is a penalty coefficient that adjusts the influence of uncertainty, and τ is the preset threshold. When the inequality is not satisfied, the system may trigger replanning or request human intervention, thereby enhancing decision robustness and safety.

C. Multi-agent Dialogic Collaboration

When multiple robots possess LLM-based "brains," they can engage in efficient negotiation via natural language [52]. For instance, the MHRC framework enables decentralized collaboration among three heterogeneous robot types: mobile robots, robotic arms, and mobile manipulation platforms [53]. Here, LLMs serve as a unified "task allocation and communication protocol," coordinating the group to accomplish complex tasks like exploration and transportation, demonstrating an advanced form of collaboration.

D. Lightweight and Human-Centered Interaction Design

To facilitate deployment of cognitive cores, research is evolving toward lightweight and interactive approaches. The InteLiPlan framework, as a lightweight LLM planner, implements robot-agnostic pipelines and introduces a human-in-the-loop mechanism allowing real-time manual intervention. It achieves high success rates and fault recovery capabilities even under onboard computing constraints [54]. The CLEAR platform emphasizes programming robotic behavior through prompt engineering (rather than fine-tuning), offering a flexible and user-friendly new paradigm for human-robot interaction [55].

In summary, through hierarchical and verification mechanisms, decision-making achieves layered structure and high resilience; multi-agent dialogue enables collaboration; and lightweight interaction design ensures the paradigm's feasibility and user-friendliness in practical systems. Collectively, these mechanisms form a robust and adaptive new decision intelligence system with LLMs as its cognitive core.

V. Continuous Learning and Adaptation

Robots deployed in the real world must adapt to evolving environments, tasks, and user preferences. The paradigm centered on LLMs as cognitive cores provides a natural framework for continual learning and lifelong learning in robots, enabling them to evolve through interactions [56], [57]. This framework is primarily achieved through three complementary mechanisms:

A. Extracting and generalizing knowledge from natural language feedback

When robots make errors, users can provide corrections in natural language. LLMs parse this feedback to distill generalizable rules or knowledge fragments (e.g., "condiments are typically stored on refrigerator door shelves") and store them in a structured knowledge base. This process can be formalized as dynamic updates to the knowledge base:

$$K_{t+1} = K_t \oplus \Phi\big(\mathcal{F}_t, s_t, a_t; \Theta\big) \tag{5}$$

Here, $\mathcal{K}_t \in K$ denotes the structured knowledge base at time step t, $\mathcal{F}_t$ represents the natural language feedback, while $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ characterize the environmental state and the executed action, respectively. The knowledge update operator $\oplus : K \times \mathcal{R} \to K$ realizes the fusion operation in the rule space. $\Phi : \mathcal{L} \times \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ denotes the rule extraction function parameterized by Θ, which maps unstructured feedback into generalizable rule representations $\mathcal{R}$. This mechanism automatically activates corresponding rules in semantically similar scenarios through the semantic retrieval function $\Psi : \mathcal{S} \times \mathcal{A} \times K \to \mathcal{R}$, thereby significantly reducing the frequency of human intervention. Similarly, incremental learning research demonstrates how large language models can learn directly from errors through interactive code generation and modification [58].

B. Automated Skill Library Construction and Curriculum Generation

The cognitive core enables robots to autonomously expand their capability boundaries. The CurricuLLM system [59] leverages LLM world knowledge to automatically design progressive training curricula for complex skills (e.g., tool usage). Furthermore, the LEAGUE++ framework [60] deeply integrates LLMs with Task-Aware Motion Planning (TAMP) and Deep Reinforcement Learning (DRL), achieving unified automation of task decomposition, skill creation, and reward generation to support continuous skill learning in long-term tasks.

C. Reinforcement Learning Fine-Tuning and Closed-Loop Optimization

While imitation learning provides a strong initial policy, reinforcement learning (RL) is crucial for fine-tuning and robustness. The cognitive core enhances this process by:

- Bridging the "Last Mile": RL fine-tunes policies from Vision-Language-Action models, adapting them to novel or edge-case scenarios for robust performance [61].
- LLM as a Learned Reward Function: Multimodal LLMs act as "preference critics," evaluating trajectory videos to provide feedback. This trains a reward model that guides RL policy optimization [62].
- Forming a Complete Learning Loop: A closed loop is created from high-level cognitive understanding (LLM) to low-level control optimization (RL), incorporating human feedback and autonomous exploration for continuous refinement of efficiency, robustness, and safety.

Through natural language interaction, autonomous exploration, and RL-based refinement, robot systems with an LLM cognitive core achieve continuous adaptation and evolution throughout their lifecycle.

VI. CONCLUSION

Large Language Models (LLMs) serve as the cognitive core driving a paradigm shift in robotics. This paper dissects LLM-centric architectures, elucidating how they reconstruct robots' multimodal understanding, decision-making, and continuous learning capabilities. By serving as a unified semantic interface, a common-sense reasoning engine, and a metacognitive coordinator, LLMs effectively bridge the gap between abstract thinking and contextual cognition in robots. This facilitates the evolution of perception toward semantic understanding, decision-making toward resilient closed-loop systems, and learning toward lifelong adaptation.

Future breakthroughs depend on multi-directional collaboration: Building embodied world models with reasoning capabilities to overcome physical common-sense limitations; Enhancing system reliability and interpretability through neuro-symbolic fusion; Upholding human-centered principles to foster symbiosis between humans, machines, and the environment. The field stands at the critical juncture of deeply integrating LLM cognitive capabilities with robotic physical bodies. This integration represents both a formidable engineering challenge and a pivotal journey in exploring the essence of intelligence to shape the next generation of collaborative partners. Its advancement urgently requires close interdisciplinary collaboration and innovation on a global scale.

REFERENCES

[1] H. Jeong, H. Lee, C. Kim, et al., "A survey of robot intelligence with large language models," *Applied Sciences*, vol. 14, no. 19, p. 8868, 2024.

[2] Y. Kim, D. Kim, J. Choi, et al., "A survey on integration of large language models with intelligent robots," *Intelligent Service Robotics*, vol. 17, no. 5, pp. 1091-1107, 2024.

[3] B. Sindhu, R. P. Prathamesh, M. B. Sameera, et al., "The Evolution of Large Language Model: Models, Applications and Challenges," in *Proceedings of the 2024 International Conference on Current Trends in Advanced Computing (ICCTAC)*, 2024, pp. 1-8. DOI: 10.1109/icctac61556.2024.10581180.

[4] K. Asuzu, H. Singh, M. Idrissi, "Human–robot interaction through joint robot planning with large language models," *Intelligent Service Robotics*, vol. 18, no. 2, pp. 261-277, 2025. DOI: 10.1007/s11370-024-00570-1.

[5] J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[6] K. Black, N. Brown, D. Driess, et al., "π_0: A Vision-Language-Action Flow Model for General Robot Control," arXiv preprint arXiv:2410.24164, 2024.

[7] J. Wang, E. Shi, H. Hu, et al., "Large language models for robotics: Opportunities, challenges, and perspectives," *Journal of Automation and Intelligence*, vol. 4, no. 1, pp. 52-64, 2025.

[8] Y. Ouyang, J. Li, Y. Li, et al., "Long-horizon locomotion and manipulation on a quadrupedal robot with large language models," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 11157-11164.

[9] M. Hu, T. Chen, Q. Chen, et al., "Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 32779-32798.

[10] Z. Luan, Y. Lai, R. Huang, et al., "Enhancing robot task planning and execution through multi-layer large language models," *Sensors*, vol. 24, no. 5, p. 1687, 2024.

[11] R. Mon-Williams, G. Li, R. Long, et al., "Embodied large language models enable robots to complete complex tasks in unpredictable environments," *Nature Machine Intelligence*, 2025, pp. 1-10.

[12] P. Mirowski, J. Love, K. Mathewson, et al., "A robot walks into a bar: Can language models serve as creativity support tools for comedy? An evaluation of LLMs' humour alignment with comedians," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1622-1636.

[13] S. S. Kannan, V. L. N. Venkatesh, B. C. Min, "Smart-LLM: Smart multi-agent robot task planning using large language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 12140-12147.

[14] L. X. Shi, B. Ichter, M. Equi, et al., "Hi robot: Open-ended instruction following with hierarchical vision-language-action models," arXiv preprint arXiv:2502.19417, 2025.

[15] Z. Li, X. Wu, H. Du, et al., "A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges," arXiv preprint arXiv:2501.02189, 2025.

[16] J. Duan, W. Yuan, W. Pumacay, et al., "Manipulate-anything: Automating real-world robots using vision-language models," arXiv preprint arXiv:2406.18915, 2024.

[17] W. Yuan, J. Duan, V. Blukis, et al., "Robopoint: A vision-language model for spatial affordance prediction for robotics," arXiv preprint arXiv:2406.10721, 2024.

[18] A. C. Cheng, Y. Ji, Z. Yang, et al., "Navila: Legged robot vision-language-action model for navigation," arXiv preprint arXiv:2412.04453, 2024.

[19] M. A. Graule, V. Isler, "GG-LLM: Geometrically grounding large language models for zero-shot human activity forecasting in human-aware task planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 568-574.

[20] T. Birr, C. Pohl, A. Younes, et al., "Autogpt+ p: Affordance-based task planning with large language models," arXiv preprint arXiv:2402.10778, 2024.

[21] M. G. Arenas, T. Xiao, S. Singh, et al., "How to prompt your robot: A promptbook for manipulation skills with code as policies," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 4340-4348.

[22] S. S. Raman, V. Cohen, I. Idrees, et al., "Cape: Corrective actions from precondition errors using large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14070-14077.

[23] Y. Wu, Z. Xiong, Y. Hu, et al., "SELP: Generating safe and efficient task plans for robot agents with large language models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 2599-2605.

[24] L. Zha, Y. Cui, L. H. Lin, et al., "Distilling and retrieving generalizable knowledge for robot manipulation via language corrections," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15172-15179.

[25] G. Tziafas, H. Kasaei, "Lifelong robot library learning: Bootstrapping composable and generalizable skills for embodied control with language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 515-522.

[26] M. Ahn, D. Dwibedi, C. Finn, et al., "Autort: Embodied foundation models for large scale orchestration of robotic agents," arXiv preprint arXiv:2401.12963, 2024.

[27] Y. Mu, T. Chen, Z. Chen, et al., "Robotwin: Dual-arm robot benchmark with generative digital twins," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27649-27660.

[28] A. Labiosa, Z. Wang, S. Agarwal, et al., "Reinforcement learning within the classical robotics stack: A case study in robot soccer," IEEE Int. Conf. Robot. Autom. (ICRA), 2025, pp. 14999-15006.

[29] F. Vahl, J. Griepenburg, J. Gutsche, et al., "SoccerDiffusion: Toward learning end-to-end humanoid robot soccer from gameplay recordings," arXiv preprint arXiv:2504.20808, 2025.

[30] M. Brienza, E. Musumeci, V. Suriani, et al., "LLCoach: Generating robot soccer plans using multi-role large language models," in Robot World Cup, Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 176-188.

[31] Y. Li, Z. Lai, W. Bao, et al., "Visual large language models for generalized and specialized applications," arXiv preprint arXiv:2501.02765, 2025.

[32] S. Liu, J. Zhang, R. X. Gao, et al., "Vision-language model-driven scene understanding and robotic object manipulation," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 21-26.

[33] Z. Zhou, Y. Zhu, M. Zhu, et al., "ChatVLA: Unified multimodal understanding and robot control with vision-language-action model," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 5377-5395.

[34] Z. Mandi, S. Jain, S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 286-299.

[35] F. Liu, F. Yan, L. Zheng, et al., "Robouniview: Visual-language model with unified view representation for robotic manipulation," arXiv preprint arXiv:2406.18977, 2024.

[36] A. Szot, B. Mazoure, H. Agrawal, et al., "Grounding multimodal large language models in actions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20198-20224, 2024.

[37] J. Gao, B. Sarkar, F. Xia, et al., "Physically grounded vision-language models for robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 12462-12469.

[38] W. Lai, T. Zhang, T. L. Lam, et al., "Vision-language model-based physical reasoning for robot liquid perception," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 9652-9659.

[39] H. Chen, Y. Yao, R. Liu, et al., "Automating robot failure recovery using vision-language models with optimized prompts," arXiv preprint arXiv:2409.03966, 2024.

[40] A. Werby, C. Huang, M. Büchner, et al., "Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[41] A. J. Sathyamoorthy, K. Weerakoon, M. Elnoor, et al., "Convoi: Context-aware navigation using vision-language models in outdoor and indoor environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 13837-13844.

[42] Z. Long, G. Killick, R. McCreadie, et al., "Robollm: Robotic vision tasks grounded on multimodal large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 12428-12435.

[43] Y. Jin, D. Li, J. Shi, et al., "RobotGPT: Robot manipulation learning from ChatGPT," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2543-2550, 2024.

[44] Y. Guo, Y. J. Wang, L. Zha, et al., "Doremi: Grounding language model by detecting and recovering from plan-execution misalignment," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 12124-12131.

[45] A. Lykov, D. Tsetserukou, "LLM-brain: AI-driven fast generation of robot behaviour tree based on large language model," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 2024, pp. 392-397.

[46] Y. Liu, L. Palmieri, S. Koch, et al., "Towards human awareness in robot task planning with large language models," arXiv preprint arXiv:2404.11267, 2024.

[47] M. Tantakoun, C. Muise, X. Zhu, "LLMs as planning modelers: A survey for leveraging large language models to construct automated planning models," in *AAAI 2025 Workshop LM4Plan*, 2025.

[48] X. Lin, Y. Wu, H. Yang, et al., "CLMASP: Coupling large language models with answer set programming for robotic task planning," arXiv preprint arXiv:2406.03367, 2024.

[49] D. Bai, I. Singh, D. Traum, et al., "Twostep: Multi-agent task planning using classical planners and large language models," arXiv preprint arXiv:2403.17246, 2024.

[50] V. Pallagani, B. C. Muppasani, K. Roy, et al., "On the prospects of incorporating large language models (LLMs) in automated planning and scheduling (APS)," in *Proceedings of the International Conference on Automated Planning and Scheduling*, 2024, vol. 34, pp. 432-444.

[51] C. Wen, J. Liang, S. Yuan, et al., "How secure are large language models (LLMs) for navigation in urban environments?" arXiv preprint arXiv:2402.09546, 2024.

[52] K. Garg, S. Zhang, J. Arkin, et al., "Foundation models to the rescue: Deadlock resolution in connected multi-robot systems," arXiv preprint arXiv:2404.06413, 2024.

[53] W. Yu, J. Peng, Y. Ying, et al., "MHRC: Closed-loop decentralized multi-heterogeneous robot collaboration with large language models," arXiv preprint arXiv:2409.16030, 2024.

[54] J. P. Macdonald, R. Mallick, A. B. Wollaber, et al., "Language, camera, autonomy! Prompt-engineered robot control for rapidly evolving deployment," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 717-721.

[55] A. Koubaa, A. Ammar, W. Boulila, "Next-generation human-robot interaction with ChatGPT and robot operating system," *Software: Practice and Experience*, vol. 55, no. 2, pp. 355-382, 2025.

[56] J. Li, M. Zhang, N. Li, et al., "Exploring the potential of large language models in self-adaptive systems," in *Proceedings of the 19th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2024, pp. 77-83.

[57] J. Zheng, S. Qiu, C. Shi, et al., "Towards lifelong learning of large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1-35, 2025.

[58] L. Bärmann, R. Kartmann, F. Peller-Konrad, et al., "Incremental learning of humanoid robot behavior from natural interaction and large language models," *Frontiers in Robotics and AI*, vol. 11, p. 1455375, 2024.

[59] K. Ryu, Q. Liao, Z. Li, et al., "Curricullm: Automatic task curricula design for learning complex robot skills using large language models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 4470-4477.

[60] Z. Li, K. Yu, S. Cheng, et al., "League++: Empowering continual robot learning through guided skill acquisition with large language models," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[61] Y. Zeng, Y. Mu, L. Shao, "Learning reward for robot skills using large language models via self-alignment," arXiv preprint arXiv:2405.07162, 2024.

[62] J. Liu, Y. Yuan, J. Hao, et al., "Enhancing robotic manipulation with AI feedback from multimodal large language models," arXiv preprint arXiv:2402.14245, 2024.