LLM as A Robotic Brain: Unifying Egocentric Memory and Control

LLM as A Robotic Brain: Unifying Egocentric Memory and Control

Jinjie Mai, Jun Chen, Bing Li, Guocheng Qian,
Mohamed Elhoseiny, Bernard Ghanem

King Abdullah University of Science and Technology (KAUST)

Abstract

Embodied AI focuses on the study and development of intelligent systems that possess a physical or virtual embodiment (*i.e.* robots) and are able to dynamically interact with their environment. Memory and control are the two essential parts of an embodied system and usually require separate frameworks to model each of them. In this paper, we propose a novel and generalizable framework called **LLM-Brain**: using **L**arge-scale **L**anguage **M**odel as a robotic **brain** to unify egocentric memory and control. The LLM-Brain framework integrates multiple multimodal language models for robotic tasks, utilizing a zero-shot learning approach. All components within LLM-Brain communicate using natural language in closed-loop multi-round dialogues that encompass perception, planning, control, and memory. The core of the system is an embodied LLM to maintain egocentric memory and control the robot. We demonstrate LLM-Brain by examining two downstream tasks: active exploration and embodied question answering. The active exploration tasks require the robot to extensively explore an unknown environment within a limited number of actions. Meanwhile, the embodied question answering tasks necessitate that the robot answers questions based on observations acquired during prior explorations.

1 Introduction

Embodied AI [14], as an emerging area of research, is devoted to studying and developing intelligent systems that possess a physical or virtual embodiment, such as robots, and can dynamically interact with their environment. For many years, reinforcement learning-based methods [5] have been dominating the field of embodied AI.

More recently, significant advances have been made in the field of Internet AI, which focuses on learning from images, videos, and texts obtained from the Internet. The impressive advancements in foundational models [2], e.g. Large Language Model (LLM), Vision Language Model (VLM), have sparked interest in integrating these large models (LM) into existing Embodied AI pipelines.

Unified methods [16, 13, 7, 10, 4] aim to train a single large network capable of generalizing across multimodal inputs and various embodied AI tasks. However, these methods demand substantial resources for joint training across all modalities and tasks.

Preprint. Work in progress

Zero-shot or few-shot methods utilize pre-trained large models directly to take advantage of world knowledge [17] without bells and whistles. Many of them [1, 8, 9] adopt LLMs as task planners to decompose high-level instructions into sub-goals in robotic affordances. These methods assume the robot's capabilities for advanced commands, which might be insufficient for tasks requiring geometric configuration of the environment. Several other methods [11, 18, 19] leverage LLMs for translating natural language into robotic plans and instructions.

Multi-model multimodal methods ensemble multiple large models, such as LLMs and VLMs, to acquire multimodal capabilities for addressing embodied tasks in perception, planning and control. Some existing methods [12, 6] finetune VLMs and LLMs, incorporating them into the reinforcement learning pipeline.

In contrast, our proposed LLM-Brain is a zero-shot framework that employs language-based communication between different LMs. Socratic models [20] design variable-embedded prompt templates and task-specific model interactions, allowing for communication across modalities. On the other hand, our approach employs 'instruction' prompts to assign roles to LLMs, facilitating automatic closed-loop communication and free-form message passing between different LMs. The most relevant work to ours, LM-Nav [15], leverages multiple LMs for vision-language navigation tasks without additional training. However, their reliance on a pre-built topology environment map significantly restricts their ability to generalize to unseen scenes.

To this end, we propose **LLM-Brain**: using **L**arge-scale **L**anguage **M**odel as a robotic **brain** to unify egocentric memory and control. Our contributions can be summarized in three key aspects:

1. We present a versatile multi-model, multimodal framework that leverages the capabilities of large models in a zero-shot fashion, enabling generalizability across various tasks. Our system's components communicate with each other using natural language, forming a closed loop of perception, planning, control, and memory. This approach promotes efficient collaboration between different modalities with explainability.
2. We treat LLM to be the central component of our system. The LLM functions as an intelligent brain, seamlessly integrating temporal robot actions and spatial environmental observations into its memory. The brain can thus control the robot to effectively follow given instructions while relying on its accumulated memory and world knowledge.
3. We benchmark our methods on multiple embodied AI tasks. The proposed framework exhibits strong scalability, allowing for straightforward application with various downstream embodied AI tasks, such as active exploration, embodied question answering, and vision-language navigation.

By developing LLM-Brain, we aim to enhance the adaptability and functionality of multimodal Embodied AI systems, extending their applicability across a wide range of open real-world scenarios.

2 Method

2.1 The pipeline

The robot's exploration powered by LLM-Brain consists of the following steps: Role Initialization, Eye-Nerve Perception, and Brain Reasoning and Control.

Role Initialization. We initialize the three agents by giving them initial prompts. In this way, each agent can understand the role they need to act.

Eye-Nerve Perception. Most VLMs are trained on high-quality images, while egocentric video frames perceived by a robot often suffer from low-quality issues caused by environment scan and

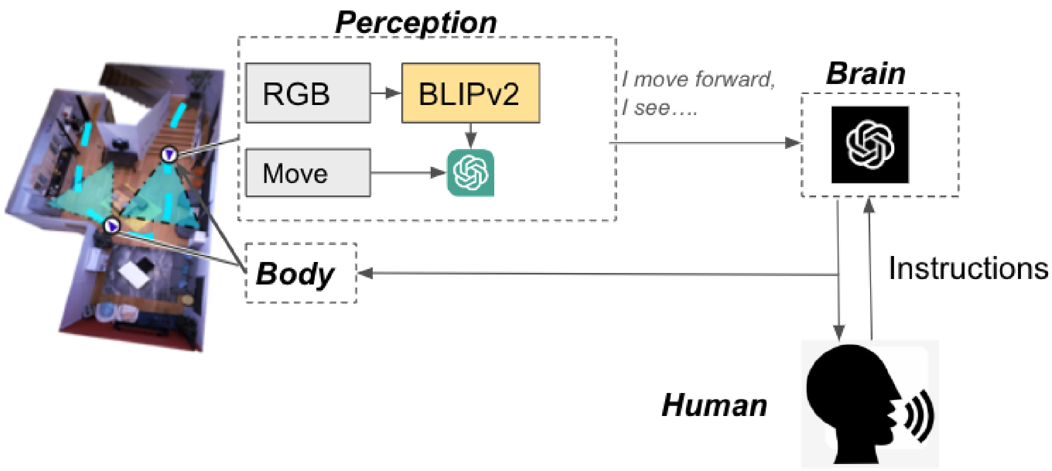

Figure 1: **Illustration of the proposed framework**.

simulator, unusual viewpoints influenced by robot motion, and lack of visual information at some indoor locations. Due to the aforementioned challenges, many advanced VLMs can not accurately extract rich environmental information. Therefore, inspired by ChatCaptioner [21, 3], we design an iterative question-answering pipeline to bridge VLM and LLM. The nerve will continuously ask the eye multiple questions and summarize the dialogue to give a detailed description of the current observation. Together with the robot action received from the robot sensor, the nerve will forward the *perception* message to the brain.

Brain Reasoning and Control. The brain of the robot behaves like the *brain* of humans. According to its memory and predefined instructions, the brain, thanks to the ability of nowadays LLM, can perform instruction-specific reasoning. It plays the role of a leader that analyzes and abstracts the information of a partially observed environment and infers the next action based on its world knowledge without human intervention. To ground the planning obtained by the reasoning into actions, we provide the brain with the context of action space: `move_forward`, `turn_left`, `turn_right`, `stop`. The brain can use these rudimentary textual commands to control the robot.

3 Experiments

As shown in Fig. 2, although our LLM-Brain is built in a zero-shot fashion, it effectively enables the robot to extensively explores an unknown environment.

Limitations. Although our early experiments initially demonstrated the feasibility of our proposed LLM-Brain, we also show that our current method still has many limitations. The most common instances of failure occur when the eye and brain do not coordinate well. For instance, it's also hard for LLM to make a good action decision if the upcoming perception is too coarse. When the text perception only contains high-level semantics and lacks directional information, the brain will get confused and make bad navigation decisions, resulting in a failed or inefficient exploration. Also, since the brain will receive the high-level semantic perception at each timestamp, it can't recognize the same instance that appeared across different observations.

Figure 2: **Illustration of a robot's exploration progress**. The left image illustrates an egocentric video frame captured by the robot's visual sensor, and the right one shows the explored map constructed by the robot.

4 Conclusion

Our proposed LLM-Brain framework represents an interesting step forward in the field of Embodied AI by introducing a versatile, multimodal system that leverages the power of LMs in a zero-shot manner. Through natural language communication and closed-loop dialogues, LLM-Brain efficiently bridges the gap between different modalities, enhancing the robot's ability to perceive, plan, control, and retain memory. The framework's scalability ensures seamless integration with various downstream tasks, including active exploration, embodied question answering, and further tasks like vision-language navigation. As a result, LLM-Brain advances the adaptability and applicability of Embodied AI systems in real-world scenarios, unlocking new potential for intelligent robotic agents. LLM-Brain is intrinsically connected by language, enabling humans to readily gain insights by examining the underlying communication and reasoning. This fosters a robotic framework that offers excellent explainability.

Future research might focus on improving the perception system for enhanced spatial understanding, refining memory mechanisms to support both short-term and long-term information retention, and conducting comprehensive benchmarking against existing reinforcement learning-based baselines within the Habitat challenges.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023.

[4] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[5] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[6] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *arXiv preprint arXiv:2206.08853*, 2022.

[7] Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia Pinel, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pages 175–187. PMLR, 2023.

[8] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

[9] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[10] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

[11] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.

[12] Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards a unified agent with foundation models. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023.

[13] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[14] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[15] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR, 2023.

[16] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[17] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. Technical Report MSR-TR-2023-8, Microsoft, February 2023.

[18] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Chatgpt empowered long-step robot control in various environments: A case application. *arXiv preprint arXiv:2304.03893*, 2023.

[19] Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023.

[20] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. So-

cratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[21] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023.