Text to Action: Large Language Models into Multimodal Agents

Yuldashev Izzatillo
Department of Computer Engineering
Pukyong National University
yuldashev.dev@gmail.com

Abstract

Multimodal agents, capable of integrating language, vision, and sensor data to perceive, reason, and act, are reshaping robotics, autonomous systems, and human-computer interaction. Recent advances in Large Language Models (LLMs) have accelerated progress in these fields by enabling semantic grounding, contextual awareness, and adaptive task execution. This report analyzes the evolving role of LLMs in three influential works: *SayCan*, *PaLM-E*, and *Gato*. Through these case studies, we explore how LLMs unify multimodal inputs, enhance generalization, and enable real
world problem-solving.

Introduction

Multimodal agents address the challenge of unifying heterogeneous streams of information such as text, images, and sensor data to execute complex tasks in real-world environments [1]. In a typical scenario, a household robot may be tasked with organizing a bookshelf: it must interpret its surroundings through visual inputs, identify objects, and execute precise physical manipulations [2]. The advent of Large Language Models (LLMs) has catalyzed a paradigm shift in how these agents process and integrate data [3]. By mapping raw sensor inputs to semantically rich representations, LLMs ground perceptual understanding in language-driven reasoning [4]. This fusion of perception and language enables systems to handle increasingly flexible and unstructured tasks, such as real-time decision-making in dynamic environments.

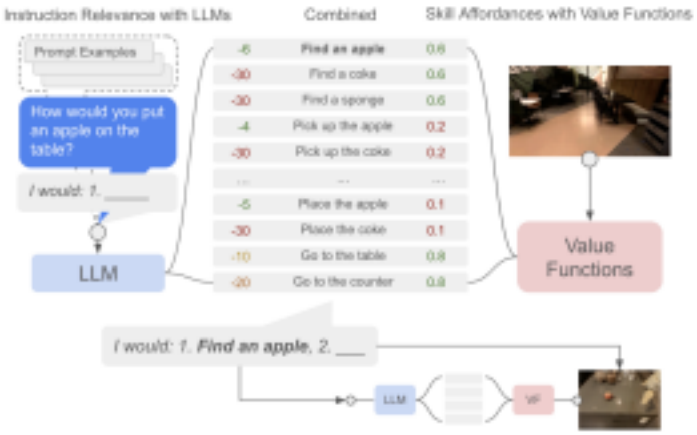

Figure 1.

Key LLM-driven approaches in Multimodal Agents

A prominent example of this approach is Google's SayCan [5]. In SayCan, an LLM interprets high-level instructions and generates step-by-step action plans for the robot to execute. This division of labor positions the language model as a cognitive core: (Figure 1) it translates abstract requests, such as "*How would you put an apple on the table*?" into actionable routines, for example, go to the table, find an apple, and pick the apple. The key insight is that natural language instructions serve as the framework for planning and reasoning, while low-level motion control remains the responsibility of specialized hardware controllers. However, a critical limitation arises in aligning language-generated plans with a robot's physical constraints. For instance, an LLM might propose actions that exceed the robot's kinematic or sensor capabilities.

The PaLM-E model [6] exemplifies the potential of large-scale LLMs in multimodal contexts. By integrating visual and sensor data directly into the architecture of a large language model, PaLM-E is designed to reason simultaneously about real-time sensory inputs and user instructions. For example, when prompted with a query like "*Is the blue block on the left?*" while processing a live camera feed, PaLM-E uses its language-based reasoning to analyze the spatial relationships between objects. It then directs a physical robot to adjust their positions if necessary. A key finding from this work is the enhanced task robustness achieved through model scaling: larger LLMs, such as the 562B-parameter PaLM variant, demonstrate fewer hallucinations and more coherent cross modal reasoning compared to smaller models.

Another milestone in this domain is Gato [7], a generalist agent that processes text, images, and actions through a single transformer architecture. Gato's versatility lies in its ability to seamlessly switch between tasks as varied as playing Atari games and controlling robotic arms. Instead of building separate models for each modality, Gato unifies diverse input streams by tokenizing them into a shared representation space, enabling a single transformer to interpret and act on multimodal data. This architecture embodies the promise of general-purpose agents capable of rapid adaptation to novel tasks.

In addition to reviewing existing literature, I conducted a small-scale experiment to observe how GPT-4o model transforms everyday household requests into step-by-step robot instructions. Each prompt such as *"I spilled my drink, can you help?"* was translated into a concise plan for navigating, identifying objects, and executing physical tasks. These results (attached in the JSON file) illustrate how LLMs can serve as high-level planners, mapping natural language instructions to feasible action sequences in real-world scenarios.

Conclusion

The integration of language-driven reasoning with multimodal perception holds transformative potential for robotics and AI. Systems like SayCan, PaLM-E, and Gato illustrate how LLMs enhance agents' ability to interpret complex instructions, adapt dynamically, and execute precise tasks in real time. However, advancing the field requires overcoming three key hurdles: data scarcity in robotics compared to language domains, computational inefficiency in scaling multimodal models, and safety assurance to mitigate risks of LLM-guided physical actions.

References

1. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal Machine Learning: A Survey and Taxonomy. IEEE TPAMI. [DOI:10.1109/TPAMI.2018.2798607]

2. Amodei, D., Olah, C., Steinhardt, J., et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

3. Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives. arXiv:2005.01643.

4. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML. [DOI:10.48550/arXiv.2103.00020]

5. Ahn, M., Brohan, A., Brown, N., et al. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691.

6. Driess, D., Xia, F., Sajjadi, M. S., et al. (2023). PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378.

7. Reed, S., Zolna, K., Parisotto, E., et al. (2022). A Generalist Agent. arXiv:2205.06175.