Enhancing Human-Robot Interaction with Multimodal Large Language Models

MATTHEW GRIMALOVSKY, HEDAYA WALTER, and JORGE ORTIZ, Rutgers University, USA

This paper presents initial work on integrating a large language model (LLM) with structured multimodal perceptions, including computer vision and sensor data, to enable more advanced contextual reasoning for human-robot interaction (HRI) systems. We implement an approach that encodes multimodal input into a structured log format for the LLM to interpret, demonstrating promising skills in querying object locations, assessing behavioral patterns, and logical reasoning. For instance, when asked about a misplaced phone, the system can infer its likely placement on a specific table by correlating action timestamps across spaces. This initial integration thus indicates the significant yet underscored potential of large language models to enhance collaborative human-robot interaction through integration with additional modalities while needing future improvement in areas like personalization and real-world robustness.

CCS Concepts: • **Computer systems organization** → **Robotic autonomy**; • **Computing methodologies** → **Logical and relational learning**; **Generative and developmental approaches**; • **Hardware** → **Sensor applications and deployments**; • **Human-centered computing** → Usability testing.

Additional Key Words and Phrases: human-robot interaction, multimodal large language models, contextual reasoning, ambient intelligence, sensor data integration, multimodal perception, logical reasoning, personalization, real-world robustness

ACM Reference Format:
Matthew Grimalovsky, Hedaya Walter, and Jorge Ortiz. 2024. Enhancing Human-Robot Interaction with Multimodal Large Language Models. In *HRI '24: 19th Annual ACM/IEEE International Conference on Human Robot Interaction (HRI), March 11-15, 2024, Boulder, Colorado.*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Large language models (LLMs) have shown increasing promise for enhancing robots' natural language and communication abilities for more intuitive human-robot interaction (HRI). However, additional multimodal integration presents opportunities to enhance contextual understanding and reasoning. By combining LLMs' language generation capabilities with computer vision, audio, inertial sensors, and other modalities, robots may move towards more experientially meaningful communication, recognition of user needs and emotions, and increasingly helpful situated interactions.

This paper presents initial work combining LLMs with structured multimodal data as a step toward more advanced ambient reasoning in HRI systems. We implement an approach that encodes computer vision classifications and sensor logs into a common representation, which an LLM processes to produce insightful narratives and responses. Our results demonstrate promising interpretive skills, including location inference, assessment of behavioral patterns, and logical reasoning about implausible scenarios. Specifically, our system showcases the capability to pinpoint the likely location of a misplaced phone by correlating action timestamps across different monitored spaces. Furthermore, when queried about an individual's work habits, the LLM deduces periods of activity and inactivity based on actions suggestive of disengagement from desk work. These promising ambient reasoning skills highlight the potential of LLMs to extract useful insights from multimodal data.

SO-HRI-24, March 11–15, 2024, Denver, CO

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

These interpretive skills could allow robots to interact more intuitively by grounding conversations in shared environmental knowledge and practical user needs. For instance, by determining periods of focused work versus disengagement, a robot assistant could politely offer help when a user appears distracted rather than interrupting important tasks. Furthermore, deducing and communicating object locations could enable more natural directives, e.g., suggesting to "check near the doorway table" for a misplaced item. While conditional and requiring extensive real-world adaptation, these promising ambient reasoning capacities indicate LLMs' untapped potential for enhancing situated human-robot collaboration and assistance. Additional multimodal integration may enrich contextual understanding further, moving towards HRI that is not just interactive but cooperative, helpful, and responsive to users' implicit situational needs.

1 MOTIVATION: MULTIMODAL SENSEMAKING AND LANGUAGE MODELS

Integrating Large Language Models (LLMs) with multimodal sensor data presents an opportunity to enhance human-robot interaction (HRI) by enabling robots to develop a more comprehensive understanding of their environment and users. By leveraging LLMs' natural language processing capabilities to interpret and contextualize data from various sensors, robots can better comprehend and respond to the complexities of human behavior and emotions. While the examples provided in this paper focus on understanding contextual information, such as object location and human activity, this is an essential foundation for more advanced reasoning in HRI. Accurately interpreting context allows robots to better understand and respond to users' needs, paving the way for more intuitive and meaningful interactions.

Sensemaking in Smart Environments Integrating large language models (LLMs) with multimodal sensor data in smart environments presents an opportunity to enhance contextual understanding and sensemaking, akin to the exploration and organization capabilities demonstrated by Sensecape [8]. This integration allows for the interpretation and meaningful interaction with the physical world through sensor data, offering a nuanced view of our surroundings and enabling dynamic, intelligent responses within these environments.

Physical-World Grounding Grounding LLMs in the physical world enhances robots' and systems' environmental interaction and understanding. KNOWNO [5] provides a framework for LLM-based planners to recognize knowledge gaps and seek help, using conformal prediction for statistical task completion guarantees with minimal human intervention. LLM+A [2] prompts LLMs for robotic tasks, predicting actions' consequences and generating affordance values to improve plan feasibility. Xu et al. [11] demonstrate leveraging LLMs for task completion by interpreting sensor data, aiming to embed intelligence into cyber-physical systems. GLAM [1] aligns LLMs with environments via online reinforcement learning, updating policies for better goal-solving performance.

Effective Communication and Interaction Dynamics Effective communication and interaction dynamics are essential for engaging HRI. LLMs enhance dialogues with advanced language capabilities, while multimodal learning models interpret non-verbal cues like gestures, enriching interactions. This synergy supports well-timed responses, demonstrated by recent studies [6, 10] and applications such as robot-assisted feeding [4], emphasizing the importance of multimodal strategies for effective communication.

Our work complements prior advancements in LLM grounding by integrating visual and sensor data to refine activity interpretation in HRI. This methodology deepens robots' comprehension of environmental contexts and human behaviors, providing precise contextual insights.

2 INTEGRATION OF LLMS AND MULTIMODALITY FOR INTERACTION

Our deployment showcases two key applications of our system in real-world contexts: 1) **Activity Logging and Interpretation:** Leveraging visual and sensory data, our system logs activities, exemplified by locating a misplaced phone through data pattern analysis without direct visual detection (Figure 1). This underscores our system's capability to derive and narrate activities from multimodal inputs. 2) **Query Response and Logical**

Reasoning: The system demonstrates refined reasoning, answering queries on an individual's work habits by interpreting spatial and usage data and logically addressing unlikely scenarios, such as the inquiry about an 'element' on a table, evidencing adept context understanding (Figure 2). These scenarios illustrate our system's adeptness at parsing multimodal data for insightful narratives and responses, indicating its utility for assistive intelligence. Transitioning to real-world deployment highlights personalization, robustness, and safety challenges, yet these deployments point towards a viable trajectory for ambient intelligence applications.

In this deployment, visual and sensory scenes are structured for interpretation by an LLM (i.e., LLama-70B [9]), with video processed by Squeezenet [3]. We fine-tune the model using Charades, a dataset centered around human activity recognition of indoor environments [7]. Multimodal data is annotated with location and time stamps, ensuring precise activity registration.

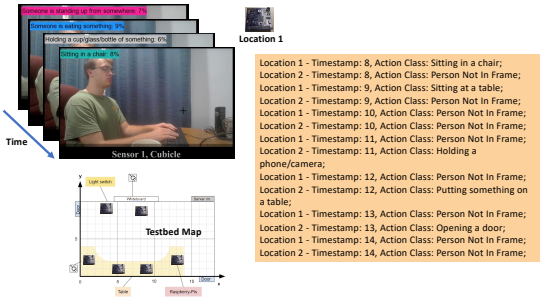

Fig. 1. The figure depicts a system that uses multimodal data to log activities in a space, which a language model then interprets to create narratives and answer queries about the events.

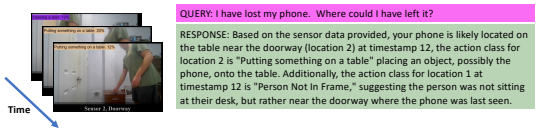

Fig. 2. We ask about a lost phone, and the LLM pinpoints its likely location near the doorway on a table at timestamp 12, using sensor data to infer the phone was placed there when the person was not at their desk.

Figure 1 illustrates our integrated system's capability to log and interpret activities within a space over time. The visual and sensor data are synchronized to create a detailed log of actions. A language model processes this log, extracting narratives and responding to queries about the recorded activities. The model is then able to identify the location of a misplaced phone despite the phone never being directly recognized in the data log. The language model infers the phone's location based on the correlation of actions logged at various timestamps, highlighting the nuanced understanding of our language model.

Further demonstrating the model's discernment, when queried about the work habits of the monitored individual, the language model deduces periods of activity and inactivity. The LLM surmises that the person was not working when absent from their desk or when engaged on their phone, indicating non-work-related behavior. To assess the model's consistency and grasp of practical scenarios, we posed improbable queries, such as whether the individual had placed an 'element'—a term ambiguously referring to a chemical element—on the table. The language model aptly responded that such an event was implausible, reflecting its logical reasoning capabilities and understanding that such items are neither transportable in a typical setting nor expected to be found in an office environment.

Our proposed system demonstrates promising capabilities in interpreting multimodal data to produce nuanced narratives and responses, which are integral for intuitive interactions. It showcases discernment in ambient reasoning, location inference, and assessment of behavioral patterns, highlighting the potential for assistive intelligence applications. However, realizing this potential requires overcoming challenges in real-world integration, such as the need for greater personalization, robustness, and adaptable safety mechanisms when transitioning from controlled environments with structured inputs to unconstrained human interaction.

3 CONCLUSION

In this paper, we have taken an initial step toward exploring the potential for Large Language Models (LLMs) and multimodal learning to enhance cognitive capabilities for Human-Robot Interaction (HRI). Our goal was to investigate integrating these methods as a promising new direction to facilitate more natural and meaningful HRI. Our proposed system demonstrates promising capabilities in interpreting multimodal data to produce nuanced narratives and responses—skills integral for intuitive interactions. We believe combining LLM's sophisticated language processing and multimodality's rich perceptual input lays the groundwork for advancing robots' understanding of contextual and emotional cues. While this initial integration shows promising capabilities, real-world deployment would involve significant challenges still to be addressed.

4 ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (NSF) as part of the Center for Smart Streetscapes, under NSF Cooperative Agreement EEC-2133516.

REFERENCES

[1] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning* (<conf-loc>, <city>Honolulu</city>, <state>Hawaii</state>, <country>USA</country>, </conf-loc>) *(ICML'23)*. JMLR.org, Article 150, 38 pages.

[2] Guangran Cheng, Chuheng Zhang, Wenzhe Cai, Li Zhao, Changyin Sun, and Jiang Bian. 2024. LLM+A: Grounding Large Language Models in Physical World with Affordance Prompting. https://openreview.net/forum?id=cbVnJa4l2o

[3] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR* abs/1602.07360 (2016). arXiv:1602.07360 http://arxiv.org/abs/1602.07360

[4] Jan Ondras, Abrar Anwar, Tong Wu, Fanjun Bu, Malte Jung, Jorge Jose Ortiz, and Tapomayukh Bhattacharjee. 2023. Human-Robot Commensality: Bite Timing Prediction for Robot-Assisted Feeding in Groups. In *Proceedings of The 6th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 205)*, Karen Liu, Dana Kulic, and Jeff Ichnowski (Eds.). PMLR, 921–933. https://proceedings.mlr.press/v205/ondras23a.html

[5] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. 2023. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. arXiv:2307.01928 [cs.RO]

[6] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300867

[7] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *CoRR* abs/1604.01753 (2016). arXiv:1604.01753 http://arxiv.org/abs/1604.01753

[8] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM. https://doi.org/10.1145/3586183.3606756

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971 cite arxiv:2302.13971.

[10] Tong Wu, Nikolas Martelaro, Simon Stent, Jorge Ortiz, and Wendy Ju. 2021. Learning When Agents Can Talk to Drivers Using the INAGT Dataset and Multisensor Fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 133 (sep 2021), 28 pages. https://doi.org/10.1145/3478125

[11] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative AI: Making LLMs Comprehend the Physical World. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications (HOTMOBILE '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3638550.3641130

Received 10 February 2024; revised 4 March 2024; accepted 4 March 2024