Latest updates: https://dl.acm.org/doi/10.1145/3776734.3796242

EXTENDED-ABSTRACT

Multimodal Large Language Models for Real-Time Situated Reasoning

GIULIO ANTONIO ABBO

SENNE LENAERTS

TONY BELPAEME

PDF Download
3776734.3796242.pdf
19 March 2026
Total Citations: 0
Total Downloads: 0

Published: 16 March 2026

Citation in BibTeX format

HRI '26: 21st ACM/IEEE International Conference on Human-Robot Interaction
March 16 - 19, 2026
Scotland, Edinburgh, UK

Conference Sponsors:
SIGAI
SIGCHI

HRI Companion '26: Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (March 2026)
https://doi.org/10.1145/3776734.3796242
ISBN: 9798400723216

Multimodal Large Language Models for Real-Time Situated Reasoning

Giulio Antonio Abbo
Ghent University – imec
Ghent, Belgium
giulioantonio.abbo@ugent.be

Senne Lenaerts
Ghent University – imec
Ghent, Belgium
senne.lenaerts@ugent.be

Tony Belpaeme
Ghent University – imec
Ghent, Belgium
tony.belpaeme@ugent.be

Abstract

In this work, we explore how multimodal large language models can support real-time context- and value-aware decision-making. To do so, we combine the GPT-4o language model with a TurtleBot 4 platform simulating a smart vacuum cleaning robot in a home. The model evaluates the environment through vision input and determines whether it is appropriate to initiate cleaning. The system highlights the ability of these models to reason about domestic activities, social norms, and user preferences and take nuanced decisions aligned with the values of the people involved, such as cleanliness, comfort, and safety. We demonstrate the system in a realistic home environment, showing its ability to infer context and values from limited visual input. Our results highlight the promise of multimodal large language models in enhancing robotic autonomy and situational awareness, while also underscoring challenges related to consistency, bias, and real-time performance.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Computer systems organization** → **Robotics**; • **Human-centered computing** → *Human computer interaction (HCI).*

Keywords

Embodied, Reasoning, LLM, VLM, Vision, TurtleBot, Vacuum Cleaner, Home Environment, Social Robot, Human-Robot Interaction

ACM Reference Format:
Giulio Antonio Abbo, Senne Lenaerts, and Tony Belpaeme. 2026. Multimodal Large Language Models for Real-Time Situated Reasoning. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3776734.3796242

1 Introduction

Strong human-robot interaction relies on effective environmental modelling, as this provides the foundation for informed decision-making. Robots must dynamically perceive and interpret their surroundings, including objects, obstacles, and human actions, while accounting for uncertainties such as sensor limitations and unexpected changes. Without a robust and adaptive model, robots risk inefficiency or even unsafe behaviour, undermining their utility in real-world applications.

This work is licensed under a Creative Commons Attribution 4.0 International License.
HRI Companion '26, Edinburgh, Scotland, UK
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3796242

Figure 1: The robot deployed in a living room, and an image from the robot's perspective, showing the unusual angle and the partially obstructed view.

Yet, modelling real-life environments presents significant challenges. For instance, in a home, the complexity of unstructured scenarios, with variables such as values and personal preferences, family dynamics, and traditions, can quickly overwhelm traditional approaches. Scaling these models to handle diverse, dynamic settings often demands meticulously handcrafted rules, which reduce the flexibility and scalability of these approaches.

Recent advancements in Large Language Models offer a compelling opportunity to equip household robots with a higher level of decision-making. These models have demonstrated impressive capabilities not only in mimicking [7] contextual reasoning [6] and common sense understanding, but also in guiding robots' behaviour [8]. When combined with multimodal inputs, such as images or sensor data, multimodal LLMs can interpret a scene and evaluate whether specific actions align with human preferences, norms, and routines [1, 2]. Furthermore, the text-based nature of these approaches produces a human-readable decision trace that explains the rationale behind each decision taken, and it can easily be integrated with additional users' indications and corrections. However, despite their potential, multimodal LLMs as decision-makers might struggle with real-time performance. Indeed, it is necessary to balance the quality and completeness of the reasoning process result with its duration. Without careful prompting, LLMs can be inconsistent in their decision outcomes, which is undesirable, especially in social robot applications.

This work explores how multimodal LLMs can be used to reason about values and preferences in real time. To do so, we combine an LLM-based reasoning framework with a TurtleBot 4 robot acting as a value-aware vacuum cleaner in a home setting.

Autonomous vacuum cleaners have become a common presence in modern households, offering convenience through scheduled or reactive cleaning routines. These robots typically rely on a combination of sensors, pre-programmed logic, and, in some cases,

basic learning algorithms to navigate environments, avoid obstacles, and perform their cleaning tasks. While increasingly effective in manoeuvring through rooms, these systems still fall short in understanding the context in which cleaning occurs. In real-world scenarios, determining whether or not cleaning is appropriate involves more than obstacle avoidance and scheduling. For instance, a person watching a movie in the living room might prefer silence, while a recently used kitchen might benefit from immediate cleaning to avoid insects taking over. These nuanced decisions depend not only on detecting objects or motion, but also on understanding activities, routines, and values that govern comfort in a shared space. A robot that can *reason* about ongoing activities and routines would be better aligned with users' expectations and preferences.

In our implementation, the robot uses image input to interpret its environment and make contextually appropriate decisions about whether to clean, wait, or return to its docking station. We observe the system in action in a real home environment, identifying strengths and limitations in the robot's contextual understanding and value alignment.

Through this work, we aim to demonstrate that multimodal LLMs provide a promising foundation for more intelligent, empathetic, and situationally aware household robots.

2 Implementation

The robot used is a TurtleBot 4, an open-source platform for research created by Open Robotics and Clearpath Robotics[1]. The robot is built on top of the iRobot *CREATE 3* mobile base, which is based on the chassis of a robot vacuum cleaner. The robot has a rich sensor suite, with our research relying on the *OAK-D-PRO* RGBD camera to collect images from the robot's point of view.

The system driving the robot has three states: *observation*, *cleaning*, and *docking* mode, with the VLM (Vision Language Model) deciding when to transition from one state to another.

In **observation mode**, the robot surveys its environment to collect a series of images. From these salient features are extracted and passed to the model, which decides which action to take. Specifically, the robot starts rotating on itself and collects a series of 10 images at regular intervals, giving a 180-degree overview of the surrounding space at one frame per second. A VLM is used to extract key elements from the image, focusing on the presence of people or pets, human activities, and contextual clues. The list of relevant elements and the original images are then used by the same VLM to perform the decision-making process. The system prompt that instructs the model how to behave is composed of three parts. First, it defines a role: "a value-aware vacuum cleaner"; then it defines the objective of "maintain cleanliness while respecting the homeowner's values, ensuring a comfortable and harmonious living environment." Finally, it describes the three modes in detail, including the capabilities and behaviour of the robot in each mode. The model is prompted to reason step-by-step on the following aspects: value alignment, "creating a safe and enjoyable environment for the house owners", time context – with the current time provided in the prompt – the action choice and its consequences, the rationale, and how the decision reflects the robot's core purpose of cleaning. The reasoning trace is then fed back into the model to take a final decision, and, finally, the entire process is summarised into a short, user-friendly paragraph to add a level of explainability. We relied on OpenAI's GPT-4o model, which accepts both image and text inputs and produces text-based outputs.

In **cleaning mode**, the robot simulates a vacuum cleaning robot by driving in a straight line until the proximity sensors of the robot detect a nearby object, upon which the robot slows down. If a collision is detected, it turns in a random direction and resumes cleaning. In the meantime, the robot collects 3 images every 0.5 seconds and uses the VLM to interpret the environment and evaluate whether any values are potentially violated. This determines whether the current cleaning task should be interrupted and the robot should transition into observation mode using a reasoning process similar to the previous mode.

Docking mode instructs the robot to return to its charging station. While in the current implementation this mode stops the robot, in the future it could be used to instruct the robot to go to another room, if the current room is busy, before finally going to the charging station.

ROS 2, the robot operating system [4], is used to implement the functionalities described. This middleware handles the communication between specialised nodes which can run on different machines. The TurtleBot 4 comes with several nodes already running, to collect images from the camera and controlling the robot. The system presented runs as a Python node on another computer, connected to the same wireless network as the robot. In addition, a graphical interface developed with PyQT5 displays the reasoning process and allows control of the robot.

3 Observations

Evaluating the value alignment of the system responses depends strongly on the subjective preferences of the people involved. As an initial evaluation, we opted for a qualitative observation of the system responses – without a quantitative evaluation – first using a small dataset, after which the system was evaluated in a realistic environment.

To observe the VLM's behaviour without the additional complications of the robot's hardware, the system was used with a set of images instead of the live feed form the robot's camera. These images are manually chosen from the YouHome Activities of Daily Living dataset [5], which comprises 31 common daily activities from 20 users, in different areas of a home. We chose images that could allow observing how well the model detects people and infers which activities they are partaking in. In addition, we also tried scenarios in which the room appears empty, or someone enters and then leaves the room in a quick succession.

By deploying the system in a home, we obtained a more comprehensive idea of the overall system performance. We put the robot in a living room, and observed its decision-making process under three conditions: when someone is watching a movie, using a phone, and around a pet dog.

In the first case, shown in Figure 1, the robot was positioned near a sofa, on top of which a person was sitting watching a television show. The robot, even from the low and partially obstructed view angle, was able to correctly recognise the activity, and decide to defer the cleaning using the noises produced as a justification. We

[1] https://clearpathrobotics.com/turtlebot-4/

then used the same setting, but with the person focused on their phone and no movie playing. In this case, the VLM reasoned that using the phone does not require silence, and to balance the users' comfort with the purpose of cleaning, it decided to start cleaning. Finally, when a dog was introduced to the robot, the system noted that the pet appeared to be moving a lot in its vicinities, and thus cleaning was not advisable for the pet's physical safety. The system also reflected that the loud noises could scare the pet, showing that its reasoning capabilities can take into account the mental wellbeing of those involved as well as their safety.

Our work has several limitations. We tested and decided not to adopt reasoning models and several prompting techniques, including learning from examples, due to the additional delays that longer prompts imply. In the future, we intend to further explore these possibilities, including the use of locally hosted models, which would alleviate the privacy issues tied to using pictures from within the home environment. The inherent bias [3] of the multimodal LLMs needs to be further studied and mitigated using additional information in the prompt, such as users' habits and preferences, which can be expressed by the users themselves through voice or text and integrated into the prompt.

4 Conclusion

We tested how well state-of-the-art multimodal language models can reason on personal preferences and values in real time. To do so, we presented an implementation of a vacuum cleaning robot based on the TurtleBot 4 platform and GPT-4o. The robot collects images from its camera and uses them to decide whether to start cleaning or not. We deployed the robot in a living room and observed its behaviour. Overall, the language model, despite the unusual point of view, the partially obstructed images, and the limited time resolution, was successful in inferring the activities taking place in the room, in using this information to reason on the values involved, and balancing the values of cleanliness and respect for the people involved.

This work aims to demonstrate that large language models, when integrated into embodied systems, go beyond task execution and begin to reason about the values and context in which their actions take place, marking a significant step toward more socially aware and adaptable domestic robots.

Acknowledgments

Funded by the Horizon Europe VALAWAI project (grant agreement number 101070930).

References

[1] Giulio Antonio Abbo and Tony Belpaeme. 2025. I Was Blind but Now I See: Implementing Vision-Enabled Dialogue in Social Robots. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE Press, 1176–1180.

[2] Giulio Antonio Abbo and Tony Belpaeme. 2025. Vision Language Models as Values Detectors. In *Value Engineering in Artificial Intelligence*, Nardine Osman and Luc Steels (Eds.). Springer Nature Switzerland, Cham, 76–86.

[3] Giulio Antonio Abbo, Tony Belpaeme, and Micol Spitale. 2026. Concerns and Values in Human-Robot Interactions: A Focus on Social Robotics. *International Journal of Social Robotics* 18, 1 (14 Jan 2026), 4. doi:10.1007/s12369-025-01351-1

[4] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. 2022. Robot Operating System 2: Design, architecture, and uses in the wild. *Science Robotics* 7, 66 (2022), eabm6074. arXiv:https://www.science.org/doi/pdf/10.1126/scirobotics.abm6074 doi:10.1126/scirobotics.abm6074

[5] Junhao Pan, Zehua Yuan, Xiaofan Zhang, and Deming Chen. 2022. YouHome System and Dataset: Making Your Home Know You Better. In *2022 IEEE International Symposium on Smart Electronic Systems (iSES)*. 414–420. doi:10.1109/iSES54909.2022.00091

[6] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with Large Language Models, a Survey. arXiv:2407.11511 [cs.AI] https://arxiv.org/abs/2407.11511

[7] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.06941 [cs.AI] https://arxiv.org/abs/2506.06941

[8] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. 2025. Gemini Robotics: Bringing AI into the Physical World. arXiv:2503.20020 [cs.RO] https://arxiv.org/abs/2503.20020

Received 2025-12-08; accepted 2026-01-12