用于实时情境推理的多模态大语言模型

Giulio Antonio Abbo、Senne Lenaerts 和 Tony Belpaeme

IDLab-AIRO，根特大学 – *imec*，比利时根特

giulioantonio.abbo@ugent.be

arXiv:2602.01880v1 [cs.RO] 2 Feb 2026

摘要

在本研究中，我们探讨多模态大语言模型如何支持实时、兼顾情境与价值的决策。为此，我们将 GPT-4o 语言模型与 TurtleBot 4 平台相结合，模拟住宅中的智能扫地机器人。模型通过视觉输入评估环境，并判断是否适合开始清洁。该系统展示了此类模型理解家务活动、社会规范和用户偏好，并据此作出细致决策的能力；这些决策符合相关人员的价值观，例如整洁、舒适和安全。我们在真实感十足的住宅环境中演示了该系统，展现其从有限视觉输入中推断情境和价值的能力。结果凸显了多模态大语言模型在提升机器人自主性和情境感知方面的潜力，同时也指出了其在一致性、偏见和实时性能方面面临的挑战。

1 引言

有效的环境建模是实现良好人机交互的基础，因为它为明智决策提供了依据。机器人必须动态感知并解读周围环境，包括物体、障碍物和人类行为，同时考虑传感器限制和意外变化等不确定因素。如果缺乏稳健且自适应的模型，机器人就可能效率低下，甚至出现不安全行为，从而削弱其在现实应用中的实用性。

然而，对现实环境进行建模面临着重大挑战。例如，在住宅中，非结构化场景错综复杂，其中价值和个人偏好、家庭关系以及传统等变量，可能很快就会让传统方法不堪重负。

于2025年12月提交至 *interactivity* 赛道，参加第21届 ACM/IEEE 国际人机交互会议，并于2026年1月获接收。



图1：部署在客厅中的机器人，以及一张机器人视角的图像，展现了不寻常的拍摄角度和部分受遮挡的视野。

要扩展这些模型，使其能够应对多样且动态的环境，往往需要精心手工编写规则，从而降低这些方法的灵活性和可扩展性。

近期，大语言模型的发展为提升家用机器人的决策能力提供了令人瞩目的契机。这些模型不仅展现出模仿 [7] 情境推理 [6] 和常识理解的出色能力，还能够指导机器人的行为 [8]。结合图像或传感器数据等多模态输入后，多模态大语言模型便可解读场景，并评估特定行动是否符合人类偏好、规范和日常习惯 [1, 2]。此外，这类方法以文本为基础，能够生成可供人理解的决策轨迹，解释每项决策背后的理由，而且还能轻松整合用户提供的补充指示和修正。然而，尽管多模态大语言模型前景广阔，将其用作决策者时仍可能难以满足实时性能要求。事实上，必须在推理过程结果的质量与完整性和耗时之间取得平衡。如果缺乏审慎的提示，大语言模型的决策结果可能前后不一致，这种情况并不可取，在社交机器人应用中尤其如此。

本研究探讨如何利用多模态大语言模型实时推理价值和偏好。为此，我们将基于大语言模型的推理框架与一台 TurtleBot 4 机器人相结合，使其充当

住宅环境中具备价值感知能力的吸尘器。

自主吸尘器已成为现代家庭中的常见设备，通过定时或按需执行清洁日常任务，为人们带来便利。这些机器人通常依靠传感器、预编程逻辑，有时还会结合基础学习算法，在环境中导航、避开障碍物并完成清洁任务。尽管它们在房间内移动时已愈发灵活，但仍无法充分理解清洁发生时的具体情境。在真实场景中，判断是否适合清洁，不能只看避障和时间安排。例如，在客厅看电影的人可能希望保持安静，而刚使用过的厨房则可能需要立即清洁，以免招来昆虫。这些细致的决策不仅取决于检测物体或运动，还取决于理解共享空间中影响舒适度的活动、日常习惯和价值观。能够对推理正在进行的活动和日常习惯的机器人，将更能符合用户的期望和偏好。

在我们的实现中，机器人利用图像输入来理解其环境，并根据具体情境判断是进行清洁、等待，还是返回停靠站。我们在真实家庭环境中观察系统运行，分析机器人在情境理解和价值对齐方面的优势与局限。

通过这项工作，我们旨在展示多模态大语言模型能够为打造更智能、更善解人意且更具情境感知能力的家用机器人提供有前景的基础。

2 实现

所用机器人为 TurtleBot 4，这是由 Open Robotics 和 Clearpath Robotics[1] 开发的开源研究平台。该机器人基于 iRobot*CREATE 3* 移动底盘构建，其底盘采用扫地机器人的机身设计。机器人配备了丰富的传感器套件；我们的研究依靠*OAK-D-PRO* RGBD 摄像头，从机器人的视角采集图像。

驱动机器人的系统有三种状态：观察、清洁和停靠模式，由视觉语言模型（VLM）决定何时从一种状态切换到另一种状态。

在**观察模式**下，机器人会巡视周围环境，收集一系列图像。从中提取显著特征并传递给模型，由模型决定采取何种行动。具体来说，机器人会绕自身旋转，并以固定间隔收集 10 张图像，以每秒一帧的速度，对周围空间进行 180 度概览。视觉语言模型会从图像中提取关键元素，重点关注是否存在人或宠物、人类活动以及情境线索。随后，同一视觉语言模型会利用相关元素列表和原始图像进行决策。指导模型行为的系统提示词由三个部分组成。首先，它定义了一个角色："一台重视价值的吸尘器"；然后，它明确了目标："在保持整洁的同时尊重住宅主人所重视的价值，确保营造舒适、和谐的居住环境。"最后，它详细描述三种模式，包括机器人在每种模式下的能力和行为。提示模型逐步推理以下方面：价值对齐、"为住宅主人营造安全愉悦的环境"、时间情境——提示词中会提供当前时间——行动选择及其后果、决策依据，以及该决策如何体现机器人清洁这一核心使命。之后，推理轨迹会被反馈给模型，以便其作出最终决策；最后，整个过程会被概括成一段简短、易于用户理解的文字，以增强可解释性。我们采用了 OpenAI 的 GPT-4o 模型，该模型能够接收图像和文本输入，并生成文本输出。

在**清洁模式**下，机器人会模拟扫地机器人，沿直线行驶，直到机器人的接近传感器检测到附近有物体，此时机器人会减速。如果检测到碰撞，机器人就会随机转向并继续清洁。与此同时，机器人每 0.5 秒采集 3 张图像，并使用视觉语言模型解读环境，评估是否有任何价值可能受到侵犯。这将决定是否应中断当前清洁任务，以及机器人是否应通过与上一模式类似的推理过程切换到观察模式。

停靠模式 指示机器人返回充电站。在当前实现中，此模式会让机器人停下；未来则可用于指示机器人前往另一个房间（如果当前房间正忙），然后再最终返回充电站。

本文使用 ROS 2（机器人操作系统） [4], 实现所述功能。该中间件负责处理可在不同机器上运行的专用节点之间的通信。TurtleBot 4 已预装多个节点，用于采集摄像头图像并控制机器人。本文介绍的系统作为 Python 节点运行在另一台计算机上，该计算机与机器人连接至同一无线网络。此外，使用 PyQT5 开发的图形界面会显示推理过程，并支持对机器人的控制。

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

3 观察结果

对系统响应的价值对齐进行评估，在很大程度上取决于相关人员的主观偏好。作为初步评估，我们选择先使用一个小型数据集对系统响应进行定性观察（不进行定量评估），随后在真实环境中评估该系统。

为了在不受机器人硬件额外影响的情况下观察视觉语言模型的行为，我们使用了一组图像，而非来自机器人摄像头的实时画面。这些图像是从YouHome日常生活活动数据集中手动选取的 [5], 。该数据集包含来自 20 位用户、涵盖住宅不同区域的 31 种常见日常活动。我们选择这些图像，以便观察模型检测人的能力，以及推断他们正在进行何种活动的能力。此外，我们还尝试了房间看起来空无一人，或有人快速进入房间后又离开的场景。

通过在住宅中部署该系统，我们得以更全面地了解系统的整体性能。我们将机器人放在客厅，并在三种情境下观察其决策过程：有人在看电影、使用手机，以及与宠物狗待在一起。

在第一个案例中（如图 1 所示），机器人被放置在沙发附近，一人正坐在沙发上观看电视节目。尽管视角较低且部分受阻，机器人仍能正确识别这一活动，并以清洁产生的噪音为由，决定暂缓清洁。随后，我们在相同的场景中进行测试，但此时这人专注于手机，且没有播放电影。在这种情况下，视觉语言模型推断，使用手机并不需要保持安静；为了兼顾用户的舒适与清洁目的，它决定开始清洁。最后，当一只狗出现在机器人附近时，系统注意到宠物似乎在周围频繁走动，因此，为了保障宠物的人身安全，不宜进行清洁。系统还指出，巨大的噪音可能会吓到宠物，这表明它的推理能力能够兼顾相关人员和动物的心理健康及安全。

我们的工作存在若干局限。我们测试了推理模型和多种提示技术，包括从示例中学习，但由于较长的提示会带来额外延迟，最终决定不采用这些方法。未来，我们计划进一步探索这些可能性，包括使用本地托管的模型，这将缓解使用住宅家庭环境中的图片所带来的隐私问题。多模态大语言模型的固有偏差 [3] 仍需进一步研究和减轻；可在提示中加入更多信息，例如用户的习惯和偏好，而这些信息可以由用户通过语音或文本表达，并整合到提示中。

4 结论

我们测试了最先进的多模态大语言模型实时推理个人偏好和价值的能力。为此，我们介绍了一种基于 TurtleBot 4 平台和 GPT-4o 的扫地机器人实现方案。机器人通过摄像头采集图像，并据此决定是否开始清洁。我们将机器人部署在客厅中，并观察其行为。总体而言，尽管视角特殊、图像部分受遮挡且时间分辨率有限，语言模型仍成功推断出房间内正在进行的活动，利用这些信息推理其中涉及的价值，并在整洁与尊重相关人员之间取得平衡。

本研究旨在证明，大语言模型集成到具身系统后，不再局限于执行任务，而是开始推理其行动所处的价值观与情境，这标志着家用机器人朝着更具社会意识和适应性的方向迈出了重要一步。

致谢

由“地平线欧洲”计划 VALAWAI 项目资助（资助协议编号：101070930）。

参考文献

[1] Giulio Antonio Abbo 和 Tony Belpaeme。2025。曾经我目不能视，如今我能看见：在社交机器人中实现视觉赋能对话。载于*2025 年 ACM/IEEE 人机交互国际会议论文集（澳大利亚墨尔本）（HRI ’25）*。IEEE Press，1176–1180。

[2] Giulio Antonio Abbo 和 Tony Belpaeme。2025。视觉语言模型作为价值观检测器。载于人工智能中的价值工程，由 Nardine Osman 和 Luc Steels 编。Springer Nature Switzerland，查姆，76–86。

[3] Giulio Antonio Abbo、Tony Belpaeme 和 Micol Spitale。2026。人机交互中的关切与价值观：聚焦社交机器人学。《国际社交机器人学杂志》18，1（2026年1月14日），4。doi:10.1007/s12369-025-01351-1

[4]Steven Macenski、Tully Foote、Brian Gerkey、Chris Lalancette 和 William Woodall。2022。机器人操作系统 2：设计、架构，

及其在真实环境中的应用。*Science Robotics* 7, 66 (2022),eabm6074. doi:10.1126/scirobotics. abm6074

[5] Junhao Pan、Zehua Yuan、Xiaofan Zhang 和 Deming Chen。2022。YouHome 系统与数据集：让您的住宅更了解您。载于 *2022 IEEE* 智能电子系统国际研讨会 *(iSES)*。414–420。 doi:10.1109/iSES54909.2022.00091

[6] Aske Plaat、Annie Wong、Suzan Verberne、Joost Broekens、Niki van Stein 和 Thomas Back。2024。大语言模型推理综述。arXiv:2407.11511 [cs.AI] https: //arxiv.org/abs/2407.11511

[7]Parshin Shojaee、Iman Mirzadeh、Keivan Alizadeh、Maxwell Horton、Samy Bengio 和 Mehrdad Farajtabar。2025。思考的幻象：从问题复杂度的视角理解推理模型的优势与局限。 arXiv:2506.06941 [cs.AI] https://arxiv.org/abs/2506.06941

[8]Gemini Robotics Team、Saminda Abeyruwan、Joshua Ainslie、Jean-Baptiste Alayrac、Montser- rat Gonzalez Arenas、Travis Armstrong、Ash- win Balakrishna、Robert Baruch、Maria Bauza、Michiel Blokzijl、Steven Bohez、Konstanti- nos 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、Clau- dio Fantacci、Cody Fong、Erik Frey、Chuyuan Fu、Marissa Giustina、Keerthana Gopalakrish- nan、Laura Graesser、Leonard Hasenclever、Nico- las Heess、Brandon Hernaez、Alexander Her- zog、R. Alex Hofer、Jan Humplik、Atil Iscen、Mithun George Jacob、Deepali Jain、Ryan Julian、Dmitry Kalashnikov、M. Emre Karagozler、Ste- fani Karp、Chase Kew、Jerad Kirkland、Sean Kir- mani、Yuheng Kuang、Thomas Lampe、Antoine Laurens、Isabel Leal、Alex X. Lee、Tsang-Wei Ed- ward Lee、Jacky Liang、Yixin Lin、Sharath Maddineni、Anirudha Majumdar、Assaf Hur- witz 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 Sori- cut、Jost Tobias Springenberg、Rachel Sterneck、Razvan Surdulescu、Jie Tan、Jonathan Tompson, Vincent Vanhoucke、Jake Varley、Grace Vesom、Giulia Vezzani、Oriol Vinyals、Ayzaan Wahid、StefanWelker、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，以及Yux- iang Zhou。2025。《Gemini Robotics：将人工智能带入物理世界》。arXiv:2503.20020 https://arxiv.org/abs/2503.20020