Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

Seeing You Seeing Me: Augmenting Human-Robot Dialogue With Vision Language Models

Thomas Sievers[0000−0002−8675−0122]

Institute of Information Systems, University of Lübeck, 23562 Lübeck, Germany
t.sievers@uni-luebeck.de

Abstract. Vision Language Models (VLMs) enable robots to visually perceive their environment as well as the actions and characteristics of their conversation partner or a human in collaboration. Ambiguities and allusions in text-based dialogue can often only be resolved if additional visual information is available. Especially for social robots deployed in everyday settings and for uncomplicated, natural use, it is essential that the robot has an understanding of situations that is appropriate to human customs. This paper presents initial experiences with the application of a Mistral AI language model with a Pepper robot for Human-Robot Interaction (HRI) in dialogue. The results show that incorporating visual information gives the dialogue more context and enables both the robot and human to take unspoken elements of the situation into account. Furthermore, using an LLM hosted in Europe provides a solution that complies with European data protection regulations.

Keywords: Social robots · Vision language model · Large language model · Human-robot interaction.

1 Introduction

Robots use many different sensors to perceive their environment, orient themselves within it, and explore it. Most of these sensors function very differently from comparable human sensory organs and give these machines superior abilities. However, when we think about collaboration between humans and robots, or simply a conversation in an informative or social setting, the robot must be able to perceive and refer to objects and situations in a human compatible way. The machine must see the same things as its human counterpart and be capable of reflecting on and discussing them.

With the emergence of Large Language Models (LLMs) and their success in Human-Robot Interaction (HRI) applications, it has become easy and almost normal to communicate with robots using natural language. But language alone sometimes does not convey everything that is important in a dialogue. Ambiguities and allusions can often only be resolved when visual information is also provided. Especially for the use of social robots in everyday situations, for example in the home environment or in care, and for uncomplicated, natural use,

it is essential that the robot has a level of comprehension that is adequate for human customs. In general, social robots must be designed and developed in such a way that they can meet the requirements of their social environment and respond appropriately and comprehensibly [11]. A human-centered perspective should improve the acceptance of social robots in HRI. The inclusion of visual information helps to correctly classify the context during an interaction.

Vision-Language Models (VLMs) are multimodal AI systems created by combining an LLM with a vision encoder that gives the LLM the ability to 'see' [9]. Such systems can be used to convey to a robot, based on visual cues, a natural language description of what is happening in front of it from its own perspective. Furthermore, these observations can be directly utilized by an LLM with additional VLM capabilities for its reasoning on the respective task in order to formulate a response to a human utterance.

This work demonstrates the use of a language model from Mistral AI with a Pepper robot [12]. Choosing an LLM hosted in Europe offers European users a solution that complies with European data protection regulations.

2 Related Work

Robots' understanding of situations in scenarios where they interact with humans is often improved through multimodal reasoning. Integrating language and vision from a robot's perspective through reflection processes helps to go beyond basic navigation and object recognition for robots in environments shared with humans [8]. This is necessary both for social encounters and for industrial environments involving human-robot collaboration [10, 17, 7].

The navigation of robots in dynamic, human-centered environments requires socially acceptable decisions based on a solid understanding of the environment, including spatial-temporal perception and the ability to interpret human intentions [16]. Munje et al. investigated whether VLMs can reliably perform the complex spatiotemporal inferences and intention interpretations required for safe and socially compliant robot navigation, and presented a dataset and benchmark for evaluating VLMs in terms of their scene understanding in real-world social robot navigation scenarios [13]. To address ethical concerns arising from biases in user data when attempting to create user-specific adaptability in VLMs for HRI, Rahimi et al. developed a framework that integrated multimodal user modeling with bias-aware optimization [14].

A tool for improving traditional text-based prompts for LLMs with real-time visual inputs was presented by Abbo et al. [2]. Text from dialogues between humans in different environments and a robot were supplemented for prompt engineering by summaries of images from the robot's perspective. For visual input, they used images from a video recorded during the conversation. Our implementation uses regularly updated individual images captured by the camera in the robot's head, with the most recent image being integrated into the prompt to the LLM.

3 Methods

The Mistral multimodal model Pixtral 12B (pixtral-12b-2409) was used for text and image processing [3]. The LLM was instructed to relate to the content of an image taken with the robot's camera equipment. The robot's front camera took a picture every four seconds to capture an up-to-date impression of what the robot saw. The interval of four seconds was chosen arbitrarily, but seemed to be a practical compromise between timeliness and resource efficiency. The latest image was sent to the Mistral Application Programming Interface (API) for chat completion together with the person's statement in the system prompt.

3.1 Humanoid Social Robot Pepper

To test the capabilities of the selected Mistral AI LLM in a human-robot dialogue, I used the humanoid social robot Pepper. Pepper was developed by Aldebaran and first released in 2015 [15]. The robot is 120 centimeters tall and optimized for HRI. It is able to engage with people through conversation, gestures and its touch screen. The robot features an open and fully programmable platform so that developers can program their own applications using software development kits (SDKs) for programming languages like C++, Python or Java respectively Kotlin [4].

The robot application forwarded the utterances of a human conversation partner, which had been received and processed by Pepper's speech-to-text system, as text input to the Mistral API, including the latest image from the robot's front camera. The returned API response was then spoken by the robot. With each API call, the whole previous dialog was transferred to the model, allowing it to constantly 'remember' what was previously said and refer to it as the dialog progressed. However, only one image was sent with each API call in order to keep data transfer and model costs low.

3.2 Prompting the LLM

I used prompts for the system role to instruct the model to execute the tasks as a completion task with zero-shot prompting [6]. The conversation and all prompts were in German.

System prompt: 'You are a robot named Pepper. The image is taken from your perspective and shows you the environment you are in and people you can interact with. Do not describe what you see in the image, but rather incorporate the information from the image into your conclusions. If there is a person present, engage in conversation based on your impressions. Be curious and focus on the person's activity or any accessories they have, and keep your comments brief. Include the entire course of the conversation in your analysis of the image. Do not repeat any details of the image that you have already mentioned!'

As Abbo et al. mentioned, it is necessary to instruct the model to keep the answers short and concise and to take into account the environment and possible accessories of the persons [2]. An instruction to be curious also improved the dialogue outcome.

4 Interactions

To test the effectiveness of the proposed use of the Mistral LLM with vision capabilities for interacting with a robot in dialogue, Pepper was positioned in five different scenarios. In addition to the different environments, the human interaction partner wore clothing and carried utensils appropriate to the context, which the robot could take into account in its remarks. Figure 1 shows the robot placed in example scenarios for interaction with a human. The robot's tablet displayed a photo of the current camera view from the robot's perspective. This allowed the robot's field of vision to be tracked for this experiment.

Fig. 1. The Pepper robot interacting in different example scenarios.

Scenario 1 in a living room with many books on the shelf in the background. The robot Pepper correctly recognized a large bookshelf in the background and a person sitting across from him at the table. The person was drinking from a cup. Pepper asked the person if they liked to read and whether they drank tea or coffee. As the conversation continued, the robot concluded that the person was probably taking a break from reading.

Scenario 2 outside on the terrace overlooking the garden. The robot referred to the beautiful weather and the lush greenery of the garden and adjacent fields. When asked specifically about objects, it mentioned a bench that was actually located at the back of the garden.

Scenario 3 in the living room with a person sitting on the couch. Pepper commented on the comfort of the couch and referred to the correctly recognized inscription on the person's T-shirt, which raised questions about its meaning (in this case, British television series).

Scenario 4 in the kitchen with one person at the stove. The robot recognized the kitchen environment, but due to its angle of view relative to the head of the interaction partner, it was unable to recognize what was on the stove. When a cooking pot was placed in the robot's field of vision, it was correctly recognized and mentioned in the dialogue.

Scenario 5 with a person working at a computer. Pepper correctly recognized a work situation at a computer with a monitor, whereby, depending on the viewing angle, with occasionally incomplete coverage of the entire scenario, sometimes a tablet and sometimes the actual laptop was recognized. The robot asked about the type of work being done on the computer and offered assistance.

5 Limitations

The inclusion of visual information from the robot's perspective in dialogue with humans enriched the naturalness of the interaction by incorporating context and allowing the robot to proactively refer to the circumstances. Situational elements such as the appearance of the environment, the person's clothing, or objects used were incorporated into the generation of an LLM utterance.

However, the language model sometimes tended to repeatedly describe the entire scene during the course of the dialogue, even when only one detail was actually asked for. Further testing and optimization of prompt engineering could remedy this situation. Occasionally, there were delays of a few seconds in the LLM's response generation, which might have been shorter without visual components. In general, a relatively fluid conversation was possible.

In this experiment, the robot Pepper captured its surroundings with the camera on its forehead between its eyes. Since Pepper also tried to keep its gaze fixed on its conversation partner at all times, the environment outside this field of vision was hardly noticed and thus escaped inclusion in the dialogue. It would be feasible for the robot to be able to avert or lower its gaze, if possible upon request, in order to perceive more details that could be relevant to the conversation.

6 Conclusion and Future Work

Particularly for use within the European Union, the ability to use powerful LLMs from European servers via a simple API integration offers the advantage of not conflicting with European data protection guidelines, as is usually the case when using OpenAI models, for example. This considerably facilitates use in many public institutions such as care facilities or schools.

An extension of the investigations with regard to group interactions or indications from the body language of the interaction partner would be interesting for future work. One could also consider the effects of a type of memory and learning mechanism on the robot's understanding of visual cues. A humanoid robot such as Pepper could possibly use its arms and hands to point at or grasp objects if a correlation could be established between recognized objects in the robot's line of sight and the orientation of its limbs.

In any case, incorporating visual information describing the scene into a robot's dialogue output opens up possibilities for more natural, context-sensitive interaction that is more appropriate for humans.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Abbas, A.N., Beleznai, C.: Talkwithmachines: Enhancing human-robot interaction through large/vision language models. In: 2024 Eighth IEEE International Conference on Robotic Computing (IRC). pp. 253–258 (2024). https://doi.org/10.1109/IRC63610.2024.00039
2. Abbo, G.A., Belpaeme, T.: I was blind but now i see: Implementing vision-enabled dialogue in social robots. In: 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 1176–1180 (2025). https://doi.org/10.1109/HRI61500.2025.10973830
3. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., Monicault, B.D., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., Jiang, A.Q., Khandelwal, K., Lacroix, T., Lample, G., Casas, D.L., Lavril, T., Scao, T.L., Lo, A., Marshall, W., Martin, L., Mensch, A., Muddireddy, P., Nemychnikova, V., Pellat, M., Platen, P.V., Raghuraman, N., Rozière, B., Sablayrolles, A., Saulnier, L., Sauvestre, R., Shang, W., Soletskyi, R., Stewart, L., Stock, P., Studnia, J., Subramanian, S., Vaze, S., Wang, T., Yang, S.: Pixtral 12b (2024), https://arxiv.org/abs/2410.07073
4. Aldebaran, United Robotics Group and Softbank Robotics: Pepper sdk for android. Tech. rep. (2025), https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/index.html
5. Asuzu, K., Singh, H., Idrissi, M.: Human–robot interaction through joint robot planning with large language models. Intelligent Service Robotics **18**, 261–277 (01 2025). https://doi.org/10.1007/s11370-024-00570-1
6. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,

Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020). https://doi.org/10.5555/3495724.3495883

7. Fan, J., Yin, Y., Wang, T., Dong, W., Zheng, P., Wang, L.: Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. Frontiers of Engineering Management **12**, 177–200 (2025). https://doi.org/10.1007/s42524-025-4136-9
8. Galatolo, A., Cumbal, R.: Look, think, understand: Multimodal reasoning for socially-aware robotics. In: ICRA 2025 Workshop: Human-Centered Robot Learning in the Era of Big Data and Large Models (2025), https://openreview.net/forum?id=lRgveD8OgE
9. Ghosh, A., Acharya, A., Saha, S., Jain, V., Chadha, A.: Exploring the frontier of vision-language models: A survey of current methodologies and future directions (2024), https://arxiv.org/abs/2404.07214
10. Kawaharazuka, K., Obinata, Y., Kanazawa, N., Okada, K., Inaba, M.: Robotic applications of pre-trained vision-language models to various recognition behaviors. In: 2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids). pp. 1–8. IEEE (2023). https://doi.org/10.1109/humanoids57100.2023.10375211
11. Mahdi, H., Akgun, S.A., Saleh, S., Dautenhahn, K.: A survey on the design and evolution of social robots — past, present and future. Robotics and Autonomous Systems **156** (2022). https://doi.org/10.1016/j.robot.2022.104193
12. Mistral AI: La plateforme - mistral ai. Tech. rep. (2025), https://console.mistral.ai/
13. Munje, M.J., Tang, C., Liu, S., Hu, Z., Zhu, Y., Cui, J., Warnell, G., Biswas, J., Stone, P.: Socialnav-SUB: Benchmarking VLMs for scene understanding in social robot navigation. In: ICRA 2025 Workshop: Human-Centered Robot Learning in the Era of Big Data and Large Models (2025), https://openreview.net/forum?id=cCuylmKVXq
14. Rahimi, H., Bahaj, A., Abrini, M., Khoramshahi, M., Ghogho, M., Chetouani, M.: User-vlm 360: Personalized vision language models with user-aware tuning for social human-robot interactions (2025), https://arxiv.org/abs/2502.10636
15. Softbank Robotics: Meet pepper. Tech. rep. (2025), https://us.softbankrobotics.com/pepper
16. Song, D., Liang, J., Payandeh, A., Raj, A.H., Xiao, X., Manocha, D.: Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. IEEE Robotics and Automation Letters **10**(1), 508–515 (2025). https://doi.org/10.1109/LRA.2024.3511409
17. Tan, R., Yang, H., Jiao, S., Shan, L., Tao, C., Jiao, R.: Smart robot manipulation using gpt-4o vision. In: 7th European Industrial Engineering and Operations Management Conference, Augsburg, Germany. IEOM Society International (2024). https://doi.org/10.46254/EU07.20240272